Intelligent Systems, Control and Automation:
Science and Engineering 90

Spyros G. Tzafestas

# Energy, Information, Feedback, Adaptation, and Self-organization

## The Fundamental Elements of Life and Society

Springer

# Intelligent Systems, Control and Automation: Science and Engineering

Volume 90

More information about this series at http://www.springer.com/series/6259

Spyros G. Tzafestas

# Energy, Information, Feedback, Adaptation, and Self-organization

## The Fundamental Elements of Life and Society

Springer

Spyros G. Tzafestas
School of Electrical and Computer
Engineering
National Technical University of Athens
Athens
Greece

ISSN 2213-8986 ISSN 2213-8994 (electronic)
Intelligent Systems, Control and Automation: Science and Engineering
ISBN 978-3-319-66998-4 ISBN 978-3-319-66999-1 (eBook)
https://doi.org/10.1007/978-3-319-66999-1

Library of Congress Control Number: 2017956737



Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

# Preface

*Society is well governed when its people obey the magistrates, and the magistrates obey the law.*

Solon

*I cannot teach anybody anything. I can only make them think.*

Socrates

*Science is the creator of prosperity.*

Plato

*Freedom is the sure possession of those alone who have the courage to defend it.*

Pericles

*Wealth consists not in having great possessions, but in having few wants.*

Epicurus

The aim of this book is to provide a comprehensive conceptual account of the five fundamental elements of life and society, *viz*., energy, information, feedback, adaptation, and self-organization. These elements inherently support any living organism, human society, or man-made system.

*Energy* is the cornerstone of everything. *Information* is included in the "program" (organized plan) of any living organism, to function over time, which is implemented by the DNA that encodes the genes and is transferred from generation to generation. It is one of the main factors of the progress of modern society which is characterized as the *"information society"*. *Feedback* (control) is a "must" for any kind of system, biological, natural, or technological, to be stable and operate according to its purpose. *Adaptation* is the capability of living organisms, species, and societies to adapt to changes that occur in their environment so as to fit to it. It

is the principle that lies behind the natural selection and evolution. *Self-organization* has many interpretations, the predominant of which is the "tendency" of natural systems to become more organized by their own, and shows more structure or order or pattern without the help or intervention of any external agent. This means that spontaneous emergence of global complex structure occurs out of local interactions.

All the above aspects of life and society have been of principal concern to humans over time, and a plethora of concepts and scientific or technological methodologies were developed and studied. The topics addressed in this book are the subject matter in a vast number of sources in the literature and the web. The book gives a collective and cohesive presentation of the fundamental issues, concepts, principles, and methods drawn from the literature, including modern applications and short historical notes of each field. The presentation is kept at a level sufficient for a clear understanding of the concepts and principles by the general scientific reader. In many cases, viz., thermodynamics, communication systems, information theory, and feedback control, the discussion includes the basic mathematical analysis aspects in some more detail which are deemed to be necessary and useful for the nonprofessionals. Unavoidably, the material provided in the book does not exhaust all the results and views available in the literature. However, it is considered to be over sufficient for disseminating the fundamental concepts and issues. The views and opinions/quotations on the delicate aspects of life and society, presented in the book, are those coined and published by the referenced authors. No attempt was made to modify or speculate them in any way.

The writing of this book was inspired by the need of a concise, cohesive, and complete presentation of the five life-and-society fundamental elements (pillars): energy, information, feedback, adaptation, and self-organization in a unique volume. Surely, besides the general reader, this book will be valuable as a source for introductory or complementary material in relevant science and engineering academic programs.

The book involves 13 chapters. Chapter 1 provides an introduction to the book presenting the background concepts of life and society, and outlining the five fundamental elements of life and society considered in the book.

Chapters 2 and 3 are devoted to the *energy*. Chapter 2 presents the basic issues of energy (historical landmarks, types, sources, and environmental impact), and Chap. 3 is devoted to thermodynamics (basic concepts, laws of thermodynamics, entropy, exergy, branches of thermodynamics, and entropy interpretations).

Chapters 4 and 5 are concerned with the *information* element. Chapter 4 introduces the concept of information and reviews the communication systems and information theory. Chapter 5 discusses information science, information technology, and information systems in enterprises and organizations.

Chapters 6 and 7 are devoted to the *feedback element*. Chapter 6 presents the concept of feedback and control, the history of its study, and the methods for linear and nonlinear control systems analysis and design developed between about 1935 and 1950 (classical control). Chapter 7 reviews the modern control techniques which are based on the state-space model, namely, Lyapunov stability,

state-feedback (eigenvalue/model matching) control, and optimal control (deterministic and stochastic). The classes of adaptive, predictive, robust, nonlinear, and intelligent control are also discussed.

Chapter 8 is concerned with the *adaptation* in biology and society including the related scientific fields of complexity and complex adaptive systems.

Chapter 9 is devoted to the final fundamental element studied in the book, i.e., the *self-organization* of natural and societal systems. The four self-organization mechanisms observed in nature are first reviewed, and the concept of self-organized criticality (edge of chaos) is then discussed. The role of *cybernetics* in the study of self-organization is also examined.

Chapters 10 through 13 are concerned with the role and impact of the five fundamental elements studied in the book on life and society discussing major issues and a variety of examples. Chapter 10 discusses the fundamental role that energy plays in life and society, starting with an examination of the three basic biochemical pathways of energy in life (photosynthesis, respiration, and metabolism) and going to the energy flow in ecosystems. The evolution of energy resources, the thermoeconomy, and the saving of energy in the human society are then investigated.

Chapter 11 deals with a number of issues that refer to the role of information in life and society. These include the substantiative and transmission roles of information in biology, and the information technology applications in modern society, such as office automation, power generation /distribution, computer-assisted manufacturing, robotics, business/e-commerce, education, medicine, and transportation.

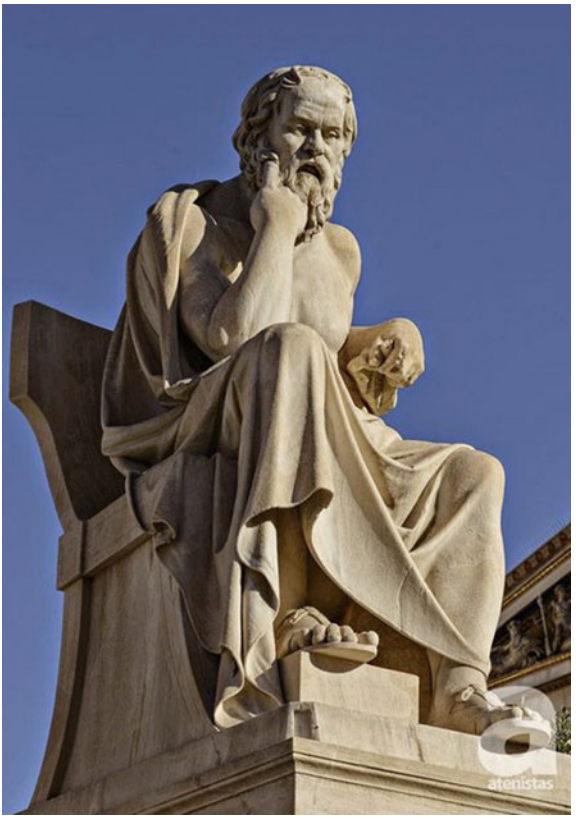
Chapter 12 reviews the role and impact of feedback in both living organisms and societal systems. Representative examples that best show the operation of negative and positive feedback in biology and society are provided. These include temperature, water, sugar, and hydrogen ion (pH) regulation, autocatalytic (autoreproduction) reactions, enzyme operation, cardiovascular–respiratory system, process control, manufacturing control, air flight and traffic control, robot control, management control, and economic control systems.

Finally, Chap. 13 provides a number of adaptation and self-organization examples and applications in life and society. These examples are adaptations of animals, ecosystems, climate change, immune systems, social–ecological systems, capital /stock market, general society system, knowledge management, and man-made self-organizing systems.

In overall, the book provides a cohesive and complete picture of the five fundamental elements: energy, information, feedback, adaptation, and self-organization, and the role they play in sustained life and society, including selected modern applications.

Athens, Greece
June 2016

Spyros G. Tzafestas




Only one thing I know, that I know nothing.

Only Absolute Truth is that there are No Absolute Truths.

**Socrates, Athens, 470-399 B.C.**

Footnote: *Statue of Socrates in front of Athens Academy*

(Sculptor: Leonidas Droses/1885. Photographer: Elias Georgouleas/2014, ‘’atenistas”: www.athenssculptures.com).

Picture taken from www.athenssculptures.com by courtesy of “athens sculptures atenistas”




Humans and Society: Synergy, hierarchy of society, social life. *Sources*
http://crossfitlando.com/wp-content/uploads/2013/04/earth-day.jpeg,
http://thesocialworkexam.com/wp-content/uploads/2011/03/Human-Behavior-Hierarchy.jpg,
http://www.urbansplash.co.uk/images/ABOUTUS_SOCIETY.jpg

# Contents


**1 Life and Human Society: The Five Fundamental Elements** . . . . . 1
1.1 Introduction . . . . . 1
1.2 What Is Life? . . . . . 2
1.2.1 General Issues . . . . . 2
1.2.2 The Living Cell . . . . . 3
1.2.3 DNA, Nucleotides, and Protein Formation . . . . . 5
1.2.4 Historical Landmarks of DNA and RNA Discoveries . . . . . 10
1.2.5 Koshland's Definition of Life . . . . . 11
1.3 The Meaning of Society . . . . . 13
1.4 Evolution of Life and Human Society . . . . . 16
1.4.1 Origin and Evolution of Life . . . . . 16
1.4.2 Evolution and the Development of Human Society . . . 21
1.5 Fundamental Elements of Some Specific Societal Aspects . . . . 28
1.5.1 Pillars of Democracy . . . . . 28
1.5.2 Pillars of Fulfilled Living . . . . . 29
1.5.3 Pillars of Sustainable Development . . . . . 30
1.6 The Five Fundamental Elements of This Book . . . . . 31
1.7 Concluding Remarks . . . . . 34
References . . . . . 35

**2 Energy I: General Issues** . . . . . 39
2.1 Introduction . . . . . 39
2.2 What is Energy? . . . . . 40
2.3 Historical Landmarks . . . . . 42
2.4 Energy Types . . . . . 45
2.4.1 Mechanical Energy . . . . . 46
2.4.2 Forms of Potential Energy . . . . . 47
2.4.3 Internal Energy in Thermodynamics . . . . . 48
2.4.4 Evidence of Energy . . . . . 50

2.5 Energy Sources . . . 51
2.5.1 Exhaustible Sources . . . 51
2.5.2 Renewable Sources . . . 54
2.5.3 Alternative Energy Sources . . . 59
2.6 Environmental Impact of Energy . . . 61
2.6.1 Impact of Exhaustible Sources . . . 61
2.6.2 Impact of Renewable Sources . . . 63
2.7 Violent Manifestations of Earth's Energy . . . 65
2.7.1 Earthquakes and Volcanoes . . . 65
2.7.2 Tornadoes and Hurricanes . . . 67
2.7.3 Tsunamis . . . 68
2.8 Concluding Remarks . . . 70
References . . . 70

**3 Energy II: Thermodynamics** . . . 73
3.1 Introduction . . . 73
3.2 Basic Concepts of Thermodynamics . . . 75
3.2.1 Intensive and Extensive Properties . . . 75
3.2.2 System and Universe . . . 76
3.2.3 System State . . . 77
3.2.4 Thermodynamic Equilibrium . . . 77
3.2.5 Temperature and Pressure . . . 78
3.2.6 Heat and Specific Heat . . . 78
3.2.7 Reversible and Irreversible Process . . . 79
3.2.8 Categories of Thermodynamic Processes . . . 80
3.2.9 Basic Concepts of Non-statistical General Physics . . . 82
3.3 The Zeroth Law of Thermodynamics . . . 83
3.4 The First Law of Thermodynamics . . . 84
3.4.1 Formulation of the Law . . . 84
3.4.2 The Thermodynamic Identity: Energy Balance . . . 86
3.5 The Entropy Concept . . . 88
3.5.1 The Classical Macroscopic Entropy . . . 89
3.5.2 The Statistical Concept of Entropy . . . 92
3.5.3 The Von Neumann Quantum-Mechanics Entropy Concept . . . 94
3.5.4 The Non-statistical General Physics Entropy Concept . . . 97
3.5.5 Rényi Entropy, Tsallis Entropy, and Other Entropy Types . . . 100
3.6 The Second Law of Thermodynamics . . . 103
3.6.1 General Formulations . . . 103
3.6.2 Formulations Through Entropy . . . 104
3.6.3 Formulation Through Exergy . . . 107
3.7 The Third Law of Thermodynamics . . . 111

3.8 The Fourth Law of Thermodynamics . . . 112
3.8.1 Lotka's Maximum Energy-Flux Principle . . . 112
3.8.2 Odum's Maximum Rate of Useful-Energy-Transformation Principle . . . 113
3.8.3 Onsager Reciprocal Relations . . . 114
3.8.4 Some Further Fourth-Law Statements . . . 115
3.9 Branches of Thermodynamics . . . 116
3.9.1 Traditional Branches . . . 117
3.9.2 Natural Systems Branches . . . 119
3.9.3 Modern Branches . . . 121
3.10 Entropy Interpretations . . . 126
3.10.1 Entropy Interpretation as Unavailable Energy . . . 126
3.10.2 Entropy Interpretation as Disorder . . . 127
3.10.3 Entropy Interpretation as Energy Dispersal . . . 128
3.10.4 Entropy Interpretation as Opposite to Potential . . . 129
3.11 Maxwell's Demon . . . 130
3.12 The Arrow of Time . . . 132
3.12.1 Psychological Arrow . . . 134
3.12.2 Thermodynamic Arrow . . . 134
3.12.3 Cosmological Arrow . . . 135
3.12.4 Quantum Arrow . . . 138
3.12.5 Electromagnetic Arrow . . . 139
3.12.6 The Causal Arrow . . . 139
3.12.7 The Helical Arrow . . . 140
3.13 Conclusions and Quotes for Thermodynamics, Entropy, and Life . . . 141
3.13.1 Thermodynamics General Quotes . . . 141
3.13.2 Entropy Quotes . . . 143
3.13.3 Life and Human Thermodynamics Quotes . . . 144
References . . . 146

**4 Information I: Communication, Transmission, and Information Theory** . . . 157
4.1 Introduction . . . 157
4.2 What Is Information? . . . 158
4.3 Historical Landmarks . . . 161
4.3.1 Pre-mechanical Period . . . 161
4.3.2 Mechanical Period . . . 162
4.3.3 Electromechanical Period . . . 162
4.3.4 Electronic Period . . . 163
4.3.5 Information Theory Landmarks . . . 164
4.3.6 Computer Networks, Multimedia, and Telematics Landmarks . . . 165

4.4 Communication Systems . . . 168
4.4.1 General Issues . . . 168
4.4.2 Shannon–Weaver Communication Model . . . 169
4.4.3 Other Communication Models . . . 170
4.4.4 Transmitter–Receiver Operations . . . 172
4.4.5 Analysis of Analog Modulation–Demodulation . . . 174
4.4.6 Pulse Modulation and Demodulation . . . 185
4.5 Information Theory . . . 190
4.5.1 General Issues . . . 190
4.5.2 Information Theory's Entropy . . . 191
4.5.3 Coding Theory . . . 198
4.5.4 Fundamental Theorems of Information Theory . . . 205
4.5.5 Jayne's Maximum Entropy Principle . . . 211
4.6 Concluding Remarks . . . 214
References . . . 215

**5 Information II: Science, Technology, and Systems** . . . 219
5.1 Introduction . . . 220
5.2 Information Science . . . 220
5.3 Information Technology . . . 225
5.3.1 Computer Science . . . 225
5.3.2 Computer Engineering . . . 238
5.3.3 Telecommunications . . . 250
5.4 Information Systems . . . 264
5.4.1 General Issues . . . 264
5.4.2 General Structure and Types of Information Systems . . . 265
5.4.3 Development of Information Systems . . . 268
5.5 Conclusions . . . 271
References . . . 272

**6 Feedback and Control I: History and Classical Methodologies** . . . 277
6.1 Introduction . . . 277
6.2 The Concept of Feedback . . . 278
6.2.1 General Definition . . . 278
6.2.2 Positive and Negative Feedback . . . 279
6.3 The Concept of Control . . . 281
6.4 Historical Landmarks of Feedback and Control . . . 283
6.4.1 Prehistoric and Early Control Period . . . 284
6.4.2 Pre-classical Control Period . . . 285
6.4.3 Classical Control Period . . . 286
6.4.4 Modern Control Period . . . 288
6.5 Classical Control . . . 290
6.5.1 Introductory Issues . . . 290
6.5.2 The Basic Feedback Control Loop . . . 291

6.5.3 System Stability . . . 294
6.5.4 System Performance Specifications . . . 295
6.5.5 Second-Order Systems . . . 297
6.6 The Root-Locus Method . . . 299
6.7 Frequency-Domain Methods . . . 301
6.7.1 Nyquist Method . . . 301
6.7.2 Bode Method . . . 305
6.7.3 Nichols Method . . . 310
6.8 Discrete-Time Systems . . . 310
6.8.1 General Issues . . . 310
6.8.2 Root Locus of Discrete-Time Systems . . . 314
6.8.3 Nyquist Criterion for Discrete-Time Systems . . . 315
6.8.4 Discrete-Time Nyquist Criterion with the Bode and Nichols Plots . . . 316
6.9 Design of Classical Compensators . . . 317
6.9.1 General Issues . . . 317
6.9.2 Design via Root Locus . . . 318
6.9.3 Design via Frequency-Domain Methods . . . 319
6.9.4 Discrete-Time Compensator Design via Root-Locus . . . 320
6.10 Ziegler-Nichols Method for PID Controller Tuning . . . 320
6.11 Nonlinear Systems: Describing Functions and Phase-Plane Methods . . . 324
6.11.1 Describing Functions . . . 324
6.11.2 Oscillations Condition . . . 326
6.11.3 Stability Investigation of Nonlinear Systems via Describing Functions and Nyquist Plots . . . 328
6.11.4 Application of Root Locus to Nonlinear Systems . . . 329
6.11.5 Phase Plane . . . 330
6.12 Concluding Remarks . . . 333
References . . . 334

**7 Feedback and Control II: Modern Methodologies** . . . 337
7.1 Introduction . . . 338
7.2 The State-Space Model . . . 338
7.2.1 General Issues . . . 338
7.2.2 Canonical Linear State-Space Models . . . 340
7.2.3 Analytical Solution of the State Equations . . . 342
7.3 Lyapunov Stability . . . 343
7.3.1 General Issues . . . 343
7.3.2 Direct Lyapunov Method . . . 344
7.4 Controllability and Observability . . . 345
7.4.1 Controllability . . . 345
7.4.2 Observability . . . 347
7.4.3 Controllability-Observability, Duality, and Kalman Decomposition . . . 348

7.5 State-Feedback Controllers . . . 349
7.5.1 General Issues . . . 349
7.5.2 Eigenvalue Placement Controller . . . 351
7.5.3 Discrete-Time Systems . . . 352
7.5.4 Decoupling Controller . . . 352
7.5.5 Model Matching Controller . . . 354
7.5.6 State-Observer Design . . . 355
7.6 Optimal and Stochastic Control . . . 356
7.6.1 General Issues: Principle of Optimality . . . 356
7.6.2 Application of the Principle of Optimality to Continuous-Time Systems . . . 358
7.6.3 Linear Systems with Quadratic Cost . . . 359
7.6.4 Pontryagin Minimum Principle . . . 360
7.6.5 Stochastic Optimal Control . . . 362
7.7 Adaptive and Predictive Control . . . 365
7.7.1 General Issues . . . 365
7.7.2 Model-Reference Adaptive Control . . . 366
7.7.3 Self-tuning Control . . . 368
7.7.4 Gain-Scheduling Control . . . 370
7.7.5 Model-Predictive Control . . . 371
7.8 Robust Control . . . 372
7.8.1 General Issues . . . 372
7.8.2 Non-stochastic Uncertainty Modeling . . . 373
7.8.3 Formulation of Robust Control Design . . . 374
7.9 Nonlinear Control . . . 375
7.9.1 State-Feedback Linearizing Control . . . 375
7.9.2 Optimal Nonlinear Control . . . 377
7.9.3 Robust Nonlinear Sliding-Mode Control . . . 377
7.10 Intelligent Control . . . 379
7.11 Control of Further System Types . . . 387
7.11.1 Control of Large-Scale Systems . . . 388
7.11.2 Control of Distributed-Parameter Systems . . . 389
7.11.3 Control of Time-Delay Systems . . . 392
7.11.4 Control of Finite-State Automata . . . 396
7.11.5 Control of Discrete-Event Systems . . . 398
7.12 Conclusions . . . 402
References . . . 403

**8 Adaptation, Complexity, and Complex Adaptive Systems** . . . 409
8.1 Introduction . . . 410
8.2 What Is Adaptation? . . . 411
8.3 Historical Note . . . 413
8.4 Adaptation Mechanisms . . . 415

8.5 Adaptation Measurement . . . 420
8.6 Complex Adaptive Systems . . . 421
8.6.1 General Issues . . . 421
8.6.2 A Concise Definition of CAS . . . 423
8.7 List of Reference Works on Complexity, Complex Systems, and Complex Adaptive Systems . . . 424
8.8 Chaos and Nonlinear Systems . . . 425
8.8.1 Fractals and Strange Attractors . . . 425
8.8.2 Solitons . . . 434
8.9 Complexity . . . 437
8.10 Emergence . . . 441
8.11 More on Complex Adaptive Systems . . . 445
8.12 Concluding Remarks . . . 454
References . . . 455

**9 Self-organization** . . . 461
9.1 Introduction . . . 462
9.2 What Is Self-organization? . . . 463
9.2.1 Definition of W. Ross Ashby . . . 463
9.2.2 Definition of Francis Heylinghen . . . 464
9.2.3 Definition of Chris Lucas . . . 464
9.2.4 Definition of Scott Camazine . . . 464
9.2.5 Definition of A.N. Whitehead . . . 465
9.2.6 Definition of M. B. L Dempster . . . 465
9.3 Mechanisms of Self-organization . . . 466
9.4 Self-organized Criticality . . . 468
9.5 Self-organization and Cybernetics . . . 470
9.6 Self-organization in Complex Adaptive Systems . . . 474
9.7 Examples of Self-organization . . . 478
9.7.1 Linguistic Self-organization . . . 480
9.7.2 Knowledge Networks . . . 481
9.7.3 Self-organizing Maps . . . 482
9.8 Concluding Remarks . . . 483
References . . . 486

**10 Energy in Life and Society** . . . 489
10.1 Introduction . . . 490
10.2 Energy and Life: Biochemical Pathways . . . 490
10.2.1 Photosynthesis . . . 491
10.2.2 Respiration . . . 494
10.2.3 Metabolism . . . 497
10.3 Energy Movement in an Ecosystem . . . 498
10.3.1 General Issues . . . 498
10.3.2 Energy Flow Through Food Chains . . . 499
10.3.3 Efficiency of Energy Flow Through Food Chains . . . 501

10.4 Energy and Human Society . . . 505
10.4.1 General Issues . . . 505
10.4.2 Evolution of Energy Resources . . . 506
10.5 Energy and Economy . . . 514
10.5.1 General Issues: Thermoeconomics . . . 514
10.5.2 Sectors of Economy . . . 516
10.6 Management of Energy . . . 518
10.7 Demand Management, Economics, and Consumption of Energy . . . 519
10.7.1 Energy Demand Management and Energy Economics . . . 519
10.7.2 Consumption of Energy . . . 520
10.8 Concluding Remarks . . . 530
References . . . 531

**11 Information in Life and Society** . . . 535
11.1 Introduction . . . 536
11.2 Information and Life . . . 537
11.2.1 General Issues . . . 537
11.2.2 Substantive Role of Information in Biology . . . 538
11.2.3 The Transmission Sense of Information in Biology . . . 540
11.3 Natural Information Processing Principles . . . 542
11.3.1 The Information Store Principle . . . 543
11.3.2 The Borrowing and Reorganizing Principle . . . 543
11.3.3 Randomness as a Genesis Principle . . . 544
11.3.4 The Narrow Limits of Change Principle . . . 545
11.3.5 The Environmental Organizing and Linking Principle . . . 545
11.4 Biocomputation . . . 546
11.5 Information and Society: Introduction . . . 548
11.6 Information Technology in Office Automation . . . 551
11.7 Computer-Based Power Generation and Distribution . . . 553
11.8 Computer-Integrated Manufacturing . . . 556
11.9 Information Technology in Business: Electronic Commerce . . . 560
11.10 Information Technology in Education . . . 561
11.11 Information Technology in Medicine . . . 563
11.12 Information and Communication Technology in Transportation . . . 564
11.13 Concluding Remarks . . . 568
References . . . 570

**12 Feedback Control in Life and Society** . . . 575
12.1 Introduction . . . 576
12.2 Feedback Control in Living Organisms . . . 577
12.2.1 General Issues . . . 577

12.2.2 Negative Feedback Biological Systems . . . . . 577
12.2.3 Positive Feedback Biological Systems . . . . . 583
12.3 Systems and Control Methods for Biological Processes . . . . . 590
12.3.1 System Modeling of Biological Processes . . . . . 590
12.4 Feedback Control in Society . . . . . 601
12.4.1 General Issues . . . . . 601
12.4.2 Hard Technological Systems . . . . . 602
12.4.3 Soft-Control Systems . . . . . 614
12.5 Concluding Remarks . . . . . 620
References . . . . . 620

**13 Adaptation and Self-organization in Life and Society** . . . . . 627
13.1 Introduction . . . . . 628
13.2 Adaptations of Animals . . . . . 629
13.3 Ecosystems as Complex Adaptive Systems . . . . . 631
13.4 Adaptation to Climate Change . . . . . 632
13.5 Adaptation of Immune and Social-Ecological Systems . . . . . 635
13.6 Stock Markets as Complex Adaptive Systems . . . . . 637
13.7 Society Is a Self-organizing System . . . . . 639
13.8 Knowledge Management in Self-organizing Social Systems . . . . . 642
13.9 Man-Made Self-organizing Controllers . . . . . 647
13.9.1 A General Methodology . . . . . 647
13.9.2 Self-organizing Traffic-Lights Control . . . . . 649
13.10 Concluding Remarks . . . . . 652
References . . . . . 656

**Index** . . . . . 661

# Chapter 1
# Life and Human Society: The Five Fundamental Elements

*The goal of life is living in agreement with nature.*
Zeno of Elea (490–435 B.C.)
*The good life is one inspired by love and guided by knowledge.*
Bertrand Russel

**Abstract** The aim of this chapter is to provide fundamental material about life and society (definition, evolution, etc.), starting with a brief presentation of cell biology, DNA/RNA, protein synthesis, and a list of the principal discoveries about DNA and RNA. The meaning of "society" is discussed, followed by the evolution of life on Earth, and the evolution of human society (physical, vital, and mental stages). The common fundamental elements (pillars) of life and society that are studied in this book, namely: energy, information, feedback, adaptation, and self-organization, are briefly introduced. As a supplement, the chapter includes a short outline of some purely societal fundamental elements that are encountered in humanity studies. These elements are: (i) pillars of democracy, (ii) pillars of fulfilled living, and (iii) pillars of sustainable development.

**Keywords** Life · Society · Molecular biology · Cell biology · Life domains
Energy · Information · Feedback · Adaptation · Self-organization
Evolution of life · History of life · Evolution of human society
Human development · DNA · Life-program · Pillarsof democracy
Pillars of fulfilled living · pillars of sustainable development

## 1.1 Introduction

This chapter serves as an introduction to the book by providing some background concepts about life and society, specifically their definitions and evolution. These concepts will help the reader to go smoothly to the five particular "*elements*" *or* "*pillars*" of life and society studied in the book. The term "*pillar*" is used in several frameworks of life and society, some of which will be discussed in Sect. 1.5. Koshland has used the term "pillar" for the definition of 'life' [1]. According to



S. G. Tzafestas, *Energy, Information, Feedback, Adaptation, and Self-organization*,
Intelligent Systems, Control and Automation: Science and Engineering 90,
https://doi.org/10.1007/978-3-319-66999-1_1

Webster (1913) the term "pillar" literally means "a firm, upright, insulated support for a superstructure; a pier, column, or post; also, a column or shaft not supporting a superstructure, as one erected for a monument or an ornament" [2, 3]. "Figuratively, that which resembles a pillar in appearance, character or office; a supporter or mainstay" (as: the Pillars of Hercules; a pillar of the state, etc.) or "anything tall and thin approximating the shape of a column or tower". In science, pillar may be called "a fundamental principle or practice".

The questions "*what is life*" and "*what is society*" were of primary concern to humankind throughout the centuries of historical record and have been studied by philosophers, scientists, biologists, sociologists, archaeologists, geographers, etc. Today we have better informed and more developed views of what is life and how it evolved since the formation of Earth 4.5 billion years ago. We know that human societies are essentially "*adaptive systems*" the elements of which, "*human populations*", strive to satisfy their varied needs and wishes. History has shown that these needs and wishes have been accomplished either by maintaining existing ways of doing things or by developing and adopting new, innovative ways. In all cases, the parts that failed to adapt were eliminated from the system, while those that succeeded survived. This is exactly the "*principle of survival of the fittest*" which holds in all biological and sociological processes.

The structure of the chapter is as follows. Section 1.2 deals with the question "*what is life*". It starts with a brief presentation of cell biology, DNA, and protein synthesis. Then it lists the main discoveries about DNA and RNA, and provides the definition of life coined by *Daniel Koshland*. Section 1.3 is concerned with the meaning of society (*Richard Jenkins'* and *Richard Alston's* views). Section 1.4 outlines the evolution of life (prokaryotes, eukaryotes, etc.) and society (physical, vital, and mental stages). It also includes a discussion of *human development* (requirements, components, economic models) and *human development index*. Section 1.5 describes briefly the fundamental elements of some societal aspects, other than the five elements that are the subject matter of the book, namely: democracy, fulfilled living, and sustainable development. Finally, Sect. 1.6 discusses the scope of the book and places the five pillars: energy, information, feedback, adaptation, and self-organization, in their proper position which is in the "intersection" of the "*biological*" and "*societal*" sets of pillars [1–118].

## 1.2 What Is Life?

### *1.2.1 General Issues*

The reply to this question appears to be simple: "A living organism is an organized entity which is able to grow and sustain itself through metabolic processes (absorption of energy), to respond to stimuli, to protect itself from external attacks or injuries, and to be reproduced". This is a very primitive definition of life not

capturing all the facets of life. Actually, many biologists have the opinion that there still does not exist a clear, definite, and complete definition of life. One of the reasons seems to be the existence of *viruses* and other *microscopic entities*. Many biologists suggest that viruses are complex organic molecules, but others consider viruses as the simplest form of life. No one knows with certainty how life began. But we know for certain that all life on Earth involves strings of **DNA** (**D**eoxyribo**N**ucleic **A**cid) that are long chains of self-replicating molecules which encode information (*genes*) and implement the so-called *life-program*. We also know that life (except of viruses) is constructed by cells, i.e., tiny containers which contain the DNA and other chemical compounds that make up the cells. The early life forms were single cells. To understand what life is and later supply an apparently complete list of *features* (or *pillars*) that define life, we first give a short review of *cell-biology* (biological cell) [4–11].

### *1.2.2 The Living Cell*

A *living organism* may be composed of a single biological cell (*single-cell organisms*) or of many cells (*multiple-cell organisms*). The biological cell can sustain its functionality through a set of *organelles* (which are "*miniature*" machines) that each have a special function. Some of them in case of the animal cells are the following Fig. 1.1 Analogous organelles exist in the plant cell (http://waynesword.palomar.edu/lmexer1a.htm).[1]

**Cell membrane or plasma membrane** This is the external layer of a cell that has a structural and protective role affecting how molecules enter or exit the cell.

- **Nucleus** This is the "*brain*" of the cell that contains the genetic information about the processes taking place in an organism. It is surrounded by the *nuclear membrane*.
- **Nucleolus** This resides inside the nucleus and is the organelle where ribosomal RNA is produced.
- **Cytoplasm** This is the fluid that surrounds the contents of a cell.
- **Mitochondrion** This is an organelle that participates in respiration (i.e., in the energy release and storage; it is also called "*powerhouse*" of the cell).
- **Ribosomes** These are packets of **RNA** (**R**ibo**N**ucleic **A**cid) and protein. They are the site of protein synthesis. Messenger RNA from the nucleus moves systematically along the ribosome where transfer RNA adds individual amino-acid molecules to the lengthening protein chain.
- **Lysosomes** These are sacs filled in with digestive enzymes.

[1](*) The term "biology" comes from the Greek "Βίος" (bios = life) and "λόγος" (logos = speech/study), (**) All web sources and references were collected during the writing of the book. Since then, some of the urls may not be valid due to change or removal.

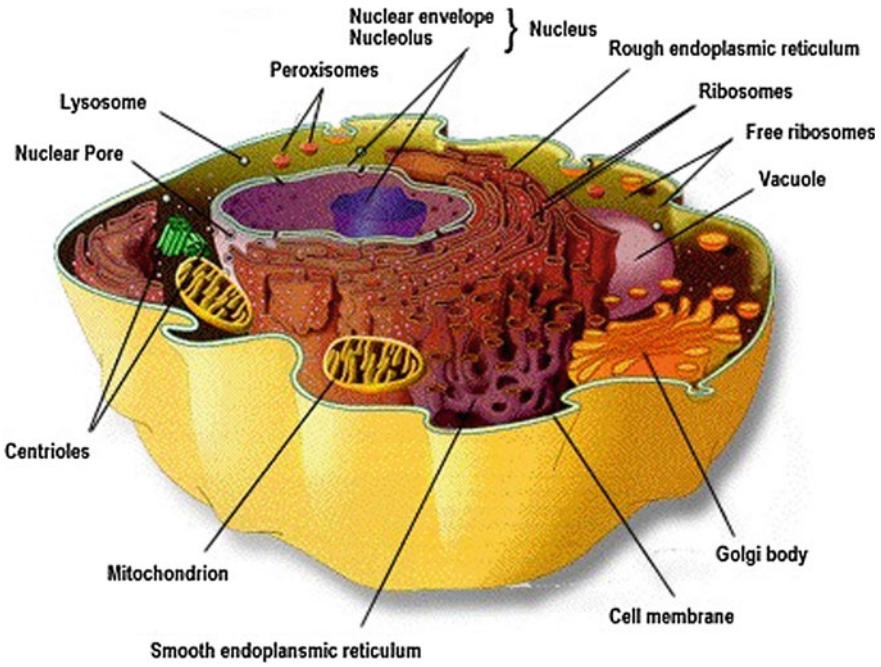



**Fig. 1.1** Schematic of eukaryote animal cell with the basic organelles. *Source* http://www.odec.ca/projects/2004/mcgo4s0/public_html/t1/animalcell3.jpg (The reader is informed that Web figures and references were collected at the time of writing the book. Since then, some of them may not be valid due to change or removal by their creators, and so they may no longer be available)

- **Golgi body/complex** They are involved in the production of glycoprotein.
- **Vacuole** Cavities filled with food being digested and waste material to go out of the cell.
- **Centrosome** A small body located near the nucleus, also called "the microtubule organizer center". It is the organelle where microtubules are made during cell division (mitosis).
- **Endoplasmic reticulum (ER)** A useful organelle, differentiated into rough ER and smooth ER, which is involved in the synthesis of protein.

The cell (or plasma) membrane, which is a semi-permeable structure composed by proteins and fat (*phospholipid*) molecules, acts as a circumferential barrier and allows only selected compounds to get in and out of a cell. The transportation of *ions* via the cell membrane into the cell is performed in three ways: *active transport* (based on concentration gradient), *passive transport* (diffusion via a carrier), and *simple diffusion* (such as *osmosis of water*). The uptake of materials from the external environment of the cell is called *absorption,* and the ejection of material is called *secretion*. A full animal-cell picture with labels is provided by *Russell Kightley Media* [21].

The cells are specialized to each perform a distinct function within an organism. Thus we have, for example:

- **Skin cells** They function as waterproof and pathogen protection from the cell's exterior environment.
- **Nerve cells** These cells, also called *neurons,* are electrically excitable cells that function within the nervous system for message transmission to and from the central nervous system.
- **Muscle cells** These cells have an elastic capability and enable flexible movement (as in our muscles).
- **White blood cells** They activate suitable digestive enzymes that break down pathogens to the molecular level, thus eliminating them.

Biological cells have the capability to break down complex molecules into simple molecules, which can then be used as building elements of other complex molecules. This is done via *pinocytosis* (e.g., drinking bacteria after breaking down them into drinkable form) or *phagocytosis* (in which the original material is eaten, after it has been broken down into a suitable form).

### *1.2.3 DNA, Nucleotides, and Protein Formation*

The type, structure, and functioning of cells are determined by *chromosomes* (*from the Greek words chroma = color and soma = body*) which reside in the cell nucleus. These chromosomes are made from DNA bonded to various proteins in the nucleus of eukaryotic cells or as a circular strand of DNA in the cytoplasm of prokaryotes and in the mitochondrion (and chloroplast) of some eukaryotes. The DNA specifies all the features of an organism, containing all the genetic material that makes what a living being is. This material (information) is transferred from generation to generation in a species, determining the offsprings' characteristics. The building blocks of DNA are the *nucleotides* which appear as four different types, namely: *adenine* (**A**), *guanine* (**G**), *thymine* (**T**), and *cytosine* (**C**). Our genome contains billions of these nucleotides in all possible permutations, located in adjacent pairs along the *double-helix* arrangement of DNA. Actually, there are two groups of bases, namely, *purines* and *pyrimidines*. Purines (adenine and guanine) have a two-ring structure, whereas pyrimidines (thymine and cytosine) have a single-ring structure. *Complementary* (or *permissible*) bases are bases that pair together in a DNA molecule. These base pairs are:

- Thymine and adenine (**TA**)
- Gyanine and cytosine (**GC**)

Thymine and cytosine cannot make a base pair, and similarly adenine and guanine cannot form a base pair.

While **DNA** resides mainly in the nucleus, the nucleic acid polymer **RNA** (*Ribonucleic acid*) is found mainly in the cytoplasm, despite the fact that it is usually synthesized in the nucleus. DNA contains the genetic codes to make RNA, and RNA contains the codes for the primary sequences of amino acids to make proteins.

The backbone of the polymer is a repeating chain of sugar-phosphate-sugar-phosphate, etc. The pentose sugar of DNA is a *deoxyribose* sugar, whereas RNA contains a *ribose* sugar. Both DNA and RNA contain a phosphate group and a nitrogenous base as shown in Fig. 1.2.

The pentose is a five-membered, puckered ring. Attached to the ring is the phosphate group (which is a phosphorous atom with four covalently attached oxygen atoms) and the nitrogenous base. Pictures of DNA and RNA models are provided in http://www.dreamstime.com/stock-images-structure-dna-rna-molecule-vector-image28618424http://www.dreamstime.com/stock-images-structure-dna-rna-molecule-vector-image28618424

In the RNA model, we have an extra-OH of the pentose sugar, and the *uracil* base (**U**) is used instead of the thymine base (**T**) used in DNA (Fig. 1.3).

The cells—which have finite life spans—pass their genetic information to new cells replicating *exactly* the DNA to be transferred to offsprings. To this end, a

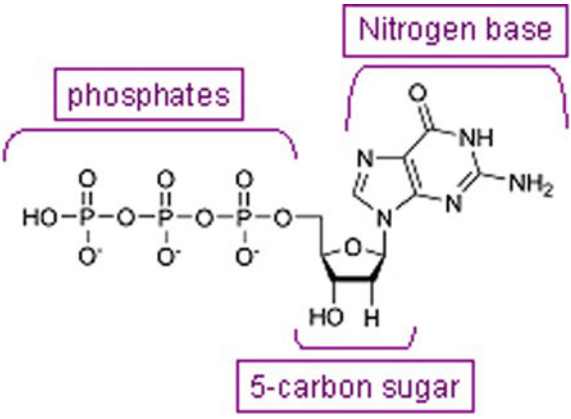
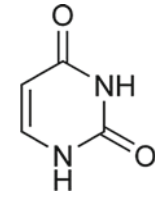



**Fig. 1.2** Structure of connected pentose sugar, phosphate group, and nitrogenous base in DNA and RNA



Uracil (U)
(RNA only)

**Fig. 1.3** Uracil base (a single-ring pyrimidine)

supply of suitable *enzymes* that stimulate the reaction process is available, together with a pool of the required nucleotides. The actual DNA acts as an *exact template*. The energy needed for this transfer is provided by **ATP** (**A**denosine **Tri**p**h**osphate) molecules (see Chap. 10).

Actually, the replication of the double-helix DNA involves two strands of DNA, each one of which produces a copy of itself. The replicated DNA has only half of the original material from its parent (i.e., it is *semi-conservative*). Therefore the two copies produced have the full (exact) DNA material contained in the two strands of the DNA involved in the replication. This is the way genetic information is transferred from cell to cell and from parent to offspring.

The sequence of the nucleotides is used to create *amino acids*, the chains of which are shaped so as to make a protein. An amino-acid molecule consists of the basic *amino group* (**$NH_2$**), the *acidic carboxylic group* (**COOH**), a *hydrogen atom* (**H**), and an *organic side group* (**R**) attached to the carbon atom. Thus, an amino acid has the structure $NH_2CHRCOOH$. Actually, there exist more than a hundred amino acids in nature, each of them differing in the R group. Twenty of them participate in protein synthesis and are differentiated into *essential* and *non-essential* amino acids. *Essential* (or indispensable) amino acids cannot be created in the body and can only be acquired via food. *Non-essential* (or dispensable) amino acids are synthesized in the body. These twenty amino acids are the following:

- **Essential** Histidine, isoleucine, leucine, lysine, methionine, phenylalanine, threonine, tryptophan, and valine.
- **Non-essential** Alanine, arginine, aspartic acid, asparagine, cysteine, glutamic acid, glutamine, glycine, proline, serine, and tyrosine.

The structure of proteins spans four levels of complexity, namely:

- *Primary structure* (the sequence of amino acids).
- *Secondary structure* (local folding sustained via short-distance interactions; hydrogen bonds).
- *Tertiary structure* (additional folding sustained via more distant interactions between alpha helices and pleated sheets).
- *Quaternary structure* (sustained by interchain interactions of more than one acid chain).

Although the sequence must determine the structure, we cannot yet predict the full structure accurately from a sequence. Structures are stable and relatively rigid. Today, there are about 4000 known protein structures determined by X-ray crystallography and 2-D NMR studies. The above four-level structure of proteins is depicted in Fig. 1.4.

The synthesis of proteins takes place in the *ribosomes* residing in the cell's cytoplasm, whereas the genetic information lies in the nucleus. Thus, the genetic information has to pass to these ribosomes. This transfer is performed by **mRNA**

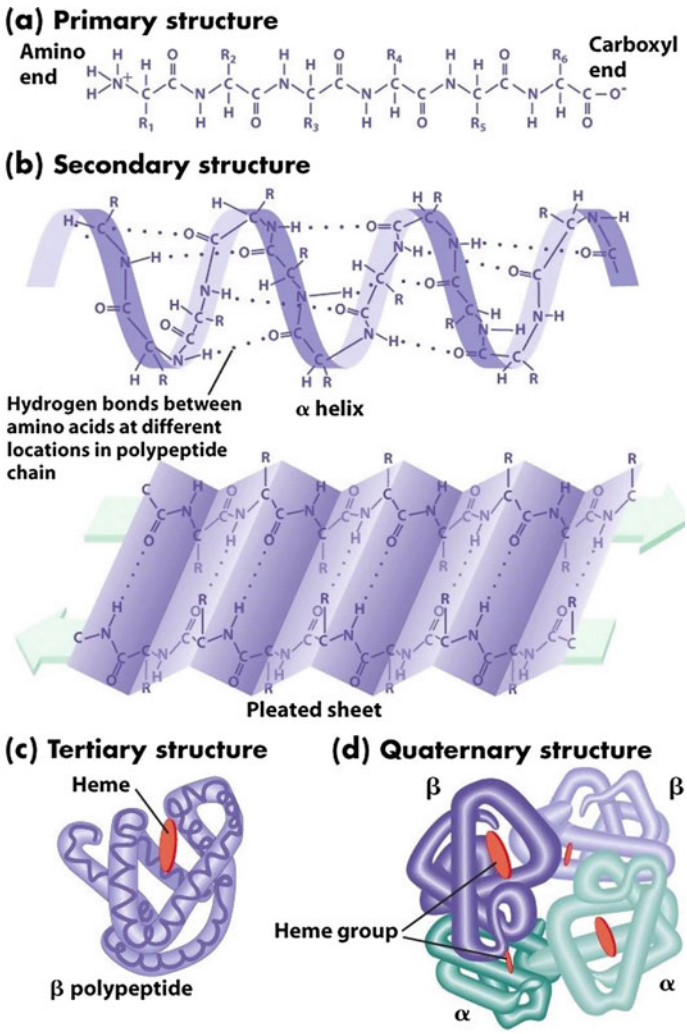



**Fig. 1.4** The four-complexity levels of proteins (*source* [18])

(**m**essenger **r**ibo**n**ucleic **a**cid), which is analogous to **DNA**, differing only in two aspects.

- In mRNA, the thymine bases are replaced by a base called *uracil* (**U**).
- The deoxyribose sugar of DNA is substituted by ribose sugar.
  The transfer is performed in the following sequence:
- Inside the cell's nucleus, genes (DNA) are transcribed into RNA. To this end, the double-helix structure of DNA uncoils for mRNA to replicate, like the DNA, the genetic sequence of which corresponds to the protein under synthesis.

- This RNA produces a mature mRNA through post-transcription modification and control.
- The mRNA is transported out of the nucleus and travels through the cytoplasm until it reaches a ribosome where it is translated into protein. Since ribosomes don't understand the mRNA code, they use their *translator*, i.e., the *transfer RNA*s (**tRNA**s). The RNAs decode the message and assemble the desired amino acids in the specified sequence to form the protein which is released into the cytoplasm for further transport and processing.

The above scheme for *protein synthesis* (known as "*dogma of molecular biology*") is pictorially illustrated in Fig. 1.5.

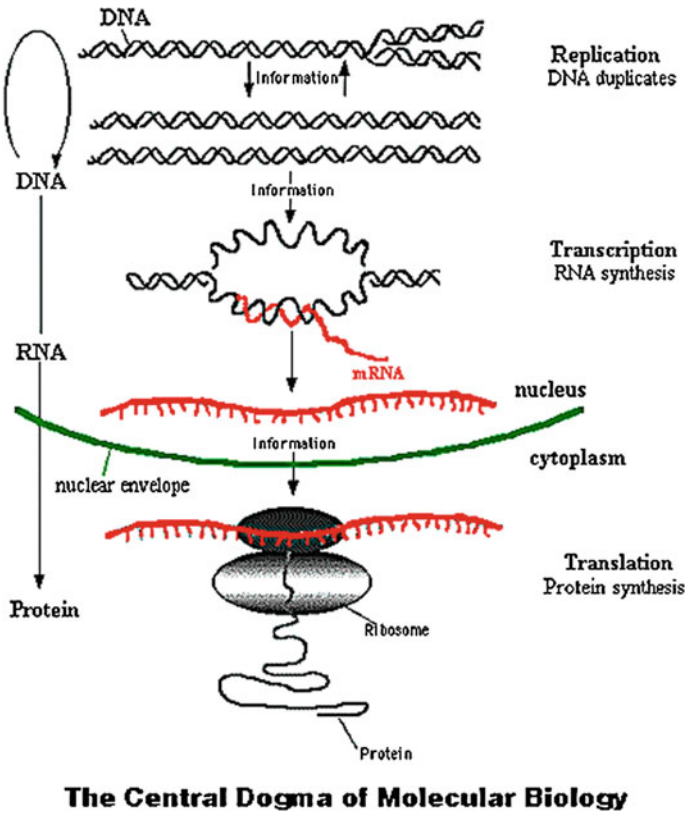



**Fig. 1.5** Synthesis of protein (*source* [9])

### 1.2.4 Historical Landmarks of DNA and RNA Discoveries

Until the 1800s, it was believed that life arose more or less spontaneously, but in 1864 *Louis Pasteur* disproved spontaneous generation. He demonstrated that, when any micro-organisms residing in a liquid are killed through boiling, the liquid becomes *sterile* (nothing grows afterwards). After *Pasteur,* the principal historical landmarks of the "*RNA world*" are here listed chronologically [12]. The complete historical evolution of DNA and RNA discoveries and studies can be found in [13–17].

1924: *Alexander Ivanovich Oparin* attributes the origin coming of the simplest single-cell life to simple organic molecules residing in the early Earth's atmosphere that was substantially different from our present atmosphere—there was no free oxygen, but there was abundant hydrogen, ammonia, methane, carbon dioxide, water, and nitrogen).

1953: *James Watson* and *Francis Crick* publish their results on the structure of DNA. They received a joint Nobel Prize for these results in 1962.

1961: *Marshal Nirenberg* and colleagues discover that messenger RNA, composed completely of the base uracil, can be translated into the amino acid phenylalanine.

1968: *Francis Crick* and *Leslie Orgel* argue that the first information molecule was RNA.

1972: *Harry Noller* suggests that ribosomal RNA plays a role in the translation of mRNA into protein.

1986: *Walter Gilbert* uses the term "*RNA world*" for the time during which RNA was the main information and catalytic molecule. *Thomas Cech* presents his discovery of *self-splicing* (catalytic RNA). In 1989, he shares a Nobel Prize with *Sidney Altman* for the catalytic RNA discovery. *Kary Mullis* presents a procedure for rapid copying of DNA and RNA sequences (p*olymerase chain reaction*). He was awarded a Nobel Prize for this in 1993.

1989: *Gerald Joyce* starts his work on simulating RNA evolution via the "polymerase hain reaction". *Jack Szostak's* lab provides evidence for self-replicating RNA.

1992: *Harry Noller's* lab provides experimental evidence for the involvement of ribosomal RNA in protein synthesis.

1993: *Gerald Joyce* presents test-tube experimental processes for RNA evolution.

1994: *Charles Wilson* (while working in Szostak's labs) creates RNA molecules that are able to perform simple cellular reactions more efficiently than the proteins, which perform it in cells.

Complete presentations of "*molecular cell biology*" are provided in [19, 20], where both *genomics* (the complete DNA sequences of many organisms), and *proteomics* (all possible shapes and functions that proteins employ) are studied. The principal topics considered include:

- The dynamic cell
- Nucleic acids and genetic code

- From gene to protein
- Protein structure and function
- Genetic analysis
- DNA replication, repair, and recombination
- RNA processing and post-transcriptional control
- The mechanism of translation
- Gene control in development
- Cell-to-cell signaling: hormones and receptors
- Genome analysis
- Epigenetics and monoallelic gene expression
- Medical molecular biology.

A useful site with biology images, videos, and cell-interactive animation is provided by *Cells Alive Com* in [22].

### *1.2.5 Koshland's Definition of Life*

With the background on molecular and cell biology provided in Sects. 1.2.2–1.2.4, we can now proceed and examine the seven fundamental elements (pillars) that define life as presented by the molecular biologist *Daniel Koshland* (2002) [1], where the term "*pillars*" is used to mean "*the essential principles* (*thermodynamic and kinetic*) *that enable a living system to operate and propagate*". These seven pillars, although essential to the distinct mechanisms by which the life's principles are implemented on Earth, may be complemented by other pillars, as well that may explain better the mechanisms of life so far known or other mechanisms to be discovered in the future for other forms of life or for life elsewhere [1]. Koshland's seven pillars defining life are the following:

- Program
- Improvisation
- Compartmentalization
- Energy
- Regeneration
- Adaptability
- Seclusion,

Which can be represented by a Temple, called as a whole by the acronym **PICERAS**. A brief description of the pillars follows Fig. 1.6.

**Program** Koshland states that "*program* is the organized plan that describes both the ingredients themselves and the kinetics of the interactions among ingredients as the living system persists through time". These interactions and processes involve the metabolic reactions that enable a living organism to function over time. Each program of a living system on Earth is implemented by the DNA which encodes the

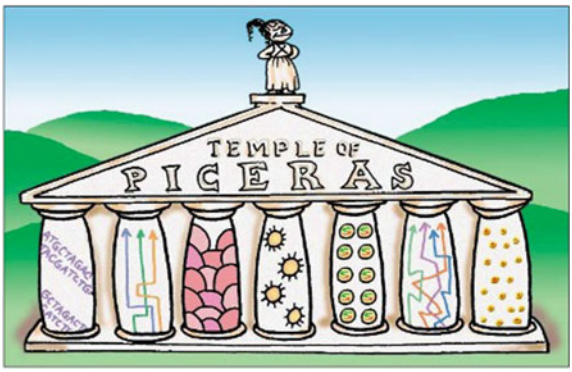



**Fig. 1.6** Koshland's seven-pillar temple for the definition of life [1]. *Source* www.astro.iag.usp.br/~amancio/aga0316_artigos/Koshland02.pdf

genes, is replicated from generation to generation, and operates through the mechanisms of nucleic acids and amino acids, as briefly described in Sect. 1.2.3.

**Improvisation** This refers to the capability of living systems to modify their programs in response to the wider environment in which they live. This modification (change) of program is realized by *mutation* and *selection* with the aid of which the program is optimized under the environmental constraints.

**Compartmentalization** This refers to the fact that all living organisms are confined to limited space, surrounded by a surface (*membrane* or *skin*), which separates the living organism from the environment. In this way, the ingredients of the organism are kept inside a definite volume and any dangerous substances (toxic or diluting) are kept outside this volume. Thus compartmentalization protects the living organism's ingredients from reactions to the external environment.

**Energy** Living systems take energy from their environment and change it from one form to another. This energy is necessary for the chemical activities or body movements of the living system to take place, during which energy quality is degraded and entropy is produced. The principal source of Earth's energy is the *Sun*, but of course other energy sources (exhaustible or non-exhaustible) exist for human life on Earth (see Chap. 2).

**Regeneration** This is the ability of living systems to replace parts of themselves that experience wear and degradation. Regeneration balances the thermodynamic losses in chemical reactions, the wear and tear of larger parts, and the decline of ingredients due to ageing. For example, the human body continually re-synthesizes and replaces its heart muscle proteins as they suffer degradation. The same is true for other constituents, such as lung sacs, kidney proteins, brain synapses, and so on. In general, living systems balance the occurring losses by synthesizing fresh molecules and parts, or importing compounds from their environment, or producing new generations to start the system over again. However, in spite of the regeneration capability, all living systems (organisms) degrade into an ultimate non-functioning state (i.e., death).

**Adaptability** This is the capability of living systems to adapt and evolve in response to changes in their environment. Improvisation is a kind of adaptability, but is too slow for many of the environmental hazards to which a living organism may be exposed. Thus Koshland considers adaptability as different from improvisation because its action is timely and does not imply any change of the program. Adaptability takes place from the molecular to behavioral level through feedback and feed-forward operations (see Chap. 8). For example, our bodies respond to depletion of nutrients (energy supplies) with hunger, which drives us to seek food, and also our appetite feedback system prevents excess eating [1]. When an animal sees a predator, it responds to the danger with hormonal changes and escape behavior. In general, influences from the environment leads to adaptive reaction through metabolic and physiological response and behavioral action.

**Seclusion** This is the ability of living systems to separate chemical pathways in order to secure that metabolic and other processes occurring simultaneously within the organism are not confused or mixed. This is achieved thanks to the *specificity* of *enzymes* that function only on the molecules for which they were designed. Such specificity holds also in DNA and RNA interactions. Seclusion is the property of life that enables numerous reactions to take place efficiently within the tiny volumes of living cells, while, at the same time, they receive and process specialized signals to accommodate changes in the environment.

Defining life is problematic especially at the level of 'bacteria', where, among others, the question "what we really mean by using the word species for a bacterium", arises. Schrodinger argued that life is not a mysterious phenomenon, but a scientifically comprehensible process that might be ultimately explained by the laws of physics and chemistry.

## 1.3 The Meaning of Society

The term *society* has its origin in the Latin *societas* from *socius,* and the French *societé* which means companion, chum, comrade, associate, or partner. At Dictionary.com, one can find several alternative meanings of the word society. Some of them are the following [23]:

- A group of humans broadly distinguished from other groups by mutual interests, participation in characteristic relationships, shared institutions, and a common culture.
- An organized group of persons, associated together for religious, benevolent, cultural, scientific, political, patriotic, or other purposes.
- A body of individuals living as members of a community.
- A highly structured system of human organization for large-scale community living that normally furnishes protection, continuity, security, and a national identity to its members.

- An organization or association of persons engaged in a common profession, activity, or interest.
- The totality of social relationships among humans.
- In biological terms, society is a closely integrated group of social organisms of the same species exhibiting division of labor.

According to sociologist *Richard Jenkins,* the term society refers to critical existential issues of humans, namely [24]:

- The way humans exchange information, including both the sensory abilities and the behavioral interaction.
- Often community-based performance and phenomena cannot be reduced to individual behavior, i.e., the society's action is "greater than the sum of its parts".
- Collectives usually have life spans, exceeding the life span of individual members.
- All aspects of human life are tied together in a collective sense.

According to *Richard Alston* [25]:

"*Society must be understood as a combination of functional, cognitive, and cultural systems*". This view is illustrated by his "triangle heuristic model of society" (Fig. 1.7).

**Cognition** is interpreted as meaning the issues that help people understanding the difference between "*what is*" versus "*what ought to be*".

**Culture** includes the social imperatives such as groups, values, status, roles, authority, ideology, etc.

**Function** includes the institutional aspects of society, namely: norms (i.e., rules that govern activity and behavior), moral statements, and sets of obligations and expectations. *Norms* are combined to produce "*roles*". A set of roles specifies the individual. *Groups* match similar persons and interests and interact by means of

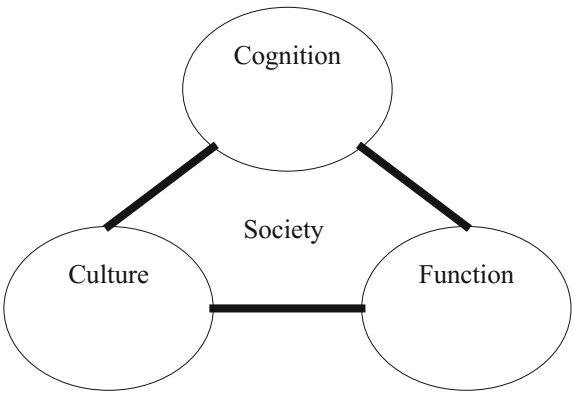



**Fig. 1.7** Alston's heuristic model of society

institutions and institutional complexes. *Institutions* are formed at the proper levels, viz. community, local, state, nation, and international, as may be required.

*Society* is sometimes separated from *culture*. According to *Clifford Geertz* "society is the organization of social relations, while culture emerges from symbolic forms (beliefs, ideologies, etc.)". In political science, the term *society* usually includes all the human relations, typically in contrast to the rulers (government, etc.) within a territory (state). An *ideology* links an individual's perceptions to those dominating in the overall society and provides a basis for consensus at the group level. In other words, ideology is a way of looking at and perceiving the world, shared by the members of a community. A society changes as any one of its elements changes persistently. Change results in stress, but it is needed for adaptation and adjustment to new internal and external conditions (see Chap. 13). When the cognitive, cultural, and functional changes, accumulated over time, are no longer compatible with the ideology that served to interpret them, and solutions to social problems are no more possible within the existing systems of information and social organization, then *revolutionary change* takes place (as, e.g., the revolutionary change occurred in Western societies between the medieval period and the modern era). The study of *human society* is still continuing. Sociologists investigate human behavior from different viewpoints such as the cultural, economic, political, and psychological perspectives, both qualitatively and quantitatively.

Basic issues included in these studies are the following:

- General individual and group behavior (genetic inheritance and social experience/organization factors and issues)
- Cultural factors and norms (tradition, beliefs, etc.)
- Social-conflict factors (internal and external)
- Social-trade-off factors (material or economic)
- Economic factors
- Political factors
- Professional- and scientific-society factors
- Interaction of societies (cultural, civil, technological, etc.)
- Immigration factors
- Social change/evolution factors.

Some free online information sources on these issues of human society diachronically are provided in [26–29]. For an early and a recent book on human society, see [30, 31].

## 1.4 Evolution of Life and Human Society

Both life and human society have taken their present form through several evolutionary processes over the past millions of years (biological life) or thousands of years (human society). Our purpose here is to briefly discuss this evolution separately, for both overall life on Earth and human society.

### *1.4.1 Origin and Evolution of Life*

Scientists estimate that the Earth's atmosphere with the proper composition of oxygen, hydrogen, carbon, and nitrogen that allowed the creation of life was present about 3.9 billion years ago. They also believe that the Sun's energy, heat, and radioactive elements originated the formation of nucleic acids and proteins with replicating genetic code. These tiny entities have then self-organized and evolved, resulting in the first simple forms of life. At about 3.8 billion years ago, the fossilization of Earth's cellular life forms started. The fossilized cells resemble current *cyanobacteria*. These cells were given the name *prokaryotes* (also called *monera*) and contain a few specialized structures with their DNA not confined to a volume defined by a membrane. According to "your Dictionary.com", the name "*prokaryote*" comes from the Greek "προ" (pro = before) and "κάρυον (καρύδι)" (caryon = walnut) [32]. See also [33]. The more complex cells of plants and animals, called *eukaryotes,* from the Greek "ευ" (eu = good) and "κάρυον"

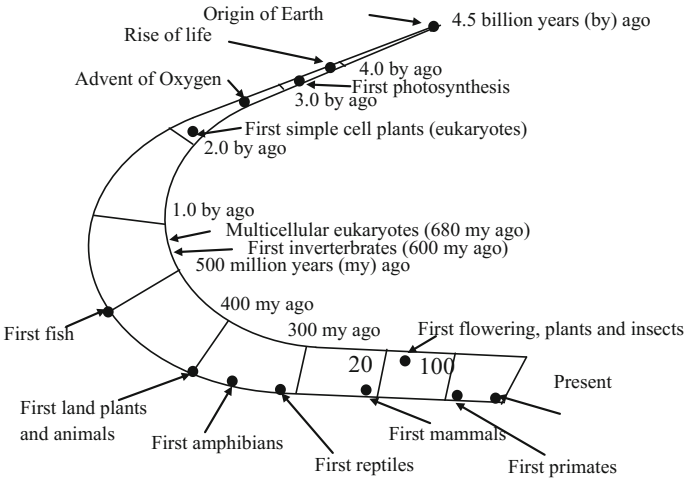



**Fig. 1.8** Approximate time of appearance of the Earth's principal species of plants and animals (Darwinian evolution of species). This estimation is, in many cases, supported by fossil evidence (by = billion years, my = million years). A more detailed pictorial illustration of the evolution of life on Earth is provided in: http://blogs.egu.eu/network/palaeoblog/files/2012/10/life-on-earth.jpg

(karyon = walnut), first appeared about 2.1 billion years ago. These cells have a nucleus bounded by a cell membrane and numerous specialized structures confined within the bounded-cell volume. The *multicellular organisms,* which are made by well-organized collections of eukaryotic cells, first appeared about 680 million years ago. At about 570 million-years ago, multi-cellular life was enormously diversified, with all but one of the *modern phylum* of animal life existing on Earth today. Fish first appeared about 500 million years ago (*Ordovician Period*) [34–37].

Figure 1.8 shows the approximate (estimated) time of the occurrence of the principal events of the evolution of life on Earth, which itself came into existence about 4.5 billion years ago. We can see that the plants and animals familiar to us (e.g., marine invertebrates such as shell-making ammonities) appeared about 540 million years ago), then fish, amphibians, reptiles, mammals, the first primates, and finally humans. The life thread that continues in the oceans indicates that the evolution of aquatic life continues up to our present time.

One of the difficulties of *Darwinian evolution of species* is, according to *Stephan Jay Gould* ([119], p. 14), the existence of "*gaps*" in the fossil record. He states: "Evolution requires intermediate forms between species and the paleontology does not provide them. The gaps must therefore be a contingent feature of the record". This difficulty was also mentioned by *Charles Darwin* himself in his book (The Origins of Species, 1859) as follows: "Geology assuredly does not reveal such finely graduated organic chain; and this perhaps, is the most obvious and gravest objection which can be urged against my theory. The explanation lies, as I believe, in the geological record".

The first known human-like primates (*hominids*) evolved in eastern Africa about 5.2 million years ago, during the so-called *Pliocene Epoch* (5.3–1.8 my), named after the Greek words "*πλείον*" (plion = more) and "*καινός*" (cenos = new) to indicate that there were more new fossil forms than in previous epochs. Most hominids lived probably in groups near (or inside) forests, and many later used tools and weapons. The oldest hominid fossils (a *jawbone teeth* and a *toe bone*) were found in Ethiopia (existing at about 5.3 my ago). A younger, almost complete hominid skeleton (known as *Lucy*) was found in Hadar (Tanzania) revealing that even the earliest hominids could walk upright on two legs. Then, we have the *Pleistocene Epoch* (1,800,000–11,700 y), where the word "pleisto" comes from the Greek "πλείστος" (most), which extended up to the beginning of the Holocene Epoch at 10,000 years ago. By the start of the *Pleistocene*, the Earth entered a cooler period of alternating glacial and interglacial phases, with arctic vegetation inside the *Arctic Circle*, and taiga coniferous evergreen forests. *Homo sapiens* appeared during the Pleistocene epoch about 400,000 years ago (evidence from archaic fossils), and the earliest modern humans appeared only 170,000 years ago. Our scientific knowledge of human evolution is improving continuously, as new fossils are discovered and described every year. In general, over the years, our view of our evolutionary past has changed as social attitudes have changed, after *Darwin's* publication entitled "*The Descent Man*" (1871).

During the *Holocene epoch* (8000 years-present), where the word "*holo*" comes from the Greek "*όλος*" (holos = entire) to indicate the appearance of entirely new fossil appearances, the Earth was relatively warm and had only small scale climate

shifts (such as the so-called "*The Little Ice Age*" which started about 660 years ago (1350) and lasted for about 300 years. The Holocene epoch is sometimes called the "*Age of Man*", although this may be misleading because modern humans evolved and spread over the Earth, influencing the global environment in ways different from any other organism, well before the Holocene period began.

On the basis of the above (Fig. 1.7), scientists divide living creatures into the following three domains:

**Archaea** These are tiny and tough prokaryotes without a nucleus. They have been recently discovered in hostile habitats like volcanic vents, hot springs, and saline pools. Although they are similar to bacteria, molecular research has shown that are biochemically and genetically very different from bacteria.

**Bacteria** Prokaryotes, small cells without nucleus. Except for *cyanobacteria* (see Chap. 10), bacteria do not contain chlorophyll. They get energy to live through the breakdown of organic matter via *fermentation* and *respiration* (i.e., they are *heterotrophs*).

**Eukaria (Eucaria)** Organisms that have an eukaryotic-cell type. This class is further divided in several life kingdoms, the four primary of which are: *Protista, Fungi, Plantae* (plants), and *Animalia* (animals). Details about them can be found in the literature [38–43]. Figures 1.9 shows a *phylogenetic* pictorial representation (*evolutionary tree*) of the above three life domains and the life kingdoms of the domain Eukaria.

Archaea and Eukarya share a common ancestor not shared by the bacteria. The eukaryotic cell probably evolved only once. Many different microbial eukaryotic (protist) groups emerged from this common ancestor [40].

The history of life and the approximate times of appearance of plants and animals (see Fig. 1.9) can also be schematically presented in the form of a 30-day calendar in which each "day" represents about 150 million years [40]. Earth came

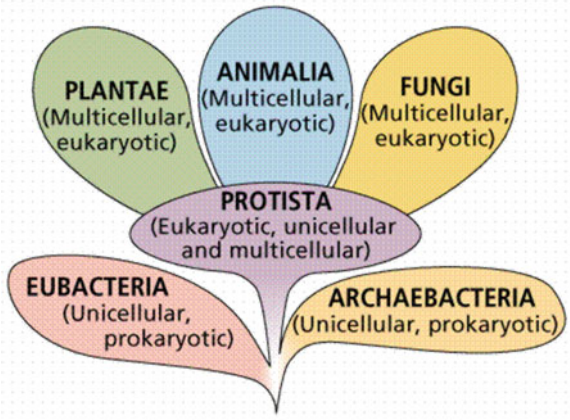



**Fig. 1.9** Phylogenetic presentation of Archaea, Bacteria and Eukaryota kingdoms (*source* [40]). A more detailed phylogenetic representation is given in: http://www.sheppardsoftware.com/content/animals/images/evolution_treeoflifechart.jpg

into existence during the first day (30 × 150 million years = 4.5 billion years ago). The origin of life is placed somewhere between the third and fourth day or about 3.0–4.0 billion years ago. Photosynthesis evolved in the 14th day, eukaryotic cells evolved in the 20th day, and multi-cellular organisms in the 24th day. Aquatic life and abundant fossils appeared in the 27th day, the first land plants and animals in the 28th day; coal-forming forests, insects, first mammals, and dinosaurs during the 29th day; and the first birds, flowering plants and the rise of mammals during the 30th day. *Homo sapiens* appeared in the last 10 min of the 30th day, and recorded history fills the last few seconds of day 30.

The theory of the origin of life and evolution of species described above is the current, dominating theory widely-accepted by most biologists and scientists. However, many biologists and physicists have published strong criticism of the Darwinian approach, from various points of view. Space limitation does not allow a detailed presentation of this criticism, which can be found in the literature. Here we only list a few quotations which summarize the conclusions and views of the respective scientists.

**H.S. Lipson**
"In fact, evolution became in a sense a scientific religion; almost all scientists have accepted it and many are prepared to "bend" their observations to fit in with it" [93].

**Sir Fred Hoyle**
"In short, there is not a shred objective evidence which supports the hypothesis that life began in an organic soup here on Earth" [94].

**David B. Kitts**
"Evolution at least in the sense of Darwin speaks of it, cannot be detected within the lifetime of a single observer" [95].

**Michael Behe**
"Intelligent Design is an explanation of the origin of life that differs from Darwin's view".

"Many systems in the cell show signs of purposeful intelligent design. What science has discovered in the cell in the past 50 years is poorly explained by a gradual theory such as Darwin's" [96].

"Scientific theories are explanations of facts. That is why they are not facts. The Darwinian theory of evolution, like all scientific theories is not a fact". "Gaps in the fossil record constitute valid objections to Darwin's theory of evolution because they are spaces for the miraculous appearance of species that have not evolved from any other source" [96].

"Biology is irreducibly complex" [96].

**Hubert P. Yockey**
"Biology is not irreducible complex because the bit string in the genome that describes a protein is finite and stops after it produces the protein (i.e., the computation is not performed indefinitely)" [101].

"Contrary to the established and current wisdom, a scenario describing the genesis of life on earth by chance and natural causes which can be accepted on the basis of fact and not faith, has not yet been written" [97].

"The fundamental consideration in evolution is the genome, not the fossil record. Gaps in the fossil record do not matter. What matters is that there are no gaps in the genome from the origin of life to the present. It is the continuity of the genome that shows the connectedness of all life-living, extinct and yet-to-be-evolved. That means there are no gaps in which species miraculously appear, as Intelligent Design falsely claims".

"The origin of life is unsolvable as a scientific problem" [101, 102].

**Collin Patterson**

"It is easy enough to make up stories of how one form gave rise to another, and to find reasons why the stages should be favored by natural selection. But such stories are not part of science, for there is no way of putting them to the test" [98].

**Klaus Dose**

"Explanatory power (of the origin of life theory) is weak … There are more questions than answers" [99].

**Freeman Dyson**

"Pathways of evolutionary development in chemical origin of life, remain unexplained and even unimaginable" [100].

**L. Harrison Matthews**

"The fact of evolution is the backbone of biology, and biology is thus in the peculiar position of being a science founded on an unproved theory—is it then a science or a faith? Belief in the theory of evolution is thus exactly parallel to belief in special creation—both are concepts which believers know to be true but neither, up to the present, has been capable of proof" [103].

**Richard Dawkins**

"We are survival machines … robot vehicles blindly programmed to preserve the selfish molecules known as genes. This is a truth which fills me with astonishment".

"Let us try to teach generosity and altruism, because we are born selfish".

"The essence of life is statistical improbability on a colossal scale".

"A universe with a God would look quite different from a universe without one. A physics, a biology, where there is a God, is bound to look different. So the most basic claims of religion are scientific. Religion is a scientific theory" [104].

**Leslie E. Orgel**

"It is extremely improbable that proteins and nucleic acids, both of which are structurally complex, arose spontaneously in the same place at the same time. Yet it also seems impossible to have one without the other. And so at first glance, one might have to conclude that life could never, in fact, have originated by chemical means. We proposed that RNA might well have come first and established what is now called the RNA world … This scenario could have occurred, we noted, if prebiotic RNA had two properties not evident today: a capacity to replicate without

the help of proteins and an ability to catalyze every step of protein synthesis [105, p. 78]. The precise events giving rise to the RNA world remain unclear. As we have seen, investigators have proposed many hypotheses, but evidence in favor of each of them is fragmentary at best. The full details of how the RNA world and life emerged, may not be revealed in the near future" [105, p. 83].

**David E. Green and Robert F. Goldberger**
"The popular conception of primitive cells as the starting point for the origin of the species is really erroneous. There was nothing functionally primitive about such cells. They contained basically the same biochemical equipment as do their modern counterparts. How, then did the precursor cell arise? The only unequivocal rejoinder to this question is that "we do not know"" [106].

**John Maddox**
"It was already clear that the genetic code is not merely an abstraction but the embodiment of life's mechanisms; the consecutive triplets of nucleotides in DNA (called codons) are inherited but they also guide the construction of proteins … So it is disappointing that the origin of the genetic code is still as obscure as the origin of life itself" [107].

**Freeman Dyson** "The more I examine the universe and study the details of its architecture, the more evidence I find that the universe in some sense must have known we were coming" [108].

More quotes and views with discussions are provided in [109–111]. The fight among creationists, intelligent design promoters, and evolutionists is naturally still being continued. Here, it is useful to note and have in our mind *Albert Einstein's* statement:

"*We still do not know one thousand of one percent of what nature has revealed to us*", where "*nature*" obviously includes living and nonliving processes and phenomena.

Also, worth to mention here is the following quotation of *Michel de Montaigue*:

*Nothing is so firmly believed as that which we least know.*

### *1.4.2 Evolution and the Development of Human Society*

As mentioned in Sect. 1.4, modern man appeared about 170,000 years ago with a hunter-gathering societal organization. From that period, human society has passed through three consecutive, largely overlapping, stages of development, namely [45, 46]:

- **Physical:** Survival stage
- **Vital:** Vital needs (trade, etc.)
- **Mental:** Knowledge and empowerment of the individual.

Each stage is distinguished by a predominant organization structure. The following provides a short description of these stages.

**Physical stage** This stage has evolved beginning with man's origin, but in particular has developed over the last 10,000 years. During this period, primitive man was a tribal inhabitant, just starting to develop social organization, using stone-based tools for producing goods for his survival. During the physical stage, people's social activity was confined (compulsorily) within the tribe, and any action outside the borders of the tribe was non-permissible. Natural phenomena (atmospheric and other) caused severe insecurities.

**Vital stage** In this stage, man stared developing the basic productive processes and controlling the physical forces surrounding him. People began to cooperate within the collectives and to be concerned about human issues above the physical ones. Traveling, cooperation, and information exchange between collectives started at this stage, aiming at improving the conditions of life. The vital stage started at around 500 years ago (i.e., approximately during the Renaissance era) and continued over the next centuries of discoveries and explorations.

**Mental stage** This stage started about 170 years ago and achieved an increasingly mature state during the last 60–70 years with the development of the so-called "information and knowledge society". During this period, a systematic codification and organization of past knowledge has had a strong start, and proper training and education methods are being developed and adopted in all areas of human concern (science, arts, technology, politics, humanities, and spiritual fields). Communications have expanded from wire and wireless communications, to computer communications, multimedia, and the Internet. The quality of life is improving, human life-expectancy is increasing, and the values of freedom, choice, and democracy are strongly institutionalized. Overall, the mental stage has presently arrived to very high standards of education, health, social involvement, social benefits, political freedom, quality of life, and an individual person's potential and wish fulfillment. It is believed by many thinkers that the mental stage is now moving towards a spiritual age, where greater efforts are made to achieve unity, peace, social cooperation, spiritual connectedness, and increased opportunities for personal success, delight, and prosperity.

Although the term *human development* **(HD)** most commonly refers to economic advancement, the term applies equally to social, political, and technological progress, all of which contribute to an optimum level of health and well-being. Recent thinkers place emphasis on HD as something distinct and different from economic growth. This distinction helps in establishing proper priorities and strategies of development. A comprehensive development theory should be *human-minded* (or human-centered), recognizing both the fact that a human being is the source and primary motive force for development, and the fact that humans are the rightful beneficiaries of social progress [44].

Development is above the process of social survival, by which a community sustains itself at the minimum level of existence (*basic needs*) without trends for

horizontal expansion or vertical progress. Development is also different from *growth,* which is the expansion or proliferation of activities at any established level of development (physical, vital, mental). Human-development theory is concerned with issues that involve the question whether "*modern societies*" represent "*progress*" over "*traditional societies*". To study this question, the development researchers and scientists go back to the earliest foundations of modernization theory with economies limited to "rural and agricultural levels".

HD is classified as follows [45]:

**Conscious development** HD proceeds from experience to comprehension and conscious understanding of the secrets of successful activities at various levels.

**Natural versus planned development** This separates the natural process of social development from the planned development processes initiated by government.

**Emergence of new social activities** These activities deal with more complex and efficient levels of organization. The development emerges from the subconscious preparedness of society that produces new conscious ideas and initiatives in the individual members. Here education and family help to assimilate these emergent social processes and embed them in the social culture.

Some key issues of HD are the following:

- HD requires energy, surplus, awareness, and ambition.
- HD is performed in several dimensions, viz. social, economic, technological, cultural, and spiritual.
- HD takes place by creating higher levels of organization and knowledge exploitation through the organized use of the best available technology at each time.
- HD is facilitated by higher levels of energy use, efficiency, productivity, creativity, quality, and accomplishment.
- HD can be achieved if the required change of attitudes, social beliefs, and life style are adopted and institutionalized.
- HD is speeded up when the awareness of opportunities spreads, aspiration increases, and infrastructure is utilized.

Going on further to our discussion on HD, we point out that *the creator of all resources is the human mind and intelligence*. HD enables a human to choose his/her priorities, i.e., it is concerned with the "*broadening of human choices*". The three fundamental components of HD that contribute synergetically to the widening of human choices are:

- Socioeconomic development
- Emancipative value change
- Democratization.

**Socioeconomic development** is the most important component of HD and includes among others technological modernization, automation, productivity enhancement, improvement of health and quality of life, rising levels of education,

increases of personal income, etc. This component increases individual resources and therefore provides people with the objective means of choice.

**Emancipative value change** is the second component of HD that contributes to human choice and is compatible with the fact that human choice does not depend only on human resources, but it is also strongly influenced by the individual's motivation and mind. In all cases, this change has the result that conformity values that subordinate human autonomy to community rules tend to be replaced by more emancipative values that are dominated by human choice.

**Democratization** is the most remarkable development of modern society. During the past five or six decades, democratization has occurred in two distinct ways, viz.: (i) many authorization regimes evolved to formal democracies by establishing democratic constitutions, and (ii) most of the existing formal democracies have applied or widened direct democratic institutions leading to rising levels of direct civic participation.

Other components of HD include, but are not limited to, communications, large-scale transportations, cultural progress, and spiritual life. All these components create rising expectations for individual personal progress, which is a fundamental indicator of the maturity of the mental stage of HD. *This is because all development reduces to the development of human beings.* Thorough discussions on all these components of human society's development are provided in [45–47].

The *socioeconomic component* combines social issues in several ways, such as those mentioned before, with economic policies that were developed and adopted in various countries over time. The term "*economics*" comes from the Greek "οικονομικός" (economicos = cheap), which is a good practice in household-budget management. The modern field of economics was initiated in the 17th and 18th centuries as the Western world started its gradual evolution from an agrarian to an industrial society.

The fundamental economic problems faced in all epochs have been the same, namely:

- Given that our resources are limited, how do we choose what to produce?
- What are the best and stable prices, and how do we guarantee the full usage of our resources?
- What policies must be followed to achieve a rising standard of life both for ourselves and future generations?

The study of these questions has led over time to the development of several distinct *economic theories* and *models* with relative merits and drawbacks [58–69, 113–116]. These are the following.

**Mercantilism** was the model used by merchants during the 16th and 17th centuries. This theory is based on the assumption that a nation's wealth was mainly the outcome of gold and silver accumulation. Mercantilists advocated the enforcement of state prices on foreign goods to restrict import trade.

**Physiocrats** developed (in France during the 18th century) the model of a circular flow of income and output. Physiocrats had the belief that agriculture was the unique source of economic wealth, in contrast to mercantilists, who supported enhancing trade at the expense of agriculture. Physiocrats endorsed the principle of *minimal government interference* in the economy (in French "*laissez-faire*").

**Classical Economics** This is widely considered as the first modern school of economic thought. It began in 1776 with *Adam Smith's* seminal work "*The Wealth of Nations*", for which he is recognized as the father of "*free market economics*". In Smith's theory, the ideal economy is a self-regulating market system that automatically satisfies the economic needs of the people. Other major developers of classical economics include *Jean-Baptiste Say, David Ricardo, Thomas Malthus,* and *John Stuart Mill*. Smith incorporated in his theory some of the physiocrats concepts (including "laissez faire"), but did not accept the position that only agriculture was productive. David Ricardo focused on the distribution of income among landowners, workers, and capitalists, in contrast to Smith who placed emphasis to the production of income. Thomas Malthus used the idea of diminishing returns to explain low living standards, and challenged the idea that a market economy tends to produce full employment. He attributed unemployment to the tendency of economy to reduce its spending by over-saving.

**Marginal Economics** While classical economics is based on the assumption that prices are determined by the cost of production, *marginal economists* gave emphasis to the fact that the prices also depend upon the level of demand, which varies according to the degree that consumers are satisfied by the offered goods and services.

**Marxist School** This school challenged the principle of classical economic theory. Actually the term "*classical economics*" is due to *Karl Marx* and refers to Ricardo's and Mill's theory of economics. This term is currently used for all economic theories followed Ricardian economics. *Marx* and *Engel* developed the so-called "*historian materialism*" [63, 64]. According to this theory, manifestations of existence and evolution were based on the necessity of labor and production. Specifically, this theory asserts that as long as the labor mode (breeding, cultivation, handicrafts) were just permitting the production of the absolutely necessary means for the existence of the particular worker, there was not the need to exploit the work of others. But as soon as man, due to the growth of productive forces and technology, was able to produce a *surplus* beyond the minimum requirements, this surplus was usually taken by other people. This process has passed through several stages, i.e., *Slavery, Feudalism,* and *Capitalism*. Marx predicted that capitalism would create misery for workers, since the struggle for profit would lead the capitalists to apply labor-saving processes, producing a "*reserve army of the unemployed*" who would eventually rebel and seize the means of production.

**Institutionalist Economics** This economic model, also known as "*Institutionalist political economy*", considers the individual economic process to be a part of the larger social process which is affected by currently established views of thought and

life styles. Therefore, institutionalist economists reject the narrow classical economic model that assumes that people are primarily driven by economic self-interest. In contrast to the "minimal government interference" (laissez-faire) of physiocrats, the institutionalist economists propose that government control and social reform are necessary for bringing about as much as possible an equal distribution of income. The institutionalist political economy model has its roots to *Thorsten Veblen's* view that there is a separation (dichotomy) between technology and the operational aspects of society. New variants of institutional economics give emphasis on the role of institutions in minimizing the economic transactions costs and regard markets as an outcome of the complex interaction between the various institutions: individuals, companies, government, and social norms. The role of *law* in economics was of major concern since 1924, the year in which *John R. Commons* published his work "*Legal Foundations of Capitalism*". Institutional economics has also broadened the area of economics by including cognitive, psychological, and behavioral issues, in addition to the standard technocratic assumptions about economic activity.

**Keynesian Economic Theory** This theory is opposite to the line of thought of classical theory according to which, in a *recession*, wages and prices would be *decreased* to restore full employment. Keynes, in his 1936 publication "*Theory of Employment, Interest and Money*", argued that falling wages, by depressing people's incomes would not allow a revival of spending. His position was that direct government intervention was necessary to increase total spending. He reasoned theoretically that the use of government spending and taxation can stabilize the economy. Specifically, government must spend and decrease taxes when private spending is insufficient and threatens a *recession*, and must reduce spending and increase taxes when private spending is too high and threatens *inflation*.

**Monetarists Economic Theory** This theory was developed in the 1970s by *Milton Friedman* who challenged the Keynesian conservative economic theory [113–116]. According to Friedman: "*inflation is always and everywhere a monetary phenomenon*". He accepted the Keynesian definition of recession, but rejected the therapy policy. He reasoned that government should maintain the money supply, increasing it slightly each year to enable the natural growth of economy. The control policies that can be used for this purpose include:

- Open-market operations
- Funding
- Monetary-base control
- Interest rate control.

Friedman coined a new definition of money as "*monetary aggregates*" that include virtually everything in the financial sector (saving deposits, money market accounts, etc.), which is different from the definition used by most economists (i.e., cash circulation and its close equivalents such as checking accounts). Friedman argued that an excessive increase in the money supply leads inherently to inflation, and that monetary authorities should solely maintain price stability. He derived a

quantified monetary rule (known as *Friedman's k-percent rule*) by which the money supply can be determined computationally, helping economic bodies to enact their best monetary policy decisions. Monetarism was most popular in the 1970s and 80s.

Two theories that historically have strongly challenged the theory of "*free market and globalization*" are the *dependency theory* [70–72], and the w*orld-system theory* [73–77]. Dependency theory (first formulated in the 1950s), argues that low levels of development in underdeveloped countries (called the *peripheral countries*) spring from their dependence on the advanced economies (of the *rich* or *core countries*), which are further enriched at the expense of the peripheral countries own wealth. The world-system theory was developed by *Immanuel Wallerstein* and lies somewhere between the theories of Weber and Marx [75–77].

A recent book that explains that "*capitalist freedom*" is a two-edged sword, is "Capitalism and Freedom: The Contradictory Character of Globalization" of *P. Nolan* [78]. He states that, although many benefits of capitalism and globalization are visible when compared to the economic organization of non-capitalist systems, in our times we experience severe undesirable results for both the natural environment and the population. Actually, as P. Nolan argues, capitalist globalization's contradictions have intensified. It is generally recognized that they visibly threaten the environment and widen global inequality, within both rich and poor countries, and between the internationalized global power elite and the mass of citizens rooted within the respective nation. The author explores comprehensively the impact of the globalized economic phenomenon on individual and social liberties

**Human Development Ranking** The major current measure for ranking human well-being and development in various countries worldwide is the so-called *human development index* **(HDI)** which is a quantitative combination of three particular indicators namely: (i) life expectancy at birth, (ii) adult literacy rate, and (iii) gross domestic product (GDP) that measures the income per capita [79, 80]. In 1990, the *United Nations Development Program* **(UNDP)** launched the *human development report* **(HDR),** which is published in more than 12 languages and takes into account, in addition to the above three components of HDI, the *Gender-related Development Index* **(GDI)**, the *Gender Empowerment Measure* **(GEM),** and the *Human Poverty Index* **(HPI)**. Each year, HDP is devoted primarily to one or more distinct challenges facing humanity. For example, the 2007–2008 HDR was primarily focused on climate change, and the 2009 HDR on the migration process and its effects on growth and health. Countries with HDIs below 0.5 are classified as "low-development" countries, and countries with an HDI $\geq$ 0.8 as "high development" countries. Details about HDI and HDR are provided in [80, 81].

## 1.5 Fundamental Elements of Some Specific Societal Aspects

As mentioned in Sect. 1.1, the term "*pillar*" is used for a fundamental element that provides a "*firm support*" in a variety of natural processes, and human activities, and creatures (material, cultural, and social). For completeness in this section, we provide a brief outline of some human-society "fundamental elements" that refer to processes beyond the ones extensively covered in this book. These are the following:

- Pillars of democracy
- Pillars of fulfilled living
- Pillars of sustainable development

### *1.5.1 Pillars of Democracy*

The term "*democracy*" comes from the Greek word "*δημοκρατία*" (demokratia: *δήμος* (demos) = public (people) + *κράτος* (kratos) = rule/power/state). The rules and procedures of democracy go back to *Athenian Democracy* (508–322 BC) in which the founder of democracy was *Pericles* [82].

One of the basic principles of Athenian Democracy ("*freedom-of-speech*" principle) is expressed by the following question addressed to the citizens in "*public forums*" and "*general assemblies*":

"*Τις αγορεύειν βούλεται;*"

"*Who wants to make a speech?*"

Today, there are many variations of democracy all over the world, and it is generally accepted that the "*pillars*" of a sound institutionalized democracy include at least the following "*musts*" [83, 84]:

- **Pillar 1—Elections:** The elections must be free and fair. No one individual or group can monopolize power over the election process.
- **Pillar 2—Political Tolerance:** Minorities must benefit equitably from the election process. The minority's civil rights must be respected, tolerated, and protected in order to secure a sustainable democracy.
- **Pillar 3—The Rule of Law:** Political action must be subject to democratic laws and take place within the bounds of a proper judiciary and regulatory framework.
- **Pillar 4—Free Press and Expression:** The nature and extent of the democratic political system are indicated by what the citizens are permitted to say, print, discuss and publicize. A free press is *a must* and provides a measure of the freedom of expression in the society.

- **Pillar 5—Accountability and Transparency:** A democratic government must be accountable and its actions must be transparent. The same is true for individual institutions. The goal of accountability and transparency is to protect the citizens from misguided policies and practices that enrich a few at the expense of the many.
- **Pillar 6—Decentralization:** Government must be close to the people governed and funding and resources must be decentralized. The local communities provide good ways to see how democracy is linked to the everyday lives of the people.
- **Pillar 7—Civil Society:** A healthy democracy provides strong vitality to its civil society. Citizens must have a role to play in participating in the public-policy making and checking the government's decision making. Forums, clubs, charities, professional unions, think tanks and other special-purpose societies are included in the civic-society umbrella.

In general, democracy can be compressed into the following rule: "*government of the people, by the people and for the people*". Unfortunately, this is not always strictly and fully followed, despite contradictory declarations by politicians and governors.

### *1.5.2 Pillars of Fulfilled Living*

*Nature* is governed by *unchanging laws* that are continuously working in our lives no matter if we know and understand them or not. In the *Freedom Technology's* site (London) [86] the following five pillars of fulfilled living are discussed:

**Pillar 1—Freedom, sovereignty and privacy:** This pillar includes processes and policies that protect and maintain what rightfully belongs to the individual.
**Pillar 2—Good bodily health:** A person is free from pain and discomfort independently of any controlling medical treatment.
**Pillar 3—Relationship principles:** Skills and practices that secure good relationships with other people and life-long love.
**Pillar 4—Absolute truths:** They help to promote true prosperity from within in agreement with the known laws of nature, as proven by reality and not simply based on traditions.
**Pillar 5—Wealth creation and financial freedom:** This can be achieved by encouraging the individual to exert his/her own skills and use profitably his/her own time, thus personally escaping the constraints of a salaried position.

### 1.5.3 Pillars of Sustainable Development

Sustainability is, according to *Robert Gilman*, "the ability of a society, ecosystem, or any similar ongoing system, to continue operating in the indefinite future without being forced into decline via exhaustion of key resources" [44]. Sustainable development usually refers to ecological sustainability, but currently other forms such as economic, societal, and cultural sustainability are noticeably entering the scene. The three pillars of sustainable development most-commonly considered are [44]:

- Economic growth
- Social progress
- Environmental protection

In [87], *Keith Nurse,* prepared a report for the *Commonwealth Secretariat,* in which he adds "*culture*" as the fourth pillar of sustainable development. Here, we will briefly describe all these pillars of sustainable development.

**Pillar 1—Economic growth:** According to Munro (1995) [87], this must be compatible with the need to strike a balance between the costs and benefits of economic activity, within the confines of the carrying capacity of the environment. Economic growth should not be achieved at the expense of intergenerational equity.
**Pillar 2—Social progress:** Social sustainability is achieved by respecting social values and norms, which are largely intangible "*ethical*" values and relate to language, education, work styles, and spiritual issues. Social progress assumes, as a prerequisite, the satisfaction of basic needs within the society (food, clothing, shelter), and equity in the distribution of resources.
**Pillar 3—Environmental protection:** The protection and preservation of the quality of the environment is the cornerstone of sustainable development. International concerns about the environment are evidenced by the UN series of summits and declarations started in 1992 (Rio de Janeiro), and continued with the Copenhagen Declaration (1995), the Kyoto Protocol (COP3, 1997) {http://unfcc.int/kyoto_protocol/items/2830.php}, the 11th UNCCC (Climate Change Conference) at Bali (India, 2007), and the UNCC (COP15) of Copenhagen (December 2009). General information about on-going events about environmental protection is provided in [88].
**Pillar 4—Culture:** Often, culture is considered narrowly, not linked to the wider sustainable-development debate. The connection of culture to development was systematically pointed out by *Serageldin* and *Martin-Brown* (1999) [87]. Today the concept of sustainable development has matured and embodies culture as one of its key elements. *Wallerstein* (1991) states that the possibilities for an ecologically sustainable future depend on how "*production cultures*" and "*consumption cultures*" are adapting to ecological, socio-political, and technological changes. Today it is understood that the cultural sector plays a dual role, i.e., it is an arena for social identity formation and an arena for economic growth potential on the basis of intelligent and technological properties and capabilities.

Closing our discussion on the *societal pillars,* we just list the pillars of the following aspects:

- *Pillars of dynamic schools* (communication and relationship, leadership and empowerment, planning and evaluation, collaboration, accountability and responsibility, consistency, and redundancy) [89].
- *Pillars of prosperity* (energy, economy, integrity, sympathy, sincerity, impartiality, and self-reliance) [90].
- *Pillars of market leaders* (vision-directed, customer-driven, innovative, flexible/adaptive, intellectual-capital oriented, and knowledge-based) [91].
- *Pillars of science* (*matter–quantum theory, life, the mind-intelligence*) [67].
- *Pillars of instructional technology* (*philosophy of technology, history of technology, and technological leadership*) [118].
- *Pillars of education* (*learning to live together, learning how to acquire knowledge, learning how to act, and learning for life*) [85].

Worth mentioning is also the working paper of the American Economic Foundation **(AEF)**: "*The Ten Pillars of Economic Wisdom*" [92].

## 1.6 The Five Fundamental Elements of This Book

The five fundamental elements of life and human society that are the subject matter of this book are:

- Energy

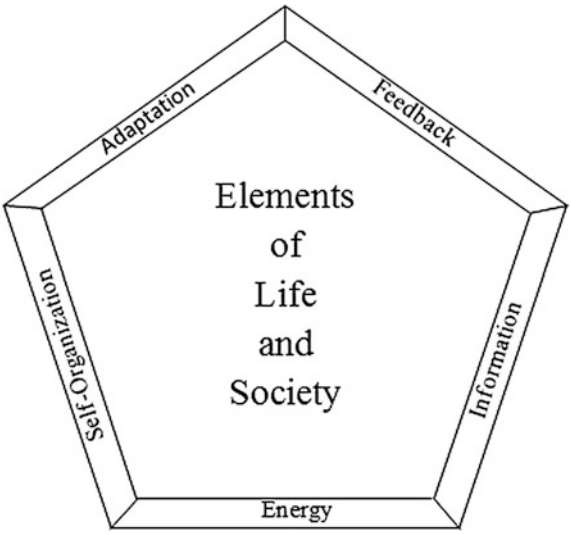



**Fig. 1.10** The pentagon representation of LHS. The energy element is the basic prerequisite of all the other elements

**Fig. 1.11** Four energy sources: the Sun (the source of life on Earth), solar energy system, electric energy, and wind energy

- Information
- Feedback
- Adaptation
- Self-Organization

These elements/pillars support both the life and the society and have also a technological content. From the seven biological pillars of life suggested by Koshland (see Sect. 1.2.5) *compartmentalization, regeneration,* and *seclusion* pillars are biological characteristics of life only, and do not directly apply to society. The *program* can be considered to be involved in the *information* pillar, and *improvisation* is included in *adaptation*. Likewise, the pillars of *democracy,*

**Fig. 1.12** Four examples for information: computer, computer network, DNA, and the human brain


**Fig. 1.13** The feedback concept

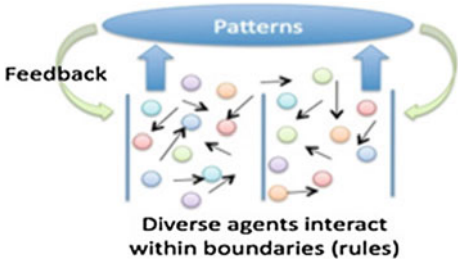



**Fig. 1.14** Representation of adaptation

*education, fulfilled living, sustainable development, prosperity,* etc., discussed in Sect. 1.5 are fundamental elements (characteristics) of society that do not have a biological nature or interpretation. The full gamut of fundamental elements of life

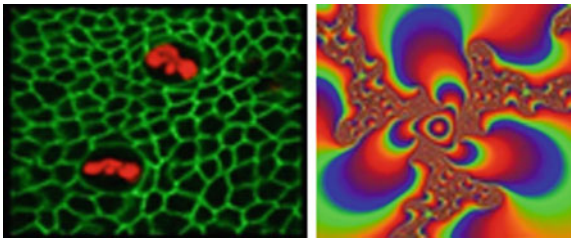


**Fig. 1.15** Two instances of self-organization

and human society involves the union of the *biological* and *societal* aspects. Therefore, considering "*life and human society*" as a "*Temple*", like Koshland, we can use a temple that has five central columns accompanied by a left-hand, side column standing for the pure biological elements and a right-hand side column that stands for the pure societal (humanistic, economic, cognitive, cultural, and functional) aspects. Here we represent the five common (central) elements of life and society by a pentagon as shown in Fig. 1.10, with energy as its base.

In this book, these five *central* (common) *fundamental elements/pillars* of LHS will be considered with emphasis on their diachronic technological issues. Chapters 1–9 discuss the conceptual and methodological aspects of the five elements, including the historical landmarks of their evolution and development. Chapters 10–13 discuss the connection, role, and impact of these elements to the life and human society, including evolutionary aspects and modern applications.

Figure 1.11 shows four sources/types of energy, Fig. 1.12 shows four pictures for information, and Figs. 1.13, 1.14 and 1.15 show pictures for feedback, adaptation, and self-organization (Figure 1.14).

## 1.7 Concluding Remarks

In this introductory chapter, the question "*what is life*" was addressed and the basic concepts of "*cell biology*" (living cell, nucleotides, DNA/RNA, proteins) were discussed. Then, the meaning of society was analyzed. The evolution of life and human society was presented in some detail, also including a number of views (expressed as quotations) about the origin of life and evolution that support or criticize Darwin's theory. Next, the fundamental elements/pillars of a number of specific societal aspects were briefly outlined. These elements have a "*pure societal*" nature and concern the aspects of democracy, education, fulfilled living, and sustainable development. The five fundamental elements: energy, information, feedback, adaptation, and self-organization, which are the subject matter of this book, are actually common elements of life and society, and as such they have been studied in "*life sciences*" and "*social sciences*", and "*engineering sciences*", often from different points of view.

In this reference book, we will present the underlying concepts, principles, methods, and applications of these five fundamental elements, in a collective way, aiming at convincingly demonstrating their roles (individual and combined), and their impact on the existence, evolution, operation, maintenance, and sustainability of life and modern human society.

## References


1. D.E. Koshland, Jr., The seven pillars of life (Special Essay). Science **295**, 2215–2216 (March, 2002) (www.sciencemag.org)
2. Meaning of Pillar. http://ardictionary.com/pillar/6196
3. Definition of Pillar. http://dictionary.net/pillar
4. Biology Online. http://www.biology-online.org
5. L. Pellegrini, The DNA Page. http://home.comcast.net/~llpelegrini/dna.html
6. DNA and RNA Structures. http://www.whatislife.com/reader/dna-rna/dna-rna.html
7. C.E. Ophardt, *DNA and RNA Introduction* (Virtual Chembook, Elmhurst College, 2003). http://www.elmhurst.edu/~chm/vchembook/580DNA.html
8. J.E. Wampler, Tutorial on Peptide and Protein Structure, 1996. http://www.bmb.uga.edu/wampler/tutorial/
9. J. Abrams, DNA, RNA and Protein: Life at its Simplest. http://www.postmodern.com/~jka/rnaworld/nfrna/nf-rnadefed.html
10. The University of Arizona, The Biology Project, January 2003. http://biology.arizona.edu (Large Molecules Problem Set)
11. J. Hadfield, DNA Structure Microbe Library. http://www.microbelibrary.org/ASMOnly/Details.asp?ID=432
12. J. Abrams, RNA World: History, an Idea of History. http://www.postmodern.com/~jka/rnaworld/rnaworld/nfrna/nf-rnaworldded.html
13. N. R. Pace, B.C. Thomas, C.R. Woese, Probing RNA structure, function and history by comparative analysis, In: *The RNA World* (Chap. 4) (Cold Spring Harbor Laboratory Press, New York, 1999)
14. J.A. Witkowski, *The Inside Story: DNA to RNA to Protein* (Cold Spring Harbor Laboratory Press, New York, 2005)
15. J.N. Davidson, Nucleic acids: the first hundred years, *Biochem. Biophys. Acta,* 17–33 (1989)
16. A. Rich, The nucleic acids: a backward glance, Prog. Nucleic Acid Res. Biol., 3–6 (1968)
17. A. Ralston, K. Shaw, mRNA: history of functional investigation. Nat. Educ. 1(1) (2008). http://www.nature.com/suitable/topicpage/mrna-history-of-functional-investigation.html
18. http://barleyworld.org/css430_09/lecture%209-09/figure-09-03.JPG
19. H. Lodish, A. Berk, S. Lawrence Zipurski, P. Matsudaira, D. Baltimore, J, Darnell, *Molecular Cell Biology*, 5th edn. (W.H. Freeman, New York, 2003)
20. L.A. Allison, *Fundamental Molecular Biology* (Wiley-Blackwell, New York, 2007)
21. Russel Kightley Media, Animal Cell Diagram with Labels. http://www.rkm.com.au/CELL/animalcell.html
22. Cells Alive, Cell Models: An Interactive Animation http://www.cellsalive.com/cells/cell_model.htm
23. Define Society at Dictionary.com http://dictionary.reference.com/browse/society?qsrc=2446
24. R. Jenkins, *Foundations of Sociology* (Palgrave MacMillan, London, 2002)
25. R.M. Alston, *The Individual vs. the Public Interest: Political Ideology and National Forest Policy* (West View Press, Boulder, CO, 1983), pp. 37–38 and 187–190, (see: http://faculty.weber.edu/ralston/Home%20Page/1740lec4.htm)

26. Studies in the Theory of Human Society (Giddings) http://www.britannica.com/EBchecked/topic/569932/Studies-in-the-Theory-of-Human-Society
27. K. Davis, *Human Society* (MacMillan Company, New York, 1949), Questia Online Library: http://www.questia.com/library/book/human-society-by-kingsleyDavis
28. F.W. Blackman, *History of Human Society,* 1926 (Free e-Book) http://manybooks.net/titles/blackmanf3061030610-8.html
29. Science for All Americans Online: *Human Society* (Chap. 7) http://www.project2061.org/publications/sfaa/online.chap7.htm
30. B. Russel, *Human Society in Ethics and Politics* (Routledge Classics, New York, 2010)
31. C. Lévi-Strauss, *The Savage Mind* (*Nature of Human Society*) (The University of Chicago Press, Chicago, 1966)
32. Definition of Prokaryote, Your Dictionary.com http://www.yourdictionary.com/prokaryote
33. H.G. Trüper, How to name a prokaryote? Etymological considerations, proposals and practical advice in prokaryote nomenclature. FEMS Microbiol. Rev. **23**(2), 231–249 (1999)
34. *Origin and Definition of Life,* Physical Geography.net, e-book (Chap. 9). http://www.physicalgeography.net/fundamentals/9a.html
35. E.-A. Viau, *The Three Domains*, http://www.world-builders.org/lessons/less/les4/domains.html
36. The Origin and Evolution of Life: A Product of Cosmic, Planetary, and Biological Processes, *NASA's Planetary Biology Program*. http://cmex.ihmc.us/VikingCD/Puzzle/Evolife.htm
37. The Evolution of Life, School of Science and Engineering, The University of Waikata. http://sci.waikata.ac.nz/evolution/HumanEvolution.shtml
38. D. Sadava, *Life Study Guide: The Science of Biology,* 9th edn. (W.H. Freeman, New York, 2009)
39. W.M. Becker, L.J. Kleinsmith, J. Hardin, G.P. Berton, *The World of the Cell,* 7th edn. (Benjamin Cummings, New York, 2008
40. W.K. Purves, D. Sadava, G.H. Orians, C. Heller, *Life: The Science of Biology*, 7th edn. (Sinauer Associates and W.H. Freeman, New York, 2004
41. W.K. Purves, *GET it Cell Biology,* CDROM (Mona Group LLC, Sunderland, MA, 1998)
42. E. Viau, *Life Emerges on Your World,* World Builders (Lesson 4) http:www.world-builders.org/lessons/less4/les4/Vles4.html
43. Biological Evolution (*About.com Biology*). http://biology.about.com/od/evolution/a/aa110207a.htm
44. S.G. Tzafestas, *Human and Nature Minding Automation* (Springer, Dordrecht/Heidelberg, 2010)
45. R. Posner, The Four Stages of Society's Evolution, Growth Online. http://www.gurusoftware.com/gurunet/Social/Topics/FourStages.htm
46. G. Jacobs, R. MacFarlane, N. Asokan, *Comprehensive Theory of Social Development,* Intl. Center for Place and Development. http://www.motherservice.org/Essays/ComprehensiveTheoryofSocialDevelopment.html
47. C. Morris, I. Adelman, *Comparative Patterns of Economic Development 1850–1914* (John Hopkins University Press, Baltimore, 1988)
48. Evolution of Human Societies (Chap. 3). http://www2.fiu.edu/~grenierg/chapter3.html
49. Principia Cybernetica Web: Social Evolution. http://pespmc1.vub.ac.be/socevol.html
50. H. Raven, T. Williams, (eds.) Nature and Human Society: The Quest of a Sustainable World, in *Proceedings of 1997 Forum on Biodiversity,* National Academy Press, Washington, DC, 1997
51. R. Nunes, The Evolution of Society: A Comprehensive View: Workers Party (NZ). http://workersparty.org.nz/resources/study-material/the-evolution-of-society
52. C. Welzel, R. Inglehart, Human Development and the Explosion of Democracy, WZB Discussion Paper FSII 01-202 (WZB, Berlin, 2001)
53. C. Welzel, R. Inglehart, The theory of human development: a cross-cultural analysis, Eur. J. Pol. Res.**42**, 341–379 (2003)

54. R.J. Estes, Trends in world social development. J. Dev. Soc. **14**(11), 11–39 (1998)
55. Human growth and development: courses on theoretical perspectives. http://www.unm.edu/~jka/courses/archive/theory1.html
56. J.D. Nagle, A. Mahr, *Democracy and Democratization* (Sage Publications, London, 1999)
57. R. Inglehart, W.E. Baker, Modernization, cultural change and the persistence of traditional values, Amer. Sociol. Rev. **65**, 19–51 (2000)
58. Classical economics, http://en.wikipedia.org/wiki/classical_economics
59. J. Buchan, *The Authentic Adam Smith: His Life and Ideas* (W.W. Norton & Company, New York, 2006)
60. G. Stolyarov II, *The Dynamic Process of Market Forces in Free Market Economic Theory* (Hellium Inc, 2002–2010). http://www.helium-com/items/112757-the-dynamic-process-of-market-economic-theory
61. A. Smith, *The Wealth of Nations* (University of Chicago Press, Chicago, 1977)
62. http://www.frbs.org/publications/education/greateconomists/grtschls.html
63. D. Kellner, *Engels, Modernity and Classical Social Theory* http://www.gscis.com/ucla.edu/faculty/kellner/essays/engelsmodernclassicalsocialtheore.pdf
64. K. Marx, *A Contribution to the Critique of Political Economy,* English Translation ed. by M. Dobb (Progress Publishers, London, 1979)
65. O. Ruhle, *Karl Marx: His Life and Works* (The Viking Press, New York, 1943
66. Historical Materialism Design: Modern, Mid-Century, Industrial Age http://historicalmaterialism.com
67. Historical Materialism, http://en.wikipedia.org/wiki/Historical_materialism
68. Friedrich Engels, http://en.wikipedia.org/wiki/Friedrich_Engels
69. Classical Economists http://en.wikipedia.org/wiki/Classical_economics
70. J.D. Cockroft, A.-G. Frank, D. Johnson (eds.), *Dependence and Underdevelopment* (Anchor Books, New York, 1972)
71. V. Ferraro, Dependency Theory: An Introduction. http://www.mtholyoke.edu/acad/intrel/depend.htm
72. T. Spybey, *Social Change, Development and Dependency: Modernity, Colonialism and the Development of the West* (Policy Press, Oxford, 1992)
73. D. Chirot, *Social Change in the Modern Era* (Harcourt Brace Jovanovich, New York, 1986)
74. D. Chirot, D.T. Hall, World system theory. Annu. Rev. Soc. **8**, 81–106 (1982)
75. I. Wallerstein, *The Modern World System I: Capitalist Agriculture and the Origins of the European World-Economy in the Sixteenth Century* (Academic Press, New York, 1974)
76. T.K. Hopkins, I. Wallerstein, *Processes of the World System* (SAGE Publications, Beverly Hills, CA, 1980)
77. I. Wallerstein, *The Essential Wallerstein* (The New York Press, New York, 2000)
78. P. Nolan, *Capitalism and Freedom: The Contradictory Character of Globalisation* (Authem Press, London, 2007)
79. M. Engineer, I. King, N. Roy, The human development index as a criterion for optimal planning. Indian Growth Dev. Rev. **1**, 172–192 (2008)
80. http://en.wikipedia.org/wiki/Human_Development_Index
81. http://hdr.undp.org/en/humander/
82. C.W. Blackwell, Athenian democracy: a brief overview, in: *Athenian Law in its Democratic Context*, ed. by A. Lanni (Center for Hellenic Studies On-Line Discussion Series Stoa, February, 2003). (http://www.stoa.org)
83. A. Panyarachum, *Building the Pillars of Democracy* (Center for International Private Enterprise (CIPE), August, 2008), pp. 1–8 (http://www.cipe.org/blog)
84. L. Diamond, *The Spirit of Democracy* (Henry Holt and Company, New York, 2008)
85. Learning: *The Treasure Within,* UNESCO Report on Education for the 21st Century, German UNESCO Commission Publication, Neuwied/Berlin, pp. 18–19, 1997 (see: http://www.dadalos.org/politik_int/bildung/saeulen.htm)
86. The Five Pillars of a Fulfilled Life, Freedom Technology. http://freedomtechnology.org/pillars.htm

87. K. Nurse, Culture as the Fourth Pillar of Sustainable Development, Commonwealth Secretariat, Malborough House, London, U.K., 2002. Also: Intl. Meeting for Small Island Developing States (SIDS), Barbados '10
88. http://www.environmental-expert.com
89. S.W. Edwards, P.E. Chapman, *Six Pillars of Dynamic Schools* (Educational Research Service (ERS), Alexandria, VA, 2009)
90. J. Allen, Eight Pillars of Prosperity, The James Allen Free Library, http://james-allen.in1woord.nl/?text=eight-pillars-of-prosperity
91. F. El-Nadi, The Six Pillars of Market Leaders. http://www.evancarmichael.com/Human-Resources/840/The-Six-Pillars-of-Market-Leaders.htm
92. The Ten Pillars of Economic Wisdom, American Economic Foundation. http://www.worldvieweyes.org/resources/Strauss/TenPillarsEconWisd.html
93. H.S. Lipson, A physicist looks at evolution. Phys. Bull. **31**, 138 (1980)
94. F. Hoyle, *The Intelligent Universe* (Holt, Rinehart & Winston, New York, 1983)
95. D.B. Kitts, Paleontology and evolutionary theory. Evolution **28**, 466 (1974)
96. (a) M. Behe, *Darwin's Black Box: The Biochemical Challenge to Evolution* (Free Press, New York/London, 1996). (b) M. Beche, *The Edge of Evolution* (Free Press, New York/London, 2007)
97. H.P. Yockey, A calculation of the probability of spontaneous biogenesis by information theory. J. Theor. Biol. **67**(396) (1977)
98. L.D. Sunderland, *Darwin's Enigma* (Master Books, San Diego, USA, 1984), p. 89
99. K. Dose, The origin of life: more questions than answers. Interdisc. Sci. Rev. **13**(4), 348–356 (1988)
100. F. Dyson, Honoring Dirac. Science **185**, 1160–1161 (1974)
101. H.P. Yockey, *Information Theory and Molecular Biology* (Cambridge University Press, Cambridge, 1992)
102. H.P. Yockey, *Information Theory, Evolution, and the Origin of Life* (Cambridge University Press, Cambridge, 2005)
103. L.H. Matthews, *Introduction to Darwin's "The Origin of Species"* (J.M. Dent & Sons Ltd, London, 1971)
104. Quotes from: http://www.brainquote.com/quotes/authors/r/richard_dawkins.html
105. L.E. Orgel, The origin of life on earth. Sci. Am. **271**, 77–83 (1994)
106. D.E. Green, R.F. Goldberger, *Molecular Insights into the Living Process* (Academic Press, New York, 1967)
107. J. Maddox, The genesis code by numbers. Nature **367**, 111 (1994)
108. F. Dyson, *Disturbing the Universe* (Harper and Row, New York/San Francisco, 1979)
109. Scientific Reality vs. Intelligent Design. http://www.idnet.com.au/files/pdf/Doubting%20Yockey.pdf
110. Evolutionists Quotes: Life from Non-Life http://www.strengthsandweakeness.org/evol_quotes.htm
111. Evolution/Creation Quotes http://www.cft.org.za/articles/evquote.htm
112. Quotations on the Origin of Life and Evolution http://www.ucc.ie/academic/undersci/pages/quotes_origin_evolution.htm
113. M. Friedman, A theoretical framework for monetary analysis. J. Polit. Econ. **78**(2), 210 (1970)
114. F. Friedman, A.J. Schwartz, *Money in Historical Perspective* (University of Chicago Press, Chicago, 1987)
115. K. Brunner, A.H. Meltzer, *Money and the Economy: Issues in Monetary Analysis* (Cambridge University Press, Cambridge, U.K., 1993)
116. P. Krugman, *Peddling Prosperity* (W.W. Norton & Co., New York, 1994)
117. M. Kaku (ed.), *Visions: How Science Will Revolutionize the 21st Century* (Anchor Books, New York, 1997)
118. L.A. Tomei, The pillars of instructional technology, in *Technology Literacy Applications in Learning Environments*, ed. by D. Carbonara (Information Science Publishing, Hershey/London, 2005), pp. 1–13
119. S.J. Gould, Evolution's eratic pace. National History **36**(5), 12–16 (1977)

# Chapter 2
# Energy I: General Issues

*The history of man is dominated by,*
*and reflects, the amount of available energy.*
Frederick Soddy
*The scientist discovers a new type of material or energy*
*and the engineer discovers a new use of it.*
Gordon Lindsay Glegg

**Abstract** Energy is the basis of everything. It is the dominant fundamental element of life and society. Its movement or transformation is always followed by a certain event, phenomenon, or dynamic process. Energy is used by humans to acquire useful minerals from earth, and construct technological creatures (buildings, transportation systems, factories, machines, etc). The energy used by end users in our society comes from exhaustible sources (coal, fuel oil, natural gas), non-exhaustible (renewable) sources (hydroelectric, wind, solar) or from alternative sources (bio-alcohol, biodiesel, liquid nitrogen, hydrogen). In this chapter, we provide a historical tour to the energy and thermodynamics studies and developments, accompanied by an exposition of the fundamental aspects of energy. These aspects include the energy concept itself, the energy types, the energy sources, and the impact of energy generation and use on the environment.

**Keywords** Energy · Energy types · Kinetic energy · Potential energy
Evidence of energy · Available energy · Sensible energy · Latent energy
Chemical energy · Nuclear energy · Energy sources · Exhaustible/renewable
energy sources · Reversible/irreversible process · Fossil/non-fossil fuels

## 2.1 Introduction

*Energy* is the most fundamental prerequisite for all living organisms on Earth and engineered (man-made) systems to live, operate, and act. It is one of the most important physical concepts discovered by human. In the elementary textbooks, energy is defined as the "*ability to do work*". To do all things we do, we need energy. More generally, things can change because of energy. For example, by



S. G. Tzafestas, *Energy, Information, Feedback, Adaptation, and Self-organization*,
Intelligent Systems, Control and Automation: Science and Engineering 90,
https://doi.org/10.1007/978-3-319-66999-1_2

taking concentrated energy in the form of oxygen plus food, we can perform both the unconscious synthesis of the complex biological substances required for our bodies and our conscious physical and mental work, returning to nature-diffused energy as body heat and less-concentrated-energy substances.

Using energy, we can acquire from the Earth useful minerals, construct powerful complex machines, etc. The energy used by end users in our society (car manufacturers, wind turbine makers, dairy farmers, and so on) comes from fossil sources (coal, fuel oil, and natural gas), and electrical energy generated using fossil, nuclear fuel, and renewable-energy sources (wood, hydroelectric, wind, solar, etc.). Humans need energy to walk, to run, to read a book, to think, and even to sleep, and so on. Nearly all buildings (homes, offices, etc.) need energy for lighting, air conditioning, water heating, space heating, and lift systems. To run our office equipment (computers, printers, fax machines, etc.), we need energy. Energy is used to light our cities, power our vehicles, industrial plants, trains, airplanes, etc. In general, we buy energy, we sell energy, try to conserve energy, convert energy, and so on. Energy can be transferred from one place to another and can be transformed from one form to another. Every time energy moves or changes form, a certain event takes place (e.g., something moves, gets warmer or cooler, breaks, falls down, and so on). Therefore, to understand the processes that occur in natural and man-made systems, we must study energy's behavior, i.e., what happens when energy moves or changes form.

This chapter gives a brief exposition of the basic issues of energy, namely the energy concept itself, energy types, energy sources, and the impact of energy on the environment, including a tour of the principal historical landmarks of energy and thermodynamics.

## 2.2 What is Energy?

Energy is a physical concept that cannot be defined in the usual concrete way. Actually, most of the authors in physics, energy, and thermodynamics bypass the definition and describe energy through the physical and mathematical properties of its various manifestations [1–8]. The term, "*energy*" comes from the Greek word "*ενέργεια: energeia*" (action, activity, operation) and "*ενεργός: energos*" (active, working). Aristotle used the term "*ενέργεια*" to clarify the definition of "*being*" as "*potency*" (δύναμις-dynamis: force) and "*action*" (energeia). The elementary definition of energy as "*the ability to do work*" is more a property, a characteristic of energy, than a definition, and applies only to mechanics. In [7], energy is introduced as follows: "While it is difficult to define energy in a general sense, it is simple to explain particular manifestations of energy". In [8] it is stated that, "Energy is inherent in all matter. Energy is something that appears in many different forms that are related to each other by the fact that conversion can be made from one form of energy to another. Although no simple definition can be given for the general term energy, E, except that it is the capacity to produce an effect, the various forms in which it appears can be defined with precision". *Richard Feynman,* a famous

physicist and Nobel Laureate, in his lectures on Physics (textbook) [4, 7] introduces energy by saying:

> It is important to realize that, in physics today, we have no knowledge of what energy is. We do not have a picture that energy comes in little blobs of a definite amount… There is a fact, or if you wish, a law, governing natural phenomena that are known to date. There is no exception to this law; it is exact, so far we know. The law is called *conservation of energy;* it states that there is a certain quantity, which we call *energy,* that does not change in manifold changes which nature undergoes. That is the most abstract idea because it is a mathematical principle; it says that there is a numerical quantity, which does not change when something happens. It is not a description of a mechanism or anything concrete; it is just a strange fact that we can calculate some number, and when we finish watching nature go through her tricks and calculate the number again, it is the same!

Following *David Watson* [9], energy can be described in a general way as follows: "*Energy is a property or feature of matter that makes things to move or change condition (state), or has the capability (potential) to make things to move or change*".

In all cases, if anything moves, changes state, or happens, there is an energy change, i.e., either a change of energy form (e.g., kinetic energy of wind or water falling energy to electrical energy) or a change of location (e.g., heat flow from one object to another object). Without energy, nothing would ever move or change or happen. The "changes of condition" include all possible changes in nature, such as a change in chemical composition due to a chemical reaction, change of thermodynamic or systemic state, change of phase (solid to liquid, liquid to vapor, and vice versa), changes in pressure, change of the position of a mass, and so on. Energy "feeds" earthquakes and volcanoes, "powers" bacteria, and "drives" tornadoes, typhoons, and tidal waves. Cosmologists have attempted and are attempting to explain the origin of our universe (cosmogony) via theories that are based on energy. Today, the theory that is accepted by most cosmologists is the *Big Bang* theory. The Big Bang consisted of an explosion of space with itself, different than the explosion of a bomb, the result of which is the outward propulsion of fragments. At the very beginning after the Big Bang (about 15 billion years ago), there was only an extremely hot plasma soup, which began to cool at about 10 into—43s after creation. At that time, an almost equal (yet asymmetrical) amount of matter and antimatter existed. These materials collided and destroyed one another, creating pure energy, but with an asymmetry in favor of matter, with the discrepancy growing larger as the universe began to expand. As the universe expanded more and more, and so cooled, common particles called *baryons* began to form that include photons, neutrinos, electrons, and quarks and became the building blocks of matter and life, as we know it [10–13]. Generally, all the cosmological and astronomical phenomena of galaxies, stars, nova, etc., are large-scale transformations of matter into energy or manifestations of energy movement or the transformation of potential energy into kinetic and radiant energy [14].

The Big Bang model of "cosmogony", like other scientific cosmogony models, is supported by the abundance of the "light elements" hydrogen and helium found in the observable universe, and by other evidence such as the movement of galaxies away from us (Hubble's Law), etc. Cosmogony is an area where science and theology meet in their effort to explain the *"supernatural"* event of the cosmos' creation [11, 15].

In our observable world, energy is the ultimate agent of change, the mother of all changes [6]. Therefore, to study energy, one must learn how it behaves and what it can do. Using this knowledge, humans build all human-made systems: numerical machines, manufacturing systems, power plants, computers, robots, control systems, aircraft, and so on.

## 2.3 Historical Landmarks

The history of energy and thermodynamics represents a substantial part of the history of physics, chemistry, and science. Here only the principal landmarks will be presented. More extensive presentations can be found in [1–3, 6, 16–21, 22–31].

**Ancient Times** The ancients related heat with fire. The Egyptians viewed heat as related to their origin mythologies. In ancient Greece (fifth century BC), *Empedocles* formulated his four-element theory according to which all substances derive from *earth, water, air,* and *fire*. It appears that the fire element of *Empedocles* is probably the principal ancestor of the *J. J. Becher Combustion theory* (1669) that was further developed and renamed as *phlogiston theory* by *Georg Ernst Stahl* (1694–1734). Phlogiston theory was later disapproved by *Antoine Lavoisier,* who discovered *oxygen* and proposed the *caloric theory*. The Greek philosopher *Heraclitus* expressed his famous proverb: "*All things are moving*" ($\sim$500 BC), and argued that the three fundamental elements of nature were *fire, earth,* and *water*. For the above proverbial expression, Heracletus is known as the "*flux and fire*" philosopher. Furthermore, *Leucippus* and *Democritus* formulated the first philosophy of *atomism* which is considered today as the primary link between thermodynamics and statistical mechanics. Atomism was further developed to the subsequent *atomic theory,* which was validated in the twentieth century by the experimental proof of the existence of *atoms*. Another stepping stone that inherently stimulated the development of modern thermodynamics seems to be a "poem" of *Parmenides* entitled "*On Nature*"in which he postulated that a *void* (today's *vacuum*) could not occur in nature. It was *Otto von Guericke* who proved Parmenides' postulation through his vacuum pump, incorporated into his celebrated "*Magdeburg Hemispheres*".

**1676–1689** *Gottfried Leibniz* developed the precursor to the energy concept with his *vis viva* (living force) quantity defined, for a system of interacting objects, by the sum [37, 38] :

$$\sum_i m_i v_i^2$$

where $m_i$ stands for the mass and $v_i$ for the velocity of the $i$th interacting object. Leibniz observed that, in many cases, *vis viva* was the same before and after the interaction. Newton developed the concept of *momentum* given by:

$$\sum_i m_i v_i$$

which is conserved in all interactions (*conservation of momentum law*). It was later established that both quantities are conserved simultaneously (e.g., in elastic collisions).

**1776** *John Smeaton* publishes the results of his experiments on momentum, kinetic energy, and work that supported the conservation of energy.

**1802–1807** *Thomas Young* uses for the first time the term *energy* (from the Greek ενέργεια–energeia) to replace Leibniz's *vis viva*. Young defined energy as $E = mv^2$.

**1819–1839** *Gustave Coriolis* and *Jean-Victor Poncelet* recalibrated *vis viva* as:

$$\frac{1}{2}mv^2$$

which determines the conversion constant for *kinetic energy (vis viva) into work*. *Coriolis* used the term *quantity of work* (quantité de travail) and *Poncelet* used the term *mechanical work* (travail mecanique). The precursor to "*potential energy*" is the term *vis mortua* (dead force) and the term "*ergal*" used by *Clausius,* who showed that the energy $U$ of a system is equal to *vis viva* $T$ plus *ergal* $J$ (i.e., $U = T + J$), and that the energy $U$ remains constant during any motion (conservation of energy). The term *potential energy* was introduced by *William Rank* in 1853 [39].

**1824** *Sadi Carnot* publishes his work "*Reflections of the Motive Power of Fire*" in which he studies the operation of steam engines using caloric theory. Through the development of the concept of *reversible process* (and postulating that such a process does not actually occur in nature) he discovers the concept of *entropy*. He stated that in, all heat engines, *work* (motive power) can be produced whenever a "*fall in caloric*" occurs between a hot and a cold body. Carnot proved that if the body of a "*working substance*" (such as a body of steam) is brought back to its original state (temperature or pressure), at the end of a complete engine-cycle (known as *Carnot cycle*), no change occurs in the condition of the "working body". Here is exactly where the development of the classical entropy concept was founded. Carnot defined the *efficiency* of an engine and proved that the upper bound of efficiency is set by his ideal engine (known as *Carnot engine*) (see Sect. 3.6.2).

**1834** *Emile Clapeyron* provides a graphical and analytic formulation of Carnot's theory which facilitates very much its comprehension.

**1842** *Julius Robert von Mayer,* a German surgeon, states for the first time the *mechanical equivalence principle of heat,* but he does not provide at this time a quantitative relationship between the two [40].

**1843** *James Prescott Joule* discovers the *mechanical equivalent* of heat (independently from Mayer) in a set of experiments on friction consequences, and formulated the first version of the *First Law of Thermodynamics*. His most well-known experiment used the now called *"Joule apparatus"*, a falling weight attached to a string causing a paddle immersed in water to rotate. He demonstrated that the loss in gravitational potential energy of the falling weight was equal to the gain in thermal energy (heat) of the water due to the friction with the paddle. Actually, the work of Joule received much wider attention, and the mechanical equivalent of heat is today known as *Joule's equivalent* (see *wikipedia*, mechanical equivalent of heat).

**1847** *Herman von Helmholtz* provides an alternative statement of the *First Law of Thermodynamics* (conservation of energy).

**1848–1849** *Lord Kelvin (William Thomson),* a British mathematician and physicist, develops further the concept of absolute zero and extends it from gases to all substances. He also coins the name "*thermo-dynamics*" in the framework of his studies of the efficiency of heat engines.

**1850** *Rudolf Clausius* develops further Carnot's formulation of the *Second Law*, and explains fully the properties of the ratio $Q/T$ (without giving it a name).

**1865** *Rudolf Clausius* presents the modern macroscopic concept of *entropy* as the dissipative energy use of a thermodynamic system (or "*working body*") of chemical species during a state change. This was in disagreement with earlier considerations based on Newton's theory that heat was a non-destructible particle possessing mass. Clausius also originated the term *enthalpy* as the total heat content of a system.

**1871** *James Clerk Maxwell*, a Scottish mathematician and physicist, formulates s*tatistical thermodynamics*, a new branch of thermodynamics that deals with the analysis of large numbers of particles at equilibrium (i.e., systems with no occurring changes) for which one can define average properties such as temperature $T$ and pressure $P$.

**1872** *Ludwig Boltzmann*, an Austrian physicist, formulates the Boltzmann equation for the temporal development of distribution functions in phase space, using the constant "$k$" known as *Boltzmann's constant*. In 1948, Boltzmann's definition of entropy was properly transferred by *Claude Shannon* (1916–2001) in the modern field of *information theory*.

**1874** *Lord Kelvin* provides a new formal statement of the *Second Law*.

**1876** *Willard Gibbs* publishes a long paper: "*On the Equilibrium of Heterogeneous Substances*", in which he introduces the *free energy equation* (now known by his name) that gives the amount of "*useful work*" attainable in chemical reactions. This equation is the result of studying phase equilibria and statistical ensembles and is considered as a grand equation of *chemical thermodynamics*.

**1884** *Boltzmann* uses thermodynamic considerations to derive the *Stefan-Boltzmann* blackbody radiant flux law (discovered in 1879 by Jozef Stefan, according to which the total radiant flux from a blackbody is proportional to the fourth power of its temperature).

**1906** *Walther Nernst* develops and formulates the *third law of thermodynamics*.

**1909** *Constantin Caratheodory* presents an *axiomatic formulation* of thermodynamics.

**1927** *John von Neumann* presents the concept of "*density matrix*" and establishes the field of *quantum statistical mechanics*.

**1961** *A. Rényi* presents a generalization of Boltzmann–Gibbs entropy which depends on a parameter "$\alpha$". A similar type of entropy for non-extensive processes is introduced by *C. Tsallis* in 1988.

**1976** *Elias P. Gyftopoulos* (MIT) with *G.N. Hatsopoulos* presents a unified quantum theory of mechanics and thermodynamics, and later (1984), with *G. P. Berreta,* derives a new equation of motion for general quantum thermodynamics, which covers both reversible and irreversible processes.

## 2.4 Energy Types

Energy exists in a variety of types or forms. Any and every human activity is based on the conversion of some energy type into another. In general, energy determines the quality of the processes and changes that occur on Earth and in the Universe, including both the material processes and the thinking ones. Energy is actually the measure of a physical or biological or man-made system; more specifically, it is the measurement of a substance's movement. The occurrence of the substance movements in several forms and their interrelationships and connections inspired the development of the energy concept as a common aspect for measuring them.

The nomenclature of the various types of energy can be based on the following features [32]:

- *How energy is perceived* (e.g., mechanical energy, electrical energy, radiation energy, etc.).
- *What carries the energy* (e.g., thermal energy).
- *The source of energy* (e.g., solar energy, wind energy, geothermal energy).

The energy that is available may not be in the required form. In order to obtain the form needed, the proper conversion must be performed. It should be noted, however, that not all available energy can be converted into another form of energy.

The major well-known types of energy are the following:

### 2.4.1 Mechanical Energy

Mechanical energy is distinguished in *kinetic and potential energy* to position and elasticity.

**Kinetic Energy** The energy stored in an object due to motion. An object of mass $m$ and linear velocity $\upsilon$ has kinetic energy equal to:

$$E_{k,\upsilon} = \frac{1}{2} m\upsilon^2$$

An object of moment of inertia $J$ and angular velocity $\omega$ has kinetic energy equal to:

$$E_{k,\omega} = \frac{1}{2} J\omega^2$$

**Potential Energy** The energy stored in an object due to its position, which is equal to the work done against a given force that changes its position from a reference position. It is given by the work of the force, with a minus sign, i.e.:

$$E_p = -\int \mathbf{F} \cdot d\mathbf{s}$$

where $\mathbf{F}$ is the force vector, $\mathbf{s}$ is the displacement vector, and $''\cdot''$ stands for the scalar (inner) product of $\mathbf{F}$ and $d\mathbf{s}$. From the above relationship, it follows that $\mathbf{F} = -dE_p/d\mathbf{s}$ i.e., $\mathbf{F}$ is the negative derivative of the potential energy$E_p(s)$.

**Elastic Potential Energy** The work needed to expand or compress a spring or any other mechanism that is governed by Hook's law:

$$F = -kx$$

where $x$ is the compression or expansion displacement and $k$ is the elastic constant of the spring. In this case, the *elastic potential energy* is found to be

$$E_{p,x} = \frac{1}{2} kx^2$$

for a linear spring, and

$$E_{p,\theta} = \frac{1}{2} k\theta^2$$

for a rotating spring with expansion or compression angle displacement $\theta$.

**Work–Energy Theorem** If a net constant force $F$ is applied to an otherwise free particle (object) of mass$m$, which has velocity$\upsilon_0$, it will accelerate with a constant acceleration $''a''$ given by Newton's law$F = ma$. The velocity of the particle at time

$t$ will be $\upsilon = \upsilon_0 + at$ and the change of position (displacement)$\Delta s = \upsilon_0 t + (1/2)at^2$. Now

$$\Delta E_{k,\upsilon} = \frac{1}{2}m\upsilon^2 - \frac{1}{2}m\upsilon_0^2 = \frac{1}{2}m(\upsilon_0 + at)^2 - \frac{1}{2}m\upsilon_0^2 = ma\left(\upsilon_0 t + \frac{1}{2}at^2\right) = (ma)\Delta s$$
$$= F \cdot \Delta s = W$$

This says that the change $\Delta E_{k,\upsilon}$ in kinetic energy is equal to the work $W = F \cdot \Delta s$ done by the force $F$ on the mass$m$, which is the so-called *"work–energy theorem"*.

### 2.4.2 Forms of Potential Energy

In general, potential energy appears in several forms, not only in the above mechanical forms. Potential energy is the energy in the matter due to its position or the arrangement of its parts and includes the following:

- Gravitational potential energy
- Chemical potential energy
- Electrical potential energy
- Magnetic potential energy

**Gravitational Potential Energy** This is the energy of objects due to their position above the ground. When an object is lifted or suspended in the air, work is done on it against the pull force of gravity. When the object succumbs to the force of gravity falling towards Earth, it converts potential energy into kinetic energy.

**Chemical Potential Energy** This is the internal energy associated with the various forms of aggregation of atoms in matter (for the internal energy, see Sect. 2.4.3). For example, the chemical arrangement (makeup) of gasoline makes it a good fuel-energy source, which, when burned (combusted), releases large quantities of energy that can, e.g., move an airplane. During combustion, chemical bonds are broken and reformed, transforming gasoline into by-products such as water and carbon dioxide, releasing energy.

**Electrical (or Electrostatic) Potential Energy** This is the potential energy $U$ stored[1]. in a given configuration of point electrical charges or in a given electrostatic field distribution. $U$is equal to the work $W$ needed to bring the charges one by one, slowly, from their infinite separation (i.e., from the zero reference) to the given configuration.

Examples of this type of energy are the following: (i) A battery has chemical potential energy along with electrical potential energy. Turning on a device that is

[1]In many cases the expression "the energy stored" is avoided because energy may be erroneously depicted as a substance contained within a substance.

battery-powered, the electrical potential energy stored in the battery is converted into other forms of energy, such as light, sound, mechanical energy, or thermal energy; (ii) A power plant or hydroelectric dam maintains electrical potential energy due to the spinning generator; (iii) A solar cell stores electrical potential energy similar to a battery, as long as sun rays are impacting on it.

**Magnetic Potential Energy** This is the potential energy $E_{mp}$ of a magnet. It is equal to the work of the magnetic force (torque) done to realign the vector of the magnetic dipole 'moment'**m**. Thus, if **B** is the magnetic field vector, $E_{mp}$ is given by:

$$E_{mp} = -|\mathbf{m}| \cdot |\mathbf{B}|$$

In the case of an inductor via in which a current $I$ is flowing, $E_{mp}$ is equal to:

$$E_{mp} = \frac{1}{2} LI^2$$

where $L$ is the *inductance* of the inductor.

### 2.4.3 Internal Energy in Thermodynamics

Internal energy is the sum of all microscopic forms of energy in a system. It can be considered as the sum of potential and kinetic energies of the molecules of the object and involves the following subtypes of energy not including the kinetic energy of the system (body) as a whole [33–46]

- Sensible energy
- Latent energy
- Chemical energy
- Nuclear energy

In particular, the sum of the sensible and latent internal energy is called thermal energy.

#### 2.4.3.1 Sensible Energy

This is the part of the internal energy in the system due to the kinetic energies of its molecules (particles). It is the heat (or thermal energy) provided to the system (body) when the heat is not used to change the state of the system (as in the latent heat). The sensible energy is actually the heat that changes the temperature of the system. The sensible energy is transported in three distinct ways, namely, via conduction, convection, and radiation (or their combinations). Convection (heat or mass convection) is the motion of molecules within fluids (liquids, gases). In fluids,

convection occurs via diffusion (random motion of particles) and advection (heat transfer by the macroscopic motion of current in the fluid). Heat convection is distinguished by two principal types as follows:

- *Free or natural convective heat transfer* caused by the circulation of fluids due to the density alterations produced by the heating process itself.
- *Forced convective heat transfer* which takes place passively due to fluid motions (current) that would occur independently of the heating process. Forced convection is sometimes called *heat advection*.

#### 2.4.3.2 Latent Energy

This is the part of internal energy associated with the phase (state) of the system. It is the amount of energy in the form of heat that is absorbed or released by a chemical compound during the process of changing phase (solid, liquid, and gas). The concept was coined by Joseph Black (starting in 1750) and the name latent comes from the Latin latere (= hidden) [46]. Today, in place of latent energy, we frequently use the term enthalpy. The two typical latent heats (or enthalpies) are as follows:

- Latent heat of fusion (melting)
- Latent heat of vaporization (boiling).

#### 2.4.3.3 Chemical Energy

Chemical energy is the internal energy due to the arrangement of atoms in the chemical compounds. It is produced via reactions that take place when the bonds between the atoms loosen or break and create new compounds. The chemical energy is released in the form of heat. Reactions that release heat are called exothermic. In general, exothermic reactions consume oxygen, and, when the bonds break or loosen, oxidation occurs almost instantly. Chemical reactions that need heat to occur usually store some of this energy as chemical energy in the bonds of the newly generated compounds. Chemical energy is a source of energy easy to assess and very efficient to store and use. The chemical energy contained in food is converted by the human body into mechanical energy and heat. The chemical energy in fossil fuel is converted into electrical energy at power plants. The chemical energy stored in a battery can provide electrical power via electrolysis.

As we will see in Sect. 2.5.3, chemical energy can be used in several ways to obtain alternative renewable-energy sources very useful for the future survival of humans.

#### 2.4.3.4 Nuclear Energy

Nuclear energy is the energy due to the strong bonds within the nucleus of the atom itself. It can be released via fission (splitting) or fusion (merging together) of the nuclei of atoms. When a nucleus is split apart, a tremendous amount of energy E is released, in both heat and radiant-energy forms, according to Einstein's matter–energy equivalence equation $E = mc^2$, where m is the (converted) mass and c is the speed of light.

A nuclear power plant employs uranium as fuel, the atoms of which are split apart in controlled, nuclear chain reactions, through appropriate control rods. In a chain reaction, the particles released by atom fission go off and strike other uranium atoms, splitting them. If a chain reaction is not controlled, then an atomic bomb may be obtained under particular conditions (e.g., pure uranium-235 or plutonium, etc.). The nuclear fission creates radioactive by-products that may harm people. Therefore, very robust concrete domes are constructed to contain the radioactive material, in the case of an accident. The large amount of heat energy released by a nuclear reaction is fed to a boiler in the core of the reactor. The boiled water around the nuclear reactor core is sent to suitable heat exchangers that heat a set of pipes filled with water to produce steam. This steam is passed via another set of pipes to turn a turbine that generates electricity.

Fusion is the process of merging (joining) together smaller nuclei to make a larger nucleus. Through the fusion of hydrogen atoms, the sun creates helium atoms and gives off heat, light, and other radiation. The difficulty of controlling nuclear fusion within a confined space is the reason why, at the moment, scientists have not yet succeed in constructing a fusion reactor for generating electricity. It should be noted that nuclear fusion creates less radioactive material than fission.

### *2.4.4 Evidence of Energy*

Mechanical motion, thermal energy, sound, and light cannot easily be classified as kinetic and potential energy since they always contain a combination of the two. For example, a pendulum has an amount of mechanical energy that is continually converted from gravitational potential energy into kinetic energy, and vice versa, as the pendulum oscillates back and forth. Thermal energy consists of both kinetic energy (the sensible energy) and the latent (phase-dependent) energy. An *LC* electric circuit is, like the pendulum, an oscillator (electric oscillator), and its energy is on average equally potential and kinetic. Therefore, it is arbitrary to characterize the magnetic energy as kinetic energy and the electrical energy as potential energy, or vice versa. The inductor can be either regarded as analogous to a mass and the capacitor as analogous to a spring, or vice versa. Likewise, the sound is made up of vibrations and contains both kinetic and potential energy. Extending the reasoning about the *LC* electric circuit to the empty space electromagnetic field, which can be regarded as an ensemble of oscillators, we easily verify that radiation energy

(energy of light) can be considered to be equally potential and kinetic [7]. This interpretation is used when we are interested in the electromagnetic Lagrangian which involves both a potential energy and a kinetic energy component. Now, in empty space, the photon (which is massless, has electric charge, and does not decay spontaneously in empty space) travels with the speed of light,$c$, and its energy $E$ and momentum are related by

$$E = pc \,(p = \text{magnitude of momentum vector } \mathbf{p})$$

The corresponding equation for particles that have mass $m$ is as follows:

$$E^2 = p^2c^2 + m^2c^4$$

where $mc^2$ is the so-called rest energy. At speeds much smaller than $c$, the kinetic energy of the particle is found to be equal to $p^2/2m$ [47, 48]. This expression is used when we are interested in the energy-versus-momentum relationship. The formula $E = pc$ says that the energy of a photon is purely kinetic. The above two, apparently different, results about the energy of light are actually consistent. In the first, the electric and magnetic degrees of freedom are transverse to the direction of motion, while in the second, the speed is along the direction of motion. In all the above cases, where the energy cannot be classified as pure kinetic or pure potential energy, we say that we have "evidence of energy" [7, 49].

## 2.5 Energy Sources

The energy sources available to humans, besides the perpetual energy of the Sun, are categorized into the following:

- **Exhaustible (non-renewable) sources:** Fossil fuels (oil, coal, natural gas) and nuclear fuel.
- **Renewable (non-exhaustive) sources:** Biomass, geothermal, hydropower, solar, wind, and other alternative non-fossil sources (bio-alcohol, bio-diesel, liquid nitrogen, hydrogen, etc.).

A short presentation of them follows [8, 16–21, 32 47, 48, 50–58]. Detailed technical issues can be found in relevant books (e.g., [59–65]).

### *2.5.1 Exhaustible Sources*

Exhaustible energy sources are finite resources, and eventually, the world will run out of them, or it will become extremely difficult and expensive to retrieve those that remain.

**Fossil Fuels** were formed hundreds of millions of years ago before the time of the dinosaurs (hence the name *fossil fuels*). The age in which they were formed is known as *Carboniferous Period* (from the carbon that is the basic constituent of fossil fuels), which was part of the *Paleozoic Eva* [32]. They were formed from the dead and decayed plants and animals, under the effects of high pressure and high temperature. The dead plants and animals sank to the bottom of the swamps of oceans. They formed layers of the so-called *"peat",* which, over the centuries, was covered by sand, clay, and several minerals, and then were transformed into a form of rock named *"sedimentary"*. As more and more rock piled on top of rock, and weighed more and more, the peat was squeezed by the very high pressures which developed, and the water was ejected. In this way, eventually, the matter became oil, coal, and natural gas.

**Oil** is used to produce petrochemical products (gasoline, petroleum, kerosene, diesel, plastic fabrics) using the distillation method. The reserves of fossil fuels are being depleted at very high rates, a fact that raises strong sustainability concerns. **Oil** supplies about 40% of world's currently used energy (Fig. 2.1).

**Coal** was formed from decomposed plants via the same process (called *coalification*), but it took comparatively less time than that of fossil fuels. Fossil fuels have been used in China from very early times (around 1,000 BC), and oil was in use by ancient Babylonians (about 6,000 years ago). The intensive use of fossil fuels started during the Industrial Revolution. The principal ingredients of coal are carbon, hydrogen, nitrogen, oxygen, and sulphur, but the actual composition varies among the various types of coal (anthracite, bituminous, lignite). It is noted that anthracite is the hardest type and lignite the softest type of coal. It is estimated that, today, coal provides about 28% of the total energy consumed by humans.

**Natural Gas** collects in large quantities and contains mainly methane and some small amounts of butane, propane, ethane, and pentane. It is thinner than air, odorless, and highly inflammable. Its typical use is for cooking in the form of liquefied petroleum. The first discoveries of natural-gas seeps were made in Iran. Natural gas is commonly located near petroleum underground. It is pumped from below ground and transported via pipelines to storage tanks. Since natural gas has no odor, before going to the pipelines and storage areas, it is mixed with a chemical to acquire a strong odor, smelling like *rotten eggs*. This odor is an indication that there is a leak, and so the users of the natural gas evacuate and avoid possible ignition of the gas. Natural gas covers around 20% of the world's energy demand.

**Nuclear Fuel** The basic fuel of nuclear power reactors is uranium (U), a very heavy metal (with a melting point of 1,132 °C). Uranium is mildly radioactive, exists in the crust of Earth and contains abundant concentrated energy. It was formed in *supernovae* about 6.6-million years ago. Today, its radioactive decay provides the principal source of heat inside the Earth, causing convection and continental drift. Uranium is 40 times more abundant than silver. It is primarily used for the production of electricity from nuclear plants, but it is also used for marine propulsion and in medicine for cancer treatment via the production of

**Fig. 2.1** Two examples of technology for oil-field extraction (http://www.inn-california.com/valleys/images/ca0159.jpg, http://www.amroll.com/artman/uploads/1/oil.jpg. The reader is informed that Web figures and references were collected at the time of the book's writing. Since then, some of the urls may not be valid due to change or removal by their creators, and so they may no longer be available

radioactive isotopes. The various processes that are used for the production of electricity from nuclear reactions are collectively called the *"nuclear fuel cycle"*. The nuclear fuel cycle starts with the mining of uranium and finishes with the disposal of the nuclear waste. With the reprocessing of used fuel as a source for nuclear energy, the stages constitute a *"true cycle"* [33]. In a number of areas on Earth, the concentration of uranium in the ground is adequately high so that the mining and use of it as nuclear fuel is economically feasible. In these cases, we say that we have *uranium ore*. The uranium ore can be recovered by *excavation* or in *situ* techniques. The decision as to which technique is to be used is based on the nature of the ore body at hand, and on safety and economic issues. *Vaclav Smil* argues that although the promoters of nuclear energy in the 1970s were saying that, by the year 2000, all electricity in U.S. would come from nuclear fission, the reality of 2008 was that coal-fired power plants produced 50% of the electricity and nuclear stations only 20%, with no operating commercial breeder reactors (The

American, A Magazine (http://www.american.com/archive/2008/november-december-magazine/). Figure 2.2 shows an example of nuclear power plant.

### 2.5.2 Renewable Sources

Fossil fuels are still the main energy source used for the growth and development of modern industries and the human society, but reliance on them presents a serious problem since they are exhaustible. Renewable-energy sources are energy resources that are naturally replenishing but flow-limited. They produce little or no pollution or greenhouse gases, and they will never run out. Today, about 50% of the energy provided by renewal sources is used to produce electricity. Renewable sources include biomass, geothermal, hydropower, solar, wind, and ocean energy.

**Biomass Energy** Biomass, which is matter otherwise thought of as garbage, has been an important source of energy from the first time humans started burning wood to cook food and warm them- selves in winter. In general, biomass is organic material produced from plants and animals and, besides the release of heat, can also be converted to electricity, biodiesel, ethanol, and methane gas. The conversion of biomass to electrical energy is performed in manufacturing industries, where left-over biomass, like wood waste or paper waste, is burned to produce steam, which is then used for the electricity generation. Waste coming from household and office contains some kinds of biomass that can be recycled for fuel and other uses, thus cutting down on the need for "landfills" into which to put garbage. The two principal ways of using the biomass for energy are the following:

- The biomass is tapped at the landfill for combustible waste products. The decomposition of the garbage produces methane gas which is immediately collected by pipelines that are installed within the landfills. This methane is then used in power plants to produce electricity. This category of biomass is known as *"landfill gas"*.
- The waste wood, agricultural waste, and other organic material from industrial and municipal wastes are collected in large trucks and transferred to a biomass power plant. Here it is properly fed into a furnace where it is burned and generates the heat used to boil water in a boiler. Then the energy of the resulting steam turns turbines and electric generators.

**Geothermal Energy** *Geothermal* comes from the Greek *"γεω: geo"* (earth) and *"θερμότητα : thermal"* (heat). So geothermal means *"earth-heat"* and refers to the energy existing inside the Earth's crust. To generate electricity from a geothermal source, deep wells are constructed and high-temperature water or steam is pumped to the surface of the Earth. Geothermal power plants are built near *geothermal reservoirs* (large areas with naturally hot water). Other uses of geothermal sources include the heating/cooling of houses (heating in winter, cooling in summer). The physics of geothermal energy is briefly as follows. The crust of the Earth floats on

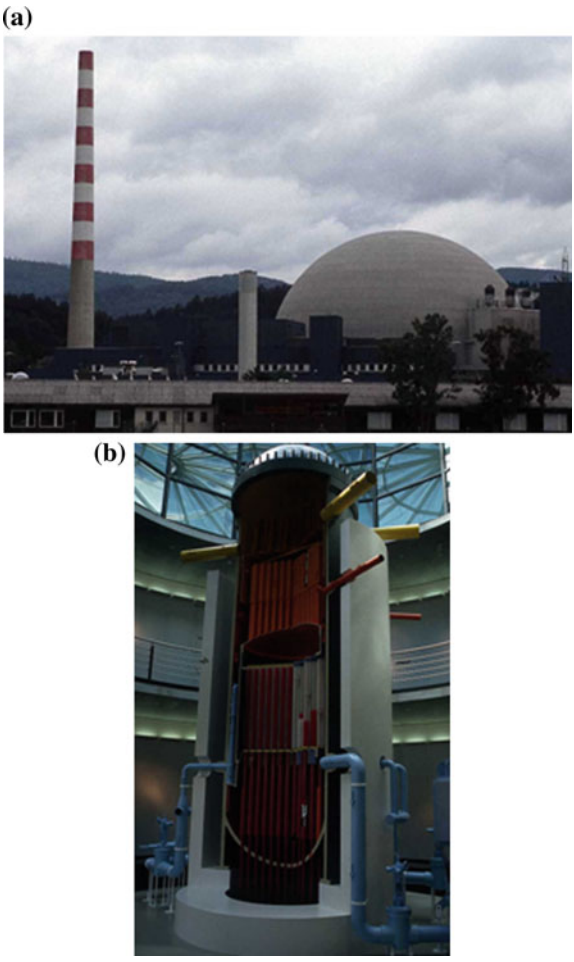



**Fig. 2.2** **a** A photo of a nuclear-power plant (http://www.picture-newsletter.com/nuclear//nuclear-plant-m82.jpg), **b** interior of the nuclear-power plant (http://www.picture-newsletter.com/nuclear/nuclear-power-nv5.jpg)

the *magma* (the hot liquid rock lying below the crust). When the magma comes to the surface of the Earth in a volcano, it is the well-known "lava". The temperature of the rock increases with the depth (about three degrees Celsius for every 100 meters below ground), and so, if we go to about 3,000–3,500 meters below the ground, the temperature of the rock would be sufficient to boil water. Actually, the hot water can have temperatures as high as 150 °Celsius or more (i.e., hotter than boiling water), but, because the water is not in contact with the atmosphere, it does not turn into steam. Hot springs are this very hot water emerging from the ground through a crack.

**Hydropower** was invented in the 1880s, and, for many decades, the moving water was used to turn wooden wheels attached to grinding wheels to grind flour or corn. These *"wooden and grinding wheels"* were the well known "water mills". Water is

a renewable resource, constantly replenished by the cycle of evaporation and precipitation. Today, moving water is extensively used for electrical-power generation, the so-called *hydro-electric power* (from the Greek "υδρο-hydro: water" and "electric"). Here, the mechanical energy of the moving (swift-falling or descending) water is used to turn the blades of a turbine. Hydroelectric power plants are built at the place of the water fall or in man-made dams (Fig. 2.3). The electric power generated depends on the amount and speed of the flowing water. In many countries, more than 10% of the total electricity is produced by hydropower.

**Solar Energy** Sunlight (solar energy) can produce electrical and thermal energy by two techniques, namely, *direct* and *indirect*. In the direct, the technique employs *solar cells* or *photovoltaic systems*. In the indirect technique, solar energy is collected by *power thermal collectors* to heat fluid and generate steam, which is fed to steam engines that produce electricity. The varying intensity of sunlight, depending on the location of the system and the weather conditions, is the major drawback of this energy source.

**Solar cells or photovoltaic (PV) cells** are made from *silicon*. When the sun's rays strike the solar cell, electrons are knocked loose and move toward the treated front surface. In this way, an electron imbalance is developed between the front and the back of the cell. Then, by connecting the two surfaces with a wire (or another connector), we make an electric current flow between the negative and positive terminals. Multiple individual solar cells are grouped to form a PV module, and many modules are arranged in arrays that are attached to special tracking systems to maximize the sunlight collection all day long. Figure 2.4 shows an example of a solar-power system.

**Solar collectors** that store heat energy may be "batch"-type collectors, while other kinds of solar collectors use circulated fluid (e.g., water or antifreeze solution) to provide the heat for storage in an insulated reservoir or for direct use. A complete solar-heating system is composed of a collector, a heat-transfer circuit, and a heat-storage device. Three basic plate-collector system types are the *passive breadbox* collector, the *active, parallel, flat-plate* collector, and the *active, serpentine, flat-plate* collector. Flat-plate collectors are usually employed to heat water or housing spaces, and other domestic buildings. To generate electricity, solar-power plants use parabolic solar reflectors of the *"parabolic dish"* or the *"parabolic through"* type. A parabolic dish concentrates the parallel sunlight rays at the focal point of the collecting lens. A kind of solar, reflector, dish concentrator can be also obtained by lining the interior of a cardboard box with aluminum foil. Parabolic troughs are cheaper than the dish. A simple way to make a parabolic trough is to use a sheet of cardboard lined with a piece of aluminum foil. The drawback of solar energy is that it is available for use only during sunny days. During nights and on cloudy days, the solar systems cannot produce energy. For this reason, some systems are of the hybrid solar and natural gas type.

**Wind Energy** Here the wind's speed is used to rotate suitable blades, which are connected to an electrical power generator. The blades of the turbine are connected

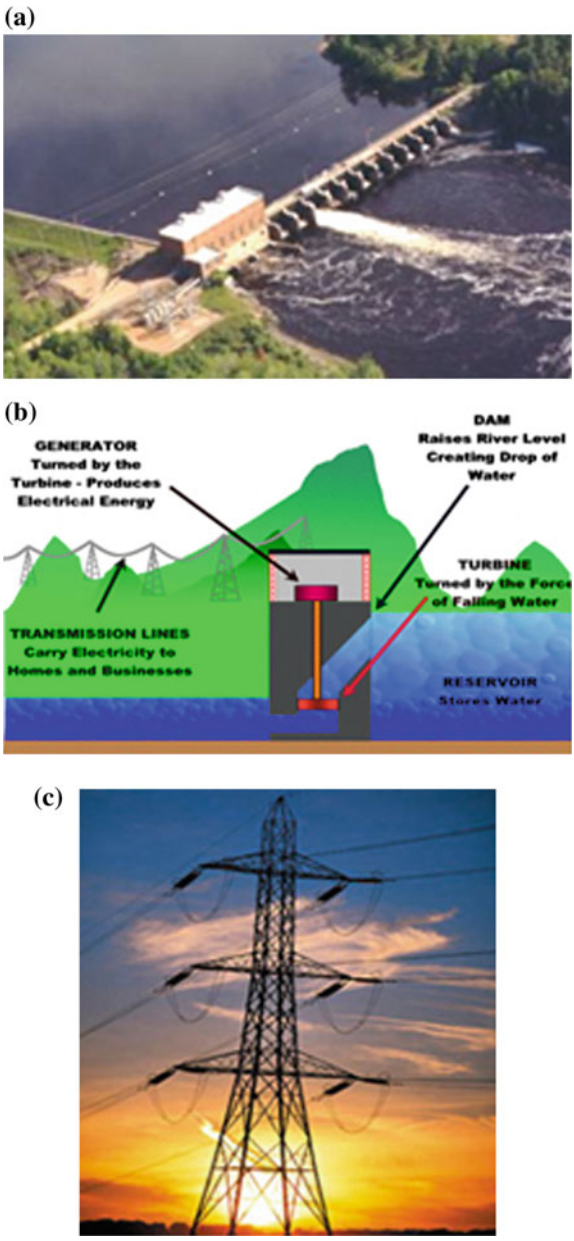



**Fig. 2.3** **a** A hydroelectric power plant example **b** Schematic hydro-power system: turbine, generator, transmission lines (*Source* http://earthsci.org/mineral/energy/hydro/hydro.htm **c** A tower for electric power transmission *Source* http://www.cbc.ca/news/background/poweroutage/electricity_terms.html

to a rotating axis via a gearbox that increases the rotation speed. This in turn rotates an electrical generator. Wind turbines or wind mills are placed in the best orientation in order to make maximum use of wind energy. Wind-power plants involve groups of wind generators (Fig. 2.5). Today, many wind-generator companies

**Fig. 2.4** A photo of a typical solar-power plant (http://www.publicdomainpictures.net/view-image.php?/image=30618&picture=solar-power-plant)

**Fig. 2.5** An example of a wind-power-generator "farm" (http://www.worsleyschool.net/science/files/windpower/page.html)

produce and install wind-power plants appropriate for both domestic and industrial systems. Again, wind generators have the drawback that they can provide sufficient amounts of electrical energy only if installed in windy areas. A wind turbine can provide sufficient power if the wind speeds are higher than 10 to 15 miles per hour, in which case each turbine produces about 50–300 kw.

**OceanEnergy** The ocean provides three main types of renewable energy, i.e., *wave energy, tidal energy,* and *thermal energy*. Of course, the ocean provides non-renewable energy as well, since, in many regions, huge amounts of oil and natural gas are buried below the seabed.

**Wave Energy** The kinetic energy of the moving ocean waves is used to rotate a turbine, and the resulting air circulation turns a generator. In other systems, the up and down motion of the wave is used to power a piston moving a cylinder up and down. The piston is used to rotate an electric generator (Fig. 2.6).

**Tidal Energy** The tides coming onto the shore are trapped in reservoirs behind dams. These reservoirs operate like standard hydropower systems, i.e., when the tide drops, the water behind the dam is let out as in conventional hydroelectric

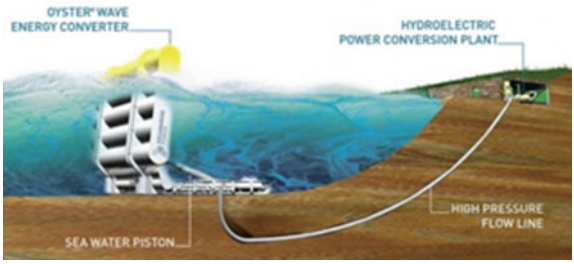



**Fig. 2.6** An example of an oyster system, a type of hydroelectric wave-energy plant (http://www.gizmag.com/tag/wave+power)

plants. In other tidal technologies under development, the natural ebb and flow of water is not changed. For the tidal energy to work efficiently, large tidal variations (at least 16 ft between low tide and high tide) are required, but very few places exist on the Earth with this tidal change. For this reason, tidal systems are of limited applicability. Three examples of tidal-power plants are the 500-kw plant installed on the Scottish shoreline of Islay, the La Rance Station in France (250 MW), which is capable to power 240,000 homes, and the Merrimack River tidal system along the Massachusetts-New-Hampshire border.

**Ocean Thermal Energy** Solar energy heats the surface water of the ocean. Thus the surface layers of the ocean are warmer than the deeper, colder layers. In tropical areas, the temperature difference between surface water and deep water can be very significant. *Ocean thermal energy conversion* **(OTEC)** *systems* exploit these temperature differences. OTEC power systems can work well if this difference is higher than 38 °F, which can occur in tropical regions. There are two types of OTEC systems: *closed* and *open*. A *closed system* converts warm surface water into steam under pressure through several heat-exchange cycles. Cold, deeper water is pumped via pipes to a condenser on the surface. The cold water condenses the steam, and the closed cycle is repeated. An *open OTEC system* turns the steam into fresh water, and new surface water is fed into the system. The generated electrical power is transmitted to the shore. Small-scale OTEC systems have been installed and tested in tropical coasts, but, in general, OTEC technology is far from economically-practical commercial use.

### *2.5.3 Alternative Energy Sources*

Some major alternative nonfossil fuels are the following:

**Bio-Alcohol** Fuels of this category are produced from plant sources. They include ethanol, butanol, methanol and propanol, and (due to their properties) can be used in car engines. The most popular is ethanol which is produced from sugar fermentation. In the US, ethanol is typically produced from corn and is being used extensively as a fuel additive. Its use as a part of a mix with gasoline or as ethanol

fuel, alone, is increasing. But, in general, bio-alcohol is not used in the majority of industries because it is more expensive than the derived from fossil fuels. Careful energy analysis has shown that there is a net loss in alcohol use, and further improvements need to be done (use of optimized crops, elimination of pesticides, etc.) to make bio-alcohol a viable renewable fuel. Of course, chemically there is not any difference between biologically produced alcohols and those produced by other sources.

**Bio-Diesel** Bio-Diesel is a renewable fuel for diesel engines produced from natural vegetable oil (such as *soybean oil*) and animal fat by a reaction with an alcohol (such as methanol or ethanol) in the presence of a catalyst. This reaction produces *mono-alkyl esters* and *glycerin,* which is removed. Bio-diesel can be mixed with petroleum-based (normal) diesel fuel in any analogy and used in all existing diesel engines without any modification (or with minor modification). Because glycerin is removed, bio-diesel is different from standard raw vegetable oil. The bio-diesel has reduced the overall emission of pollutants and possesses good lubrication characteristics. Currently, it has been introduced in all types of locomotion all over the world.

**Liquid Nitrogen** This type of fuel can only be used in cars equipped with nitrogen power combustion. These cars have a circuitry similar to electric cars, but the batteries are replaced by nitrogen fuel tanks. Liquid nitrogen is inert, odorless, colorless, non-corrosive, nonflammable, and extremely cold. Nitrogen's concentration in the atmosphere is 78.03% by volume and 75.5% by weight. Although it is inert and does not support combustion, it is not life supporting, since when it is combined with oxygen to form oxides of nitrogen, it may reduce the concentration of oxygen in the air below that needed for life. It is noted that, at low oxygen concentrations, unconsciousness and death may come very quickly and without warning. The use of liquid nitrogen (a *cryogenic liquid*) requires special protective measures (such as a full-face shield over safety glasses, loose-fitting thermally-insulated gloves, long sleeve shirts, trousers without cuffs, and safety shoes).

**Hydrogen** Hydrogen is an energy carrier, like electricity, and may be produced from many sources (water, fossil, biomass, etc.). Hydrogen can be obtained as a by-product of many chemical reactions. The hydrogen fuel cell converts the chemical energy stored in a hydrogen molecule into electrical energy. The most economic method of producing hydrogen is *steam reforming,* which is employed in industries to extract hydrogen atoms from methane ($CH_4$). This method has the disadvantage that, together with the hydrogen, *greenhouse gases* **(GHG)** that are the main cause of global warming are emitted. A second method of hydrogen production is *electrolysis*, a process that splits hydrogen from water. This method does not produce GHG emissions but is very expensive. New methods and technologies are currently under development, e.g. it has been discovered that some *algae* and *bacteria* produce hydrogen. The advantage of hydrogen fuel is that a

mass of hydrogen contains 2.8 times the energy in the same mass of gasoline. Although hydrogen fuel is already in use, it's excessively high production cost makes it, at the moment, not commercially viable. Most hydrogen is used in processing foods, refining, and metal treatment. Today, there are many hydrogen-fueled vehicles (buses and cars) powered by electric motors.

The use of renewable-energy sources (including the human-made ones) is steadily increasing because of their environmental advantages, i.e., reduced greenhouse-gas emissions. It is estimated that currently about 20% of the world's energy is produced from renewable sources, with biomass being the dominant renewable-energy source in developing countries.

## 2.6 Environmental Impact of Energy

The production, transportation, and use of energy (of any kind) have a visible, substantial impact on the environment, which includes air and water pollution and solid-waste disposal. This impact takes place at every stage of the energy cycle, from energy-extraction methods to the ways the raw resources are transported and used by humans in the industrial and domestic sectors. In the previous section, we saw that the energy sources are categorized into exhaustible and renewable sources. Therefore, here we will discuss the energy's impact on the environment separately for each one of these two energy categories.

### *2.6.1 Impact of Exhaustible Sources*

Technically, fossil fuels are the incompletely oxidized and decayed animal and vegetable materials (petroleum, coal, and natural gas) that can be burned or otherwise used to produce heat. Up to the Industrial Revolution, most energy sources were used for cooking and heating, with only small quantities used in industry. The Industrial Revolution impelled an increased utilization of conventional fuels (wood and coal) and initiated the development of new ones. Energy consumption is not equally distributed throughout the world. The developed countries, which represent only 20% of the world's population, consume about 80% of the natural gas, 60–65% of the oil, and 50% of the annual coal production [59]. Combustion of these fossil fuels is one of the principal factors contributing to GHG emissions into the atmosphere. But, the most serious long-term economic and environmental problem posed to the world seems to be the high consumption rate of natural sources. As the quantity of these energy resources becomes smaller and smaller, their cost will increase making products that use them much more costly, and nations will fight to maintain access to them. According to [57], the amount of nonrenewable sources remaining available as of 2003 was as follows:

**Oil** About 1000 billion barrels (sufficient for about 38 years)
**Natural gas** Approximately 5400 trillion cubic ft (sufficient for about 59 years).
**Coal** About 1000 billion metric tons (sufficient for about 245 years).

The principal types of harmful outcomes of the conversion of fossil fuels to energy are the following [56–58]:

- *Air pollution*
- *Water pollution*
- *Solid-waste disposal*

**Air pollution** affects the formation of urban smog, acid rain, ozone thinning, and global warming. The main cause of global warming is considered to be the carbon dioxide that is released by the combustion of fossil fuels. Other emissions released by this combustion include carbon monoxide, hydrocarbons, etc. A very injurious gas with long-term effects is nitrous oxide, which is released when coal is burned. About 50% of the nitrous oxide and 70% of the sulfur oxide coming directly from the burning of coal. Smog can cause human diseases and can also affect the sustainability of crops, because smog seeps via the protective layer of the leaves and destroys critical cell membranes. Acid rain (rain which is more acidic than conventional rain) damages lakes, forests crops, and monuments. Today, the least harmful fuel to the environment is natural gas since it releases very little carbon dioxide and other GHGs.

**Water pollution** (surface and ground) can occur during the extraction of oil, coal, and gas that typically exist underground and below groundwater reserves. The drilling process can break the natural barriers between the fossil-fuel and the groundwater reserves. Also, water supplies may be contaminated by fossil fuel during transportation and storage (broken pipes or storage tanks). But water pollution is mainly due to industry which disposes process wastewaters, cooling waters, spent process chemicals, and other contaminants into surface water, either *directly* (by piping them to a nearby lake, river, or steam), or *indirectly* (by adding them to a public sewer which eventually leads to a water body). The treatment of these wastes, so that they cease to be hazardous for human health and the environment, is excessively costly.

**Solid-waste disposal** The conversion of fossil fuels may also lead to accumulated solid waste. Solid-waste disposal also comes from agricultural wastes (such as crop residues or manure from animal feeding), which pose problems in rural areas. Other solid wastes originate from industry (process wastes) and domestic operations (institutional, household, and commercial wastes). The collection, transport, processing, recycling, disposal, and monitoring of waste materials is collectively called *"waste management and control"*.

Other types of environmental impacts of fossil fuels are [48]:

- *Land subsidence*
- *Land and wildlife disruption*

- *Drilling-mud releases*

**Land subsidence** This is due to the large holes left underground when oil and gas are taken out of underground reserves. Naturally, if there is no more mass to support the land above, it can collapse, with serious environmental and property damage.

**Land and wildlife disruption** The process of extracting fossil fuels from the Earth has large-scale infrastructure requirements. These include the construction of new roads, storage tanks, oil and gas wells, and other numerous other constructions. These infrastructure developments are usually done in rural and wilderness areas, and so they have serious impacts on plants, trees, and wildlife in general.

**Drilling-mud releases** These include the drilling fluids or muds employed for lubrication, which contain several harmful chemicals (toxic or non-toxic). The contamination of these muds occurs both in the immediate area of drills and in the wider vicinity due to their subsequent dumping.

**Nuclear energy** Nuclear energy poses a special environmental impact due to the production of long-life (thousands of years) radioactive wastes, either as spent-fuel quantities or as remainders of end-of life, dismantled power plants that have been operated for over 35–40 years.

### *2.6.2 Impact of Renewable Sources*

Although renewable resources are more environment friendly than fossil fuels, they are not appropriate for all applications and places. There are still several issues that must be taken into account from an environmental viewpoint because they are not readily utilizable in their natural forms [58, 63–65]. A brief account of these issues for each kind of renewable energy resources follows.

**Biomass Energy** This category of energy has more serious environmental impacts than any other renewable type of energy, except hydropower. Combustion of biomass releases carbon dioxide into the atmosphere producing air pollution and contributing to global warming. In addition, nitrogen oxides and particulates (soot and ash) are emitted. Unwise cutting of trees and plants (forests, peat, and so on) may lead to sterile soil, through the increased surface run-off and the resulting wind erosion. Another uncertain factor involved in biomass impact is that there is no unique biomass technology. Actually, many technologies exist for energy production from biomass, each with different environmental impacts. The production of energy from biomass needs to be done with care, since the reduction of plants and trees implies that less carbon dioxide is absorbed, thereby increasing the greenhouse phenomenon. Overall, biomass emissions are similar to those of coal-based plants, but with much smaller quantities of sulfur dioxide and harmful metals (such as mercury and cadmium). The most serious impact of biomass-based energy is due to the emission of particulates that must be restricted via suitable

filters. Nevertheless, the major advantage of replacing fossil fuels by biomass is that, if done in a sustainable way, it can reduce considerably the emission of GHGs.

**Geothermal Energy** All geothermal-energy-resource types suffer from a common category of environmental concerns. This includes the potential release of water or gas from underground reserves that contain toxic substances, air and water pollution, siting and land subsidence, and noise pollution from the high-pressure steam release. Local climate changes may also occur due to the release of heat.

**Hydropower** In many countries, large and small hydropower plants have a strong impact on fish populations (e.g., salmon, trout, etc.). Young fish are forced to travel downstream via several power plants at the risk of being killed by turbine blades at each plant. To reduce this impact, national laws have been enacted that prohibit new hydropower plants to be installed, enforce a considerable reduction of peak-power output in existing ones, direct water around the turbines during the times of the year when the fish are traveling, use screens that keep fish away from turbine blades, or flash underwater lights to direct night-migrating fish around the turbines. Also, the vast reservoirs needed for large hydropower stations flood broad expanses of farmland, affecting forest and wildlife populations and causing severe changes in the ecosystem and human economic activity. River ecosystem changes (upstream and downstream) can also be caused by the hydropower dams. In modern reservoirs, the mercury naturally existing in the soil is released by chemical reactions. Although existing and new hydropower stations have been modernized to minimize these negative consequences, for environmental protection, it is not anticipated that hydropower will increase in total in the near future by more than 10–20%. On the contrary, it may remain constant or be reduced in the long term due to lessening rainfall, the policies of protecting and restoring perilous wildlife and fish, and the increasing demand for drinking and agricultural water.

**Solar Energy** Solar cells and collectors do not produce waste or gas emissions, but the substances used in photovoltaic cells (e.g., cadmium or arsenic) are hazardous for the humans that are exposed to them. The silicon used in PVs, which is usually inert, might be harmful if breathed in as dust, and so proper protection measures should be taken. Solar systems themselves are manufactured and installed through the consumption of fossil fuels, which produce pollution. But the quantities of fossil fuels needed for this are much smaller than the corresponding quantities consumed for other comparable fossil-based energy systems. Also, the land needed for large-scale power plants (utilities) pose problems similar to other types of energy production.

**Wind Energy** Wind turbines, similar to solar cells, do not pollute the land, water, or air, but they result in visual and sound pollution in the landscape, which may affect wildlife. Wind turbines provide a good solution to the energy needed in agricultural/farming areas, but in other cases they may create difficult conflicts in land use (e.g., in forest areas, tree clearing may be required and, in other places, existing roads have to be cut). In many areas, where large-scale wind-power systems are installed, a massive death or injury of birds due to collisions with the

rotating wind turbines or electrocution has been observed. To face this and other similar problems, special measures should be taken, acceptable by the communities concerned.

## 2.7 Violent Manifestations of Earth's Energy

Our exposition of energy' properties and issues is complemented in this section with a short discussion of earthquakes, volcanoes, tornadoes, and hurricanes [66–74]. These violent geological and meteorological phenomena have shown, over human history, severe destructive effects that have killed a huge number of people and destroyed their property.

### *2.7.1 Earthquakes and Volcanoes*

The structure of the Earth involves the core (about 400 miles below the ground), the mantle which is outside the core (about 1800 miles thick), and the crust. The core is made from superheated metals, and the mantle consists of semi-molten and semi-solid rock. On top of the mantle float the so-called tectonic plates (large semi-rigid slabs) that constitute the greater part of the earth's crust and have a thickness of only three to 45 miles. The Earth's continents are the visible surfaces (and land masses) of the plates. In most cases, the edges of the tectonic plates (i.e., the borders between them), which are regions of geological turbulence and create visible fault lines (breaches), extend along the shorelines. The tectonic plates move very slowly (typically less than five inches per year), but there are periods without movement which are signals of danger. This is because, when no or little movement occurs, energy accumulates and is stored that then presses on the plates. When the plates can no longer withstand the tension, they break down and an earthquake takes place. Earthquakes are usually measured using the *Richter logarithmic scale* (or its improvement, called the *moment magnitude scale*). On the average, there is a severe earthquake ($\sim$9.00 Richter), and about 150 moderate earthquakes ($\sim$6–8 Richter) per year (since 2000). Figure 2.7 shows the collapse of houses caused by a strong earthquake in the area of Kobe (Japan).

**Volcanoes** are typically located on the edges of Earth's tectonic plates (e.g., those of Iceland and Japan), but in many cases they are located within the body of a plate, e.g., those of Hawaii (Fig. 2.8). Actually, a volcano is the aperture from which magma and other materials erupt from the ground. But, in common terminology, people talking about volcanoes mean the area around the aperture, where the volcanic materials have solidified. Volcanic eruptions are violent demonstrations of Earth's energy–mass movement. Sometimes, volcanic eruptions start with emissions of steam and gases from small apertures in the ground, accompanied by a

**Fig. 2.7** Collapsed houses in Kobe caused by the 2002 earthquake (http://images.nationalgeographic.com/wpf/media-live/photos/000/002/cache/kobe-house-collapse_262_600x450.jpg)

**Fig. 2.8** The Kilauea volcano of Hawaii (http://www.solarnavigator.net/volcanoes.htm)

dense sequence of small earthquakes. In other cases, magma comes to the surface as fluid lava which either flows continuously or shoots straight up in the form of glowing fountains.

The study of volcanoes, their action, and their products is a multidisciplinary field called "*volcanology*" [69]. A list of the major volcanoes of the world is provided in [66]. It is noted that it may be possible to use nonexplosive volcanoes to harvest geothermal energy, as, e.g., has been done in Iceland [70].

### 2.7.2 Tornadoes and Hurricanes

Normal weather phenomena include wind, clouds, rain, snow, fog, and dust storms. Less common phenomena include tornadoes, hurricanes, and ice storms. Tornadoes and hurricanes are weather phenomena belonging to the class of *natural vortexes*. The atmosphere of Earth is a complex and chaotic system in which small changes somewhere can create large variations elsewhere, see Sect. 8.8 ("*Butterfly Effect*"). The weather is shaped in the stratosphere and affects the weather below it in the troposphere. However, due to the chaotic nature of the atmosphere, it is not possible to specify precisely how this occurs. A tornado is the result of a violent movement of energy (heat) when cool air and warm air meet, forcing warm air to rise very quickly. A tornado consists of a violent windstorm with a twisting, funnel-shaped cloud, usually followed by thunderstorms. Tornadoes in the *Fujita scale* are classified not by their size but by their intensity and the damage they cause. Large tornadoes may be weak and small ones may be strong. An example of a tornado is shown in Fig. 2.9.

*Hurricanes*, known as *typhoons* in Asia, are *tropical storms* or *cyclones* that are much broader than tornadoes (Fig. 2.10). Tropical cyclones are typically formed over ocean water heated to a temperature of 26 °C and lying within about 5° of latitude from the Equator, but they can occur at other places as well. The ocean water is heated and it evaporates taking energy from the ocean. As the warmed air rises, a vacuum is formed from the resulting low-pressure system, and thus tropical storms are created. The formation of tornadoes and hurricanes can be explained by Archimedes principle, the rotational force, and the Coriolis force [72–74]. In particular, the Coriolis force creates the rotation of air around the center (a calm area called the "*eye*") in a cyclonic direction (clockwise in the Southern Hemisphere and anti-clockwise in the Northern Hemisphere). The rising water vapor cools and condenses, releasing latent energy and warming further the surrounding air. Actually, this process is a "positive feedback" process that reinforces the existence of the phenomenon. According to K.M.I. Osborne [74], tropical cyclones help to

**Fig. 2.9** A tornado (http://www.chaseday.com/tornadoes-02.htm)

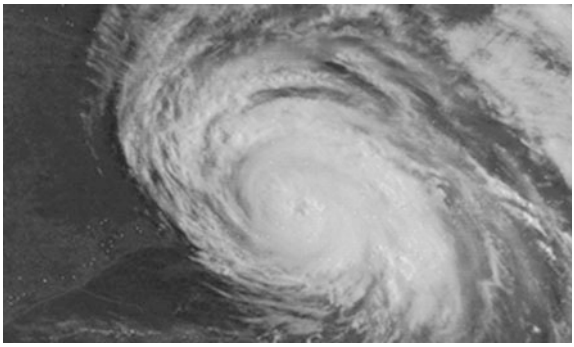


**Fig. 2.10** A satellite photo of a hurricane (Andrew) (http://ww2010.atmos.uiuc.edu/(Gh)/home.rxml)

cool the ocean by drawing heat out and converting it to wind (mechanical energy) and so ensuring that no area of the ocean becomes overheated. The increase of the overall amount of energy in the Earth's atmosphere caused by accumulating GHGs will be followed by the formation of more and more cyclones.

### *2.7.3 Tsunamis*

Tsunamis can be generated whenever large water masses are violently displaced from their equilibrium position, caused by any disturbance. The word tsunami is Japanese written in English and composed of the word "*tsu*" (meaning "*harbor*") and "*nami*" (meaning "*wave*"). Tsunamis are different from wind-generated waves since they have long periods and wavelengths like shallow-water waves. A wave turns out to be a shallow-water wave when the ratio between the water depth and its wavelength becomes very small. The traveling speed of shallow-water waves is equal to the product of the acceleration of gravity (9.81 m/s) and the water depth. For example, in the Pacific Ocean (with a typical depth 4000 m), a tsunami travels with speed about 200 m/s or greater than 700 km/h. The rate at which a wave loses its energy is inversely proportional to its wavelength. Thus, a tsunami can travel (at high speed) long distances with very small energy loss. Figure 2.11a illustrates how an earthquake generates a tsunami. The water column is disturbed by the uplift or subsidence of the seafloor. Figure 2.11b is a photo of an actual tsunami impinging on a shoreline in India.

The tsunami's energy flux (which depends on wave height and wave speed) is almost constant, and so, since the tsunami's speed decreases as it approaches the land, (i.e., water becomes shallower), its height grows to become several meters near the coast. Figure 2.12 shows the tsunami that occurred after an earthquake in Eastern Australia (near Solomon Islands).

A recent tsunami is the 7–20 m tsunami generated by the magnitude 9.0 quake that hit northeastern Japan on March 11, 2011. This tsunami reached Australia, North America, and South America after a few hours.

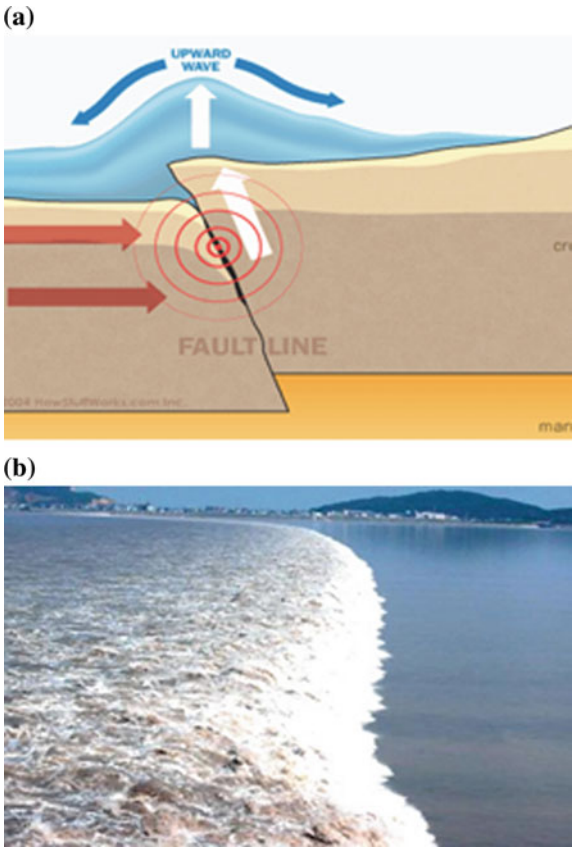



**Fig. 2.11** **a** Earthquake-tsunami genesis (http://static.howstuffworks.com **b** An actual earthquake-generated tsunami (http://www.snopes.com/photos/tsunami/tsunami1.asp

**Fig. 2.12** The Australian tsunami at Lord Howe and Norfolk Islands (http://livesaildie.com/2007/04/02/tsunami-threat-to-eastern-australia)

## 2.8 Concluding Remarks

Energy is the basis of everything and its movement or transformation is always followed by a certain event and dynamic process. For this reason, and the fact that energy cannot be formally defined, David Watson [9] calls energy the "*mysterious everything*". This chapter has discussed the fundamental aspects of energy, viz., the energy types (mechanical, electrical, chemical, and nuclear), the energy sources (exhaustible, non-exhaustible, alternative sources), and the impact of energy on the environment. An overview of the major, violent, natural phenomena of physical energy on our planet was also provided. The actual study of the energy movement and conversion, collectively called "*thermodynamics*", will be the subject of the next chapter. The role of energy in life and its flow in nature, society, and technology will be discussed in Chap. 10.

## References

1. C. Cleveland (ed.), *Encyclopedia of Energy* (Elsevier, Amsterdam, Netherlands, 2004)
2. H.M. Jones, *The Age of Energy* (Viking, New York, 1970)
3. V. Smil, *General Energetics* (MIT Press, Cambridge, MA, 1991)
4. R. Feynman, R. Leighton, M. Sands, The Feynman Lectures on Physics, vol. 3, 2nd edn. (Addison-Wesley, 2005). (Also: Wikipedia, The Feynman Lectures on Physics)
5. J.R. von Mayer, Remarks on the forces of inorganic nature. Annalen der Chemie and Pharmacie **43**, 233 (1842)
6. Buzzle.Com-Intelligent Life on the Web: Type of Energy. http://buzzle.com/articles/types-of-energy.html
7. http://en.wikipedia.org/wiki/Energy
8. The Energy Story, http://energyquest.ca.gov/chapter08.html
9. D. Watson, What is the Definition of Energy? http://www.ftexploring.com/energy/definition.html
10. C. LaRocco, B. Rothstein, The Big Bang: It Sure was Big. http://www.umich.edu/~gs265bigbang.htm
11. All About Science, Big Bang Theory An Overview. http://www.big-bang-theory.com/
12. X.T. Trinh, *The Birth of the Universe: The Big Bang and After* (Harry N. Abrams Inc, New York, 1993)
13. Wapedia, Physical Cosmology. http://wapedia.mobi/en/Physical_cosmology
14. W.J. Kaufmann III, *Galaxies and Quasars* (W.H. Freeman & Co., San Francisco, 1979)
15. M. Eastman, Cosmos and Creator.http://www.khouse.org/articles/1999/233/
16. The Energy Story, http://energyquest.ca.gov/story/chapter10.html
17. http://www.eohm.com/alternative_energy_definitions.html
18. http://www.jc-solarhomes.com/solar_hot_water.htm
19. http://en.wikipedia.org/wiki/Solar_collector
20. http://www.jc-solarhomes.com/fair/parabola20.htm
21. http://www.buzzle.com/articles/renewable-sources-of-electricity.html
22. V.M. Faires, *Theory and Practice of Heat Engines* (MacMillan, New York, 1955)
23. V.M. Faires, C.M. Simmang, *Thermodynamics*, 6th edn. (Macmillan, New York, 1978)
24. D. Watson, Energy Changes Make Things Happen? http://www.ftexploring.com/energy/energy-1.htm

25. Wikipedia, Timeline of Thermodynamics, Statistical Mechanics, and Random Processes (Article ID: 58777)
26. Wapedia, Wiki History of Thermodynamics. http://wapedia.mobi.en/History_of_thermodynamics
27. http://www.eoht.info/page/Energy
28. http://www.upscale.utoronto.ca/PVB/Harrison/ConceptOfEnergy/ConceptOfEnergy.html
29. I. Mueller, *History of Thermodynamics: The Doctrine of Energy and Entropy* (Springer, New York, 2007)
30. http://physics.uoregon.edu/∼soper/Light/photons.html
31. M. Alonso, E.J. Finn, Fundamental University Physics, vol. III (Addison-Wesley, Reading, MA, 1968)
32. http://www.masstech.org/cleanenergy/wavetidal/overview.htm
33. G.N. Lewis, M. Randall, *Thermodynamics (revised by K Pitzer and LBrewer)* (McGraw-Hill, New York, 1961)
34. Internal Energy, http://www.answers.com/topic/internal-energy
35. N. Sato, *Chemical Energy and Exergy: An Introduction to Chemical Thermodynamics for Engineers* (Elsevier, Amsterdam, 2004)
36. S. Skogestad, Chemical and Energy Process Engineering (CRC Press, Taylor & Francis Group, Boca Raton, 2008)
37. Committee for the Workshop on Energy and Transportation, *Energy and Transportation: Challenges for the Chemical Sciences in the 21st Century* (National Academies Press, Washington, DC, 2003)
38. P. Perrot, *A to Z of Thermodynamics* (Oxford University Press, Oxford, 1998)
39. http://www.britannica.com/EBchecked/topic/108679/chemical-energy
40. http://www.absoluteastronomy.com/topics/Sensible_heat
41. http://www.absoluteastronomy.com/topics/Latent_heat
42. http://www.ifpaenergyconference.com/Chemical-Energy.html
43. http://www.energyquest.ca.gov/story/chapter13.html
44. http://www.library.thinkquest.org/3471/nuclear_energy_body.html
45. J. Daintith, Oxford Dictionary of Chemistry (Oxford University Press, Oxford, 2004)
46. W.F. Magie, A Source Book in Physics (Contains excerpts on Latent Heat, by J. Robinson, University of Edinburgh (1803) from J. Black, Lectures on the Elements of Chemistry (1803), (McGraw-Hill, New York, 1935)
47. http://www.masstech.org/cleanenergy/wavetidal/overview.htm
48. http://tonto.eia.doe.gov/kids/energy.cfm?page=hydrogen_home_basics
49. http://www.uwsp.edu/cnr/wcee/keep/Mod1/Whatis/energyforms.htm
50. http://www.ucmp.berkeley.edu/help/timeform.html
51. http://www.world-nuclear.org/info/inf03.html
52. http://environment.about.com/od/renewableenergy/tp/renew_energy.htm
53. Minerals Management Service—US Department of the Interior, Ocean Energy. http://www.doi.gov/, http://www.mms.gov/mmskids/PDFs/OceanEnergyMMS.pdf
54. About.com, Hybrid Cars & Alt Fuels. http://alternativefuels.about.com/od/resources/a/energy.htm
55. http://www.hawaii.gov/dbedt/ert/otec-nelha/otec.html
56. Fossil Fuel and its Impact on the Environment, http://www.essortment.com/all/fossilfuelimpa_rhxu.htm
57. Other Environmental Impacts of Fossil Fuels,http://www.masstech.org/cleanenergy/important/envother.htm
58. http://www.ucsusa.org/clean_energy/technology_and_impacts/environmental-impacts-of.html
59. J.H. Gibbons, P.D. Blair, H.L. Gwin, Strategies for energy use. Sci. Am **261**(3), 136–143 (1989)
60. R.D. Weber, Energy Information Guide: Fossil Fuels, Energy Information Pr
61. C. Twist, Facts on Fossil Fuels (Gralier Educational Associates, 1998)

62. P. Goodman, Fossil Fuels: Looking at Energy (Hodder Wayland, Paris, France, 2001)
63. M.A. Rossen, Utilization of Non-Fossil Fuel Energy Options to Mitigate Climate Change and Environmental Impact (Plenary Lecture). http://www.wseas.us/conferences/2009/cambridge/ee/Plenary1.htm
64. W.H. Kemp, The Renewable Energy Handbook: A Guide to Rural Independence, Off-Grid and Sustainable Living (Gazelle Drake Publishing, Cardiff, UK, 2006)
65. G. Sharpe, J. Hendee, W. Sharpe, *Introduction to Forest and Renewable Resources* (McGraw-Hill, New York, 2002)
66. http://www.britannica.com/EBchecked/topic/632130/volcano
67. http://www.realtruth.org/articles/090907-006-weather.html
68. http://www.spacegrant.hawai.edu. (volcanoes)
69. http://en.wikipedia.org/wiki/Volcanology
70. http://answers.yahoo.com/question/Index?qid=20080619033354AAeGOq2,Also: 20080314213713AAeGOq2
71. Meteorology, http://www.sciencedaily.com/articles/w/weather.htm
72. The Physics of Tornadoes and Hurricanes, http://www.physics.ubc.ca/outreach/phys420/p420_04/sean/
73 http://www2010.atmos.uiuc.edu/(Gh)/guides/mtr/hurr.home.rxml
74. K.M.J. Osborne, The Physics of Hurricanes: Understanding Tropical Cyclones. http://physics.suite101.com/article.cfm/the_physics_of_hurricanes

# Chapter 3
# Energy II: Thermodynamics

*The law that entropy always increases—the second law of thermodynamics—holds, I think, the supreme position among the laws.*

Arthur Stanley Eddington

*My position is perfectly definite. Gravitation, motion, heat, light, electricity, and chemical action are one and the same object in various forms of manifestation.*

Julius Robert Mayer

**Abstract** Broadly speaking, thermodynamics is the study of the relation of heat and other forms of energy (mechanical, electrical, radiant, etc), and the conversion of one form to another, as well as their relation to matter in the Universe. This chapter gives an overview of the major concepts, laws, and branches of thermodynamics that have been developed and studied over the years since the Carnot times. Specifically, this chapter defines the basic physical concepts of thermodynamics, with emphasis on the fundamental concept of entropy, and presents the four laws of thermodynamics. Particular aspects studied are the entropy interpretations (unavailable energy, disorder, energy dispersal, opposite to potential), the Maxwell demon, and the types of arrow of time (psychological, thermodynamic, cosmological, quantum, electromagnetic, causal, and helical arrows). This chapter ends with a number of seminal quotes on thermodynamics, entropy, and life that express the opinions of the founders and other eminent contributors and thinkers in the thermodynamics field.

**Keywords** Thermodynamics · Thermodynamics branches · Thermodynamic equilibrium · Thermodynamics laws · Entropy · Enthalpy · Exergy Statistical entropy · Quantum mechanics entropy · Non-statistical physics entropy Entropy interpretations · Maxwell's demon · Arrow of time

## 3.1 Introduction

*Thermodynamics* is the field of science that investigates the phenomena of energy movement in natural and man-made systems. Historically, it was developed as a science in the eighteenth and nineteenth centuries. In 1798, Count-Rumford



S. G. Tzafestas, *Energy, Information, Feedback, Adaptation, and Self-organization*, Intelligent Systems, Control and Automation: Science and Engineering 90, https://doi.org/10.1007/978-3-319-66999-1_3

(Benjamin Thomson) initiated the investigation of the conversion of work into heat through his cannon-boring experiments and Sir Humphry-Davy did the same via his ice-rubbing experiments. The concept of entropy was discovered by Sadi Carnot (1824), and the term *thermodynamics* [in the forms thermodynamic (1849) and thermodynamics (1854)] was coined by William Thomson (Lord Kelvin) in the framework of his studies of the heat engines' efficiency [1]. The name "thermodynamics" comes from the Greek words "Θερμο" (thermo = heat) and "δυναμική" (dynamics = study of forces, power, and work).

Expressed in another way, thermodynamics is the study of the relation of heat and other forms of energy (electrical, mechanical, radiant, etc.) and the conversion of one form into another, as well as their relationship to matter in the universe.

In the twentieth century, thermodynamics has evolved into a fundamental branch of physics dealing with the study of the equilibrium properties of physical systems and the phenomena occurring during the transition toward equilibrium. Thermodynamics is presently used in many areas of science and life from physics and engineering, to biochemistry, computer science, human chemistry, business, and cosmology.

The methodological approaches of thermodynamics include the following [2–15]

- **Classical macroscopic approach** which starts from principles that are established by experiment and develops further using standard properties of physical objects and substances (Galilei, Black, Carnot, von Mayer, Gibbs, Clapeyron, Joule, Helmholtz, Thomson, Clausius, Planck, Nernst).
- **Statistical microscopic approach** which is based on statistical mechanics in which the thermodynamic systems are investigated through the statistical performance of their ingredients (Maxwell, Boltzmann, Max Planck) [16–18].
- **Quantum-mechanics approach** which is based on the density matrix (operator) concept of a mixed quantum system (Von Neumann) [19].
- **Axiomatic approach** in which all postulates of thermodynamics are derived from a set of basic mathematical axioms (Caratheodory) [12].
- **Non-statistical general theory of physics approach** in which entropy is considered to be an inherent, non-statistical property of any system (small or large) in any state (thermodynamic equilibrium or non-equilibrium), or a pure microscopic non-statistical property of matter (Gyftopoulos, Beretta) [10].

The objectives of this chapter are:

- To present the four laws of thermodynamics and the definitions of entropy: classical, statistical, quantum-mechanics, non-statistical, Renyi, and Tsallis formulations.
- To outline the branches of thermodynamics, namely, traditional, natural systems, and modern branches.
- To discuss entropy interpretations, namely: entropy as unavailable energy, entropy as disorder, entropy as energy dispersal, and entropy as the opposite of potential.
- To review the Maxwell's Demon and its exorcisms and discuss the concept of time's arrow, including all of its known types; psychological, thermodynamic, cosmological, quantum, electromagnetic, causal, and helical.

- To list a number of important quotations and statements about thermodynamics, entropy, and life, as expressed by established and newer thinkers in the field.

## 3.2 Basic Concepts of Thermodynamics

As a preparation for the material to be presented in this chapter, the following basic concepts of physics and thermodynamics are first introduced [3–10, 20–26]:

- Intensive and extensive properties
- System
- System state
- Universe
- Thermodynamic equilibrium
- Temperature
- Pressure
- Heat and specific heat
- Reversible and irreversible process
- Categories of thermodynamic processes
- Basic concepts of non-statistical general physics.

### *3.2.1 Intensive and Extensive Properties*

A physical property (or quantity or variable) is called an *intensive* (*or bulk*) *property* (or quantity or variable) if it is independent of the size of the system or the amount of material contained in it, i.e., if it is *scale-invariant*. Otherwise, if it depends on the size of the system or the amount of the material contained in it, is said to be an *extensive property* (or quantity or variable).

Examples of intensive and extensive properties are given in Table 3.1.

The ratio of two extensive quantities is an intensive quantity. For example: "Density = mass/volume".

All the extensive properties of Table 3.1, except volume, depend on the amount of material, and so, in the case of homogeneous substances if expressed on a per mass basis, yield corresponding intensive properties (e.g., specific internal energy, specific entropy, specific heat capacity at constant volume or at constant pressure, etc.).

Actually, an extensive property of a system is the sum of the respective properties of all subsystems that are contained in the system. For example, the internal energy of a system is equal to the sum of the internal energies of all its subsystems, and the same is true for the volume, etc.

**Table 3.1** A partial list of basic intensive and extensive properties

| Intensive properties | Particle number |
|---|---|
| Density | Specific gravity |
| Concentration | Specific energy |
| Temperature | Specific heat capacity |
| Pressure | -At constant pressure |
| Velocity | -At constant volume |
| Elasticity | Chemical potential |
| Viscosity | Electrical resistivity |
| Melting point | Boiling point |
| Extensive properties | Particle number |
| Mass | Internal energy |
| Volume | Heat capacity |
| Entropy | -At constant pressure |
| Enthalpy | -At constant volume |
| Exergy | Gibbs free energy |
| Stiffness | Helmholtz free energy |

### 3.2.2 System and Universe

*System* in conventional thermodynamics is any fixed mass of a pure substance bounded by a closed, impenetrable, and flexible surface (e.g., a collection of molecules of air, water, or combustion gas contained in a cylinder with a fitted piston). A *universe* is an *isolated system* since no energy or matter can go in or out (Fig. 3.1).

A *closed system* can exchange energy but no matter. An *open system* can exchange both energy and matter with its exterior. The system itself in Fig. 3.1 may exchange energy and/or matter with its surrounding through the boundary. It is remarked that the term "*universe*" is used in thermodynamics not in the cosmological sense, but simply to incorporate the system and all matter of its environment that may interact with the system. Synonymous terms for the "system" concept used throughout the history of thermodynamics are "*working substance*" (Carnot), "*working body*" (Clausius), and "*working medium*" (Ksenzhek).

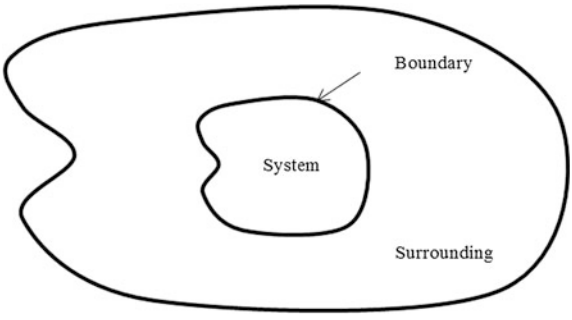



**Fig. 3.1** An isolated system (universe) composed of an open system and its surrounding

### 3.2.3 System State

The *state* of a system is specified by the values of all the variables that describe the performance of the system. These variables are usually dependent on each other. This dependence is described by one or more equations that are called "*state equations*". In the case of a gas, the state variables are the temperature $T$, the pressure $p$, the volume $V$, and the number of moles $n$ in the gas. The experiment showed that only three of these four variables are independent, and so we must have a single state equation. The best-known state equation of an ideal gas (where the molecules are assumed point masses, and there are no intermolecular forces) is:

$$pV = nRT \tag{3.1}$$

where $R$ is the so-called "*universal gas constant*" ($R$ = 8.314472 J $K^{-1}$ $mol^{-1}$).

Actually, no real gas is governed by the ideal gas equation for all temperatures and pressures, but in the limit as pressure tends to zero. This is because the effective pressure of the gas is higher than the measured pressure p and is equal to:

$$p_{\text{eff}} = p + an^2/V^2 \tag{3.2}$$

due to the intermolecular forces that attract the peripheral molecules to the interior, which are proportional to $n^2/V^2$.

Likewise, the effective volume $V_{\text{eff}}$ is smaller than $V$ because of the volume occupied by the molecules themselves, i.e.:

$$V_{\text{eff}} = V - bn \tag{3.3}$$

Therefore, the ideal gas equation now becomes:

$$p_{\text{eff}} V_{\text{eff}} = (p + an^2/V^2)(V - bn) = nRT \tag{3.4}$$

where the parameters $a$ and $b$ for each gas are determined experimentally. This equation is known as the *van der Waals equation*.

### 3.2.4 Thermodynamic Equilibrium

We say that a system is in *thermodynamic equilibrium* if no state change occurs for a long period of time. In general, the state (or condition) of a system is identified by the values of temperature, pressure, and volume (see below), which are the macroscopic thermodynamic properties.

An equilibrium state of a system consisting of a *single phase* of a pure substance can be determined by two thermodynamic properties, e.g., temperature and pressure.

### 3.2.5 Temperature and Pressure

Temperature $T$ is a physical property that provides a measure of the average kinetic energy of the atoms contained in any physical object. High average kinetic energy, i.e., fast motion of the atoms, implies high temperature and vice versa. Actually, temperature expresses the tendency of the system or object to provide or release energy spontaneously. The upper bound of temperature is the temperature at the Big Bang, and the lower bound is the absolute zero (see Sect. 3.7).

The temperature of an object (system) is measured by a *thermometer* that is brought in intimate and prolonged contact with the system, such that the object and the thermometer are in thermodynamic equilibrium. For example, a mercury-in-glass thermometer operates on the basis of thermal expansion or contraction of mercury within a glass bulb. After the equilibrium is achieved, the length of the narrow column of mercury connected to the bulb indicates the actual temperature of the object.

The *pressure* is a thermodynamic property that provides an alternative way to measure the changes in the state of a liquid or gaseous system. To this end, a *manometer* is connected to the system and the level of the free surface of its fluid is observed. The free surface of the manometer goes up or down according to the increase or decrease of the pressure (force per unit area) acting on the manometer interface.

As we saw in Sect. 3.2.1, the temperature and pressure are intensive physical properties (i.e., they do not change with the mass of the system), but can vary from point to point in the system. For example, a thermometer placed at different positions in the system may indicate different temperatures. Only if the system is in thermodynamic equilibrium are the temperatures at all points of the system the same and equal to the unique temperature of the system. This means that a system has a *unique temperature* only if it is in *equilibrium*.

### 3.2.6 Heat and Specific Heat

Heat, denoted by $Q$, is the *thermal energy* of a system. Therefore heat has the dimensions of energy and is measured in "*calories*". The "*calorie*" is defined as the amount of heat needed to heat 1 gr of water by one degree Celsius. The relation of calorie and Joule (mechanical equivalent of heat) is:

$$1\,\text{calorie} = 4.186\,\text{Joule}$$

The heat transfer to or from a system can change the state of the system and does the work. Two systems (or objects) in contact are said to be in thermodynamic equilibrium, or simply in equilibrium if they have the same temperature. If they

have different temperatures, then they are not in equilibrium, because heat will flow from the object of higher temperature to the object that has a lower temperature, until the two objects reach a common temperature, i.e., until they reach thermodynamic equilibrium. As described in Sect. 2.4.3.2, the heat flow that enters or leaves a system and does not lead to a temperature change in it, e.g., the heat flows that accompany phase alterations, such as boiling or freezing, are the so-called *latent heat* (the amount of heat absorbed or released by a substance during a state change).

The relation of heat, mass, and temperature is the so-called "specific heat" formula:

$$Q = mc\Delta \mathrm{T} \tag{3.5}$$

where $Q$ is the amount of heat, $m$ is the mass, $c$ is the *specific heat* of the material, and $\Delta T$ the temperature change. The quantity:

$$C = mc \tag{3.6}$$

is called *heat capacity* of the substance of concern with mass $m$. The field of measuring the amount of heat involved in thermodynamic processes (e.g., chemical reactions, state changes, mixing compounds, change of phase, etc.), or the heat capacities of substances is called "*calorimetry*". The instruments that are used to quantitatively measure the heat required or generated during a process are called "*calorimeters*" which may be of one the following types: constant volume (or "*bomb*") calorimeters, adiabatic calorimeters, constant pressure calorimeters, isothermal titration calorimeters, X-ray microcalorimeters, etc. [27]. To determine the heat capacity of a calorimeter, we transfer a known quantity of heat into it and measure its temperature increase via a suitably sensitive thermometer.

### 3.2.7 Reversible and Irreversible Process

The concept of a "reversible process" is best illustrated by a gas inside a frictionless piston in contact with an unlimited surrounding of temperature $T$ (known as "heat reservoir", Fig. 3.2).

The "gas-piston-heat reservoir" system is a "thermodynamic universe", as described in Sect. 3.2.2.

By pushing very slowly on the piston in small steps, the gas will compress at constant temperature $T$, and pulling it back, slowly as before, the gas will expand and return to its initial state. This procedure is defined to be a *reversible process* and is characterized by thermodynamic equilibrium at each successive step in both the compression and expansion cases. This is so because a small step in one direction (say compression) can be exactly reversed by a similar small step in the opposite direction (i.e., expansion). If the piston is moved sharply, then a turbulent gas

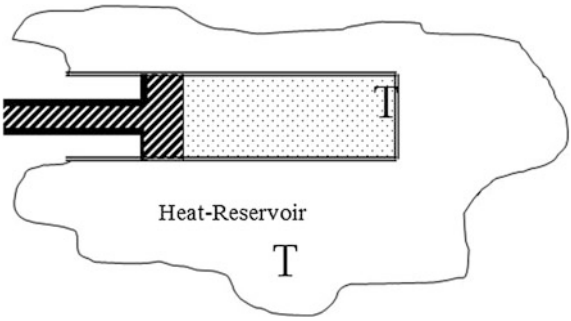



**Fig. 3.2** Gas in a piston for illustrating a reversible process

motion takes place that cannot be done in the opposite direction in the same way, i.e., passing through the same consecutive gas states. This process is called an *irreversible process*.

In general, if the temperature and pressure changes (gradients) in a system process are always small, this process can be regarded as a sequence of *near-equilibrium* states. If all these states can be restored in reverse order, the process is called an *internally reversible* process. If the same is true for the system surrounding, then the process is said to be *externally reversible*. The process is said to be a *reversible process* if it is both internally and externally reversible. Actually, in reality, all processes are irreversible because they do not satisfy the reversibility requirements. Irreversibility is due to pressure, temperature, velocity, and composition gradients caused by chemical reaction, heat transfer, fluid and solid friction, and high work rates exerted on the system.

### 3.2.8 Categories of Thermodynamic Processes

The thermodynamic processes are distinguished in the following principal categories (Fig. 3.3):

- Adiabatic processes
- Isothermal processes
- Isochoric processes
- Isobaric processes
- Isentropic processes

An **adiabatic** (**or isocaloric**) **process** is a process in which no heat transfer takes place into or out of the system. The term adiabatic comes from the Greek composite word "αδιαβατικός" ("a" = not, "δια" = through, "βατικός" = passing) and literally, means "*impassable*", i.e., heat not passing (not transferred). Since no heat transfer occurs, it is also called *isocaloric* (with equal calories), although there also exist "isocaloric" processes that are not adiabatic.

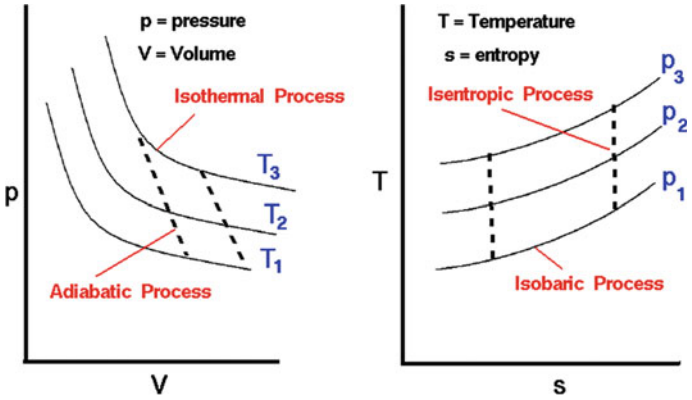



**Fig. 3.3** Categories of thermodynamic processes. (http://www.grc.nasa.gov/WWW/K-12/airplane/Images/pvtsplot.gif (*)) (The reader is informed that Web figures and references were collected at the time of the writing the book. Since then, some may no longer be valid due to change or removal by their creators, and so they may no longer be useful)

The state equation for a reversible adiabatic process of an ideal gas is:

$$PV^{\gamma} = \text{constant} \tag{3.7}$$

where $\gamma$ is the *adiabatic index* given by:

$$\gamma = c_p/c_v = (\alpha + 1)/\alpha \tag{3.8}$$

Here, $c_p$ and $c_v$ are the specific heats of the gas for constant pressure and constant volume, and "$\alpha$" is the number of degrees of freedom divided by 2, i.e., $\alpha = 3/2$ for monatomic gas and $\alpha = 5/2$ for a diatomic gas. A reversible adiabatic process occurs at constant entropy and is called an *isentropic process*.

An **isothermal process** is a process without any change in temperature. To assure that the temperature is kept constant, the system must be strongly insulated or the heat transfer into or out of the system must take place at a sufficiently slow rate so that thermal equilibrium is maintained. During an isothermal process, there is a change in internal energy, heat, and work.

An **isochoric process** is a process with no change in volume and so no work done by the system. The term *isochoric* comes from the Greek "*ισοχωρικός*" (ισο (iso) = equal + χωρικός = space/volume). To keep the volume constant is easily done by placing the system in a sealed container that does not expand or contract.

An **isobaric process** is a process where no change in pressure occurs [the process takes place while maintaining constant pressure units (bars)].

An **isentropic process** is a process in which the entropy is kept constant.

In practice, we may have more than one type of processes within a single process, such as, e.g., in the case where both volume and pressure change such that no change in temperature or heat occurs (adiabatic and isothermal process).

### 3.2.9 Basic Concepts of Non-statistical General Physics

Gyftopoulos, Berreta, and Hatsopoulos in a series of publications [10, 28–33] have shown that thermodynamics is a well-founded non-statistical theory of physics, and they presented two novel avenues for formulating the entropy concepts and the first two laws of thermodynamics. As preparation, we briefly outline here the definitions of the basic elements of this approach (system, properties, state, energy, equation of motion, types of states).

- **System**: A collection of constituents (particles and radiations) subject to both internal and external forces possibly depending on geometrical features, but independent of coordinates of constituents not belonging to the collection at hand.
  It is assumed that the system is both *separable* and *uncorrelated* with its environment, i.e., the environment includes everything that is not contained in the system. The amounts of the r constituents and the s parameters of the external forces are given by the vectors $\mathbf{n} = [n_1, n_2, \ldots, n_r]$ and $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_s]$, respectively. Examples of parameters are the volume ($\beta_1 = V$) and the potential of an external (electrical, magnetic, gravitational) field ($\beta_2 = \varphi$).
- **Properties**: The condition of the system at some instant in time needs also the knowledge of the values of a complete set of independent attributes that are called *properties* and are determined by proper measurement operations or procedures at each time.
- **State**: The state of the system at a given instant consists of the values of the amounts of the constituents, the values of the parameters, and the values of a complete set of properties, under the assumption that the measurement results do not have any correlation with measurements in any other systems in its environment. The system state may change over time as a result of internal forces (i.e., spontaneously) or because of interactions with other systems or both.
- **Energy**: The energy of any well-defined system $A$ in any well-defined state $A_1$ is a property denoted by $E_1$, which is evaluated by a weighing process connecting the state $A_1$ to a reference state $A_0$ that has an energy $E_0$ (arbitrarily assigned), i.e.,

$$E_1 = E_0 + mg(z_1 - z_0) \tag{3.9}$$

where $m$ is the mass of the weight, $g$ the gravitational acceleration, and $z_1$, $z_0$ the corresponding elevations. Energy is additive, i.e., the energy of a composite sum is the sum of the subsystems' energies. The energy change of a system is positive if it

is transferred from the environment into the system, symbolically denoted by an arrow as:

$$(E_2 - E_1)_{\text{system}A} = E^{A\leftarrow} \tag{3.10}$$

- **Equation of motion**: This is the relationship that describes the time evolution of the system's state (spontaneous, or forced, or both) subject to the condition that the external forces do not violate the definition of system. Examples of equations of motion are: Newton's equation $F = ma$, and Schrödinger's equation (i.e., the quantum-mechanical analog of Newton's or Hamilton's equations). These equations, and particularly Schrodinger's equation, are applied to reversible processes that evolve unitarily. Because not all physical processes are reversible and not all processes evolve unitarily, Gyftopoulos and Berreta developed a complete equation of motion which is applied to all cases.
- **Types of states**: The states are classified according to their time evolution in: *unsteady* and *steady states*, *non-equilibrium states* (i.e., states that change spontaneously), *equilibrium states* (i.e., states not changing while the system is isolated), *unstable equilibrium states* (i.e., equilibrium states that may be caused to act spontaneously by short-term interactions having zero or infinitesimal effect on the system's environment), and *stable equilibrium states* (i.e., states that can be changed only by interactions not leaving any net effect in the system's environment). These state concepts seem identical to those of mechanics, but here they include a much wider spectrum of states due to the first and second laws of thermodynamics. Experience has shown that starting from an unstable equilibrium or from a non-equilibrium state, energy can be moved out of the system causing a mechanical effect but not leaving any other net change in the environment's state. However, starting from a stable equilibrium state, the above mechanical effect is not produced. This is a striking outcome of the first and second laws of thermodynamics.

## 3.3 The Zeroth Law of Thermodynamics

Over the years, the laws of thermodynamics have been formulated in a variety of ways. The Institute of Human Thermodynamics collects ten variations of the zeroth law, 40 variations of the first law, 125 variations of the second law, 30 variations of the third law, and 20 variations of the fourth law [34]. In particular, the second law and the interpretations of the concept of entropy, which refers to changes in the status quo of a system, has created a long-lasting scientific debate with many different and opposing opinions and misrepresentations [35, 36].

The *zeroth law of thermodynamics*, which is known as the law (or principle) of *thermodynamical equilibrium,* has its origin in the work of the Scottish physicist Joseph Black at the end of the eighteenth century [37–39]. He observed the tendency of heat to diffuse itself from any hotter body to a cooler one nearby until it is distributed among them such that none are disposed to take more heat from the others. The heat is therefore brought into equilibrium. Black continued his argumentation by saying that "one of the most general laws of heat is that all bodies communicating freely with each other, without being subject to unequal external action, acquire the same temperature (as indicated by a thermometer), namely the temperature of the surrounding medium." This is actually a precursor of what is now considered as the "*combined law of thermodynamics*" [40].

Most modern textbooks of thermodynamics use either Maxwell's formulation (1872) or Adkins' formulation (1983), as follows:

**Maxwell**
*Two systems A and B in thermal equilibrium with a third system C are in thermal equilibrium with each another.*

**Adkins**
*If two systems are separately in thermal equilibrium with a third, then they must also be in thermal equilibrium with each other.*

Although the *zeroth law* underlies the definition of temperature and asserts the fundamental fact of thermodynamics that, when two systems are brought into contact, an exchange of energy between them takes place until they are in thermal equilibrium (i.e., until they reach a common temperature), it was only stated explicitly much later at the time when the first three laws were formulated and widely used. This is why Ralph Fowler coined the term *zeroth law* in the 1920s. Clearly, "thermal equilibrium" is a binary relation that is both a *transitive* and an *equivalence* relation.

## 3.4 The First Law of Thermodynamics

### *3.4.1 Formulation of the Law*

The *first law* of thermodynamics is widely known as the "*law of conservation of energy*", and states:

> "*Energy can be moved from one system (or place) to another in many forms, but the total amount does not change, i.e., energy cannot be created or destroyed.*"

In other words, energy is conserved during any and every event, happening, process or transformation. Energy cannot come into existence from anywhere and cannot go out of existence into anywhere. The most general formulation of the first law is due to Clausius (1865) and states:

"*The energy of the World is constant*"

Historically, the first law was firstly informally stated by Germain Hess (1840) [34] in the form:

"*The heat absorbed or evolved in any chemical reaction is a fixed quantity and is independent of the path of the reaction or the number of steps taken to obtain the reaction*", which is known as *Hess's law*.

Later, Julius Robert von Mayer (1841) stated that "*Energy can be neither created nor destroyed*", and James Joule formalized the equivalence of mechanical work and heat with his experiments on the consequences of friction (1843). The first law was explicitly formulated for the first time in 1850 by Rudolf Clausius in the form [2, 34, 41]:

"*There is a state function E, called energy, whose differential equals the work exchanged with the surroundings during an adiabatic process.*"

In the exposition of thermodynamics by Gyftopoulos and Berreta via the non-statistical general physics theory, which is applicable to all systems (large and small including one- particle or one-spin systems), the first law is stated as follows [10]:

"*Any two states $A_1$ and $A_2$ of a system A may always be the initial and final state of a weight process in which the only effect external to the system is the change $z_1 - z_2$ of the weight's elevation from the initial to the final state (a purely mechanical effect).*"

This means that, for a given weight "mg", the value of the quantity mg $(z_1 - z_2)$ is determined only by the end states of the system. Of course, instead of the weight-elevation change, one can use many other effects in the statement of the first law. As a result of the above first law, many theorems can be rigorously proven for a system $A$ such as: (i) To each system's state there corresponds a function $E$, called energy, the change of which from state $A_1$ to $A_2$ is proportional to $z_2 - z_1$; (ii) During any spontaneous changes of state (occurring in an isolated system), E remains constant; and (iii) During interactions, the energy change $E_2 - E_1$ is balanced by the energy exchanged with the systems that interact with the system $A$.

It is remarked that the first law does not provide any means to *quantify* the effects of friction and dissipation.

Another related physical law states that "*matter cannot be created or destroyed*" and so the amount of matter in the universe is fixed (*conservation of matter*). But according to Einstein's equivalence of matter (mass) and energy expressed by the equation:

$$E = mc^2,$$

where $E$ is the energy contained in the mass $m$ and $c$ is the velocity of light in vacuum, the conservation of energy law is essentially the same as the conservation of matter law [42, 43]. Therefore, overall, "*the total amount of matter* (*mass*) *and energy available in the Universe is constant* [44]. Conversion of one type of matter

into another type is always accompanied by the conversion of one form of energy into another. Usually, heat is leveled or absorbed. Of course, many energy transformations do not involve chemical changes. For example, electrical energy can be converted into mechanical, heat, or light energy, without any chemical change. Mechanical energy is converted into electrical energy in a generator. Potential and kinetic energy can be converted into one another. Many other conversions are possible, but all of the energy and mass involved in any change always appears in some form after the change is completed.

### 3.4.2 The Thermodynamic Identity: Energy Balance

According to the *first law*, in any movement of energy, "energy entering a system" is equal to "energy stored" plus "energy going out". For the human body, this is illustrated in Fig. 3.4. Figure 3.5 is a diagram of the process of maintaining the normal temperature in the human body through energy balance.

Let $\Delta U$ be the change in *internal energy* of a system during a process, $Q$ the heat *flowing into* the system ($Q > 0$) or *flowing out of* the system ($Q < 0$), W the *mechanical work is* done on the system, and $W'$ any other energy added to the system. Then, by the first law we obtain the following *thermodynamic identity* or *energy balance* relation:

$$\Delta U = Q + W + W' \tag{3.11}$$

In Eq. (3.11), the work done on the system has a positive sign ($W > 0$), and the work done by the system has a negative sign ($W < 0$). The same holds also for $W'$. The standard unit for all quantities $U$, $Q$, $W$, and $W'$ is the "joule" (in SI system), but in many cases, the units used may be the "calorie" or the British Thermal Unit (BTU).

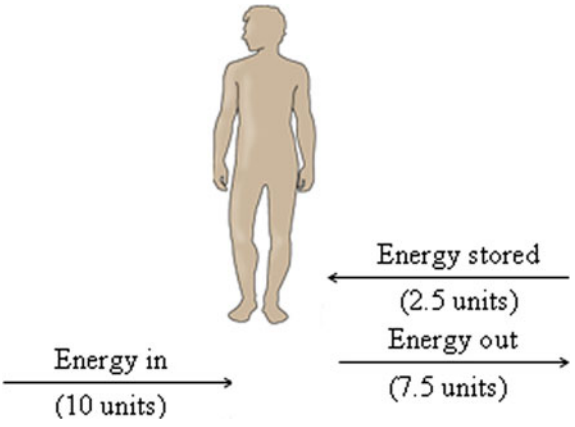



**Fig. 3.4** Illustration of the first law acting on the human body

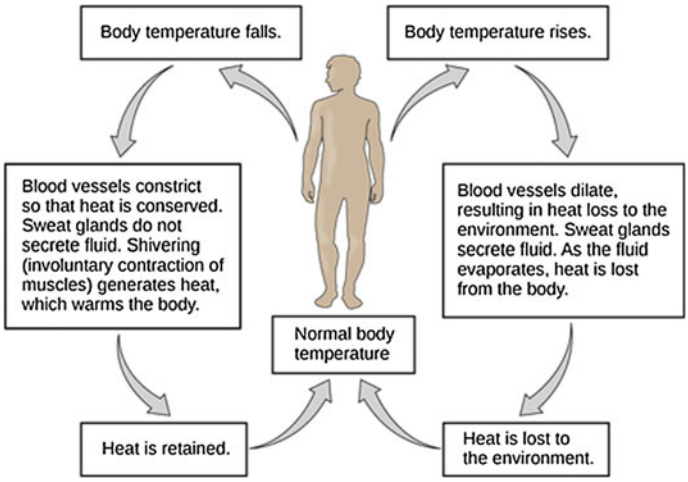



**Fig. 3.5** Human-body temperature regulation through energy balance. (http://cnx.org/resources/e685e1239aa6339d60f40ad3654e7c4d17c8e2e8/Figure_16_01_01.png)

In infinitesimal form, Eq. (3.11) can be written as:

$$\mathrm{d}U = \delta Q + \delta W + \delta W' \tag{3.12}$$

where the symbol "d" denotes an exact differential and "δ" indicates an infinitesimal increment which is not an exact differential. It is recalled that the integral of the exact differential of a quantity is given by the difference of the values of this quantity at its limits, while is not true for an inexact differential where the value of the integral depends on the path of integration. Thus we have:

$$\int_{U_i}^{U_f} \mathrm{d}U = U_f - U_i \tag{3.13}$$

The exact differential applies to physical properties (quantities) that are state functions of a system and characterize the state of the system as does the internal energy.

Now, consider a non-viscous fluid, for which the mechanical work is done on the system is given by

$$\delta W = p\mathrm{d}V \tag{3.14}$$

where $p$ is the pressure and $V$ is the volume of the fluid. Note that $p$ and $V$ are the so-called *generalized force and generalized displacement, respectively*. Referring to the "gas-in-piston system" of Fig. 3.2, Eq. (3.13) follows directly as:

$$\delta W = F\mathrm{d}x = pA\mathrm{d}x = p\mathrm{d}V \tag{3.15}$$

where $F$ is the force on the piston, $A$ is the area of the piston, and $\mathrm{d}V = A\mathrm{d}x$ is the differential volume change of the gas. Using Eqs. (3.12) and (3.14) and assuming that $\delta W' = 0$, in (3.12), we get:

$$\mathrm{d}U = \delta Q - p\mathrm{d}V \tag{3.16}$$

where the default direction of work was taken from the working fluid to the surrounding (e.g., the gas loses energy equal to the work it does on the surroundings).

**Definition**: A system which, after undergoing arbitrary change due to heat and work, returns to its initial state, is said to have participated in a *cyclic process*.

For a cyclic process (in which the final value $U_f$ of $U$ is equal to the initial value $U_i$), Eq. (3.13) becomes:

$$\oint \mathrm{d}U = U_f - U_i = 0 \tag{3.17}$$

where the special integral symbol indicates integration over a single cycle. Applying

Equations (3.17)–(3.12) (with $\delta W' = 0$, without loss of generality) gives:

$$\oint \delta Q = -\oint \delta W \tag{3.18}$$

where $-\delta W$ is work done on the system. Typically, most heat engines operate on cyclic processes, and it is in many cases convenient to calculate the net work of a cycle using Eq. (3.18) with heat additions and losses instead of using work directly.

For an arbitrary (non-cyclic) process, we get:

$$Q = U_f - U_i - W \tag{3.19}$$

where $Q$ and $-W$, respectively, are the net heat transferred and the net work done on the system via the process.

## 3.5 The Entropy Concept

The second law of thermodynamics has been formulated through the *classical* (macroscopic) and the *statistical* and *quantum-mechanics* (microscopic) concepts of *entropy* [26], and through the concept of *exergy* (energy availability or quality), which recently has gained great popularity because it lacks the ambiguities and controversies that occurred in the various interpretations of entropy [3].

### 3.5.1 The Classical Macroscopic Entropy

The term *entropy* was coined by *Rudolf Clausius* in 1865 [11] and comes from the Greek εντροπία (εν = in + τροπή = a turning) and literally means "*a turning towards*". The concept of entropy was expanded further by Maxwell [17]. Clausius had observed that the ratio of heat exchanged to absolute temperature was constant in reversible, or ideal, heat cycles. He concluded that the conserved ratio must correspond to a real, physical quantity $S$ and he called it "entropy". Thus, Clausius (classical) definition of entropy is:

$$S = Q/T \tag{3.20}$$

where $Q$ is the heat content (thermal energy) of the system ant $T$ is the absolute (Kelvin) temperature of the system. Of course, not every conserved ratio corresponds necessarily to a real physical quantity. Entropy defined in this way surely does not possess intuitive clarity. The entropy concept appears to have been introduced in physics by fortune or accident, and, in spite of Maxwell's initial reservations [17, 45], it has prevailed and is established.

As mentioned above, the definition Eq. (3.20) of entropy holds for a reversible process. This is justified by the fact that the temperature $T$ has sense only if the system is in thermodynamic equilibrium (or very near to it), in which case the system can have "one" temperature (and not many simultaneous temperatures). To determine the entropy for a *nonreversible process,* we set up a reversible process that starts and ends at the same initial and final state, respectively, with non-reversible systems. Then, we compute the entropy of this equivalent reversible process, which, because entropy does not depend on the way it takes its values, is equal to the entropy change of the original non-reversible process.

Under the assumption that $T$ is constant and the heat flow is reversible (i.e., in the case of an isothermal and reversible process), Eq. (3.20) gives the change $\Delta S$ of $S$ as:

$$\Delta S = \Delta Q/T \tag{3.21}$$

where $\Delta Q$ is the change (finite increment) of $Q$, i.e., $\Delta Q = Q_1 - Q_2$ with $Q_1$ and $Q_2$ being the heat contents (internal thermal energy) of the system at two different equilibrium states 1 and 2. Similarly $\Delta S = S_1 - S_2$, where $S_1$ and $S_2$ are the entropies of the system at the equilibrium states 1 and 2, respectively. If $\Delta Q$ is positive, then so is $\Delta S$, i.e., if the heat content $Q$ of the system goes up (with constant $T$) then its entropy goes up. If $Q$ goes down, the same happens to $S$.

If the temperature of the process is not constant, then Eq. (3.21) must be used in differential form as:

$$\mathrm{d}S = \delta Q/T \tag{3.22}$$

where $\delta Q$ denotes a "path-dependent" (inexact) differential. In this case, the finite increment $\Delta S$ in Eq. (3.21) takes the form:

$$\Delta S = \int \frac{\delta Q}{T} \tag{3.23}$$

Using the differential entropy relation Eq. (3.22) in (3.16), we get the following equation:

$$\mathrm{d}U = T\mathrm{d}S - p\mathrm{d}V \tag{3.24}$$

which is known as the "*fundamental thermodynamic equation*".

In chemistry, we deal with *open systems* where heat, work, and mass flow cross their boundaries. In these systems, the total entropy flow $\mathrm{d}S/\mathrm{d}T$ is equal to the following:

$$\frac{\mathrm{d}S}{\mathrm{d}t} = \frac{1}{T}\frac{\mathrm{d}Q}{\mathrm{d}t} + \sum_{i=1}^{L} \frac{\mathrm{d}m_i}{\mathrm{d}t} S_i^* + \frac{\mathrm{d}S_{\mathrm{int}}}{\mathrm{d}t} \tag{3.25}$$

where $\mathrm{d}Q/\mathrm{d}t$ is the *heat flow* (that causes the entropy change), $\mathrm{d}m_i/\mathrm{d}t$ is the *mass flow* (entering or leaving the system), $(\mathrm{d}Q/\mathrm{d}t)/T$ is the *entropy flow* due to heat across the boundary, $S_i^*$ is entropy per unit mass, and $\mathrm{d}S_{\mathrm{in}}/\mathrm{dt}$ is the *rate of internally generated entropy* of the system.

In case of multiple flows, the term $\mathrm{d}Q/T$ is replaced by $\Sigma_j \mathrm{d}Q_j/T_j$, where $T_j$ is the temperature at which $Q_j$ flows. It is remarked that the workflow does not contribute to the change of entropy.

Three very useful thermodynamic entities used in chemistry are the following:

$$\text{Enthalpy: } H = U + pV \tag{3.26}$$

$$\text{Helmholtz free energy: } A = U - TS \tag{3.27}$$

$$\text{Gibbs free energy: } G = H - TS \tag{3.28}$$

The differential forms of $H$, $A$, and $G$ are found as follows:

$$\begin{aligned} \mathrm{d}H &= \mathrm{d}U + p\mathrm{d}V + V\mathrm{d}p \\ &= T\mathrm{d}S + V\mathrm{d}p \quad (\text{by } 3.24) \end{aligned} \tag{3.29}$$

$$\begin{aligned} \mathrm{d}A &= \mathrm{d}U - T\mathrm{d}S - S\mathrm{d}T \\ &= -S\mathrm{d}T - p\mathrm{d}V \quad (\text{by } 3.24) \end{aligned} \tag{3.30}$$

$$\begin{aligned} \mathrm{d}G &= \mathrm{d}H - T\mathrm{d}S - S\mathrm{d}T \\ &= -S\mathrm{d}T + V\mathrm{d}p \quad (\text{by } 3.24) \end{aligned} \tag{3.31}$$

From Eqs. (3.24), (3.29), (3.30), and (3.31), it follows that:

$$U = U(S,V), H = H(S,p), A = A(T,V), G = G(T,p) \tag{3.32}$$

Therefore:

$$\mathrm{d}U = (\partial U/\partial S)_V \mathrm{d}S + (\partial U/\partial V)_S \mathrm{d}V \tag{3.33}$$

$$\mathrm{d}H = (\partial H/\partial S)_p + (\partial H/\partial p)_S \mathrm{d}p \tag{3.34}$$

$$\mathrm{d}A = (\partial A/\partial T)_V \mathrm{d}T + (\partial A/\partial V)_T \mathrm{d}V \tag{3.35}$$

$$\mathrm{d}G = (\partial G/\partial T)_p \mathrm{d}T + (\partial G/\partial p)_T \mathrm{d}p \tag{3.36}$$

Comparing Eqs. (3.24), (3.29), (3.30) and (3.31) with Eqs. (3.33), (3.34), (3.35) and (3.36), respectively, we obtain the following alternative thermodynamic definitions of $T$, $p$, $V$, and $S$:

$$T = (\partial U/\partial S)_V, \quad T = (\partial H/\partial S)_p \tag{3.37a}$$

$$p = -(\partial U/\partial V)_S, \quad p = -(\partial A/\partial V)_T \tag{3.37b}$$

$$V = (\partial H/\partial p)_S, \quad V = (\partial G/\partial p)_T \tag{3.37c}$$

$$S = -(\partial A/\partial T)_p, \quad S = -(\partial G/\partial T)_p \tag{3.37d}$$

It is remarked that the volume is not the most appropriate independent variable because in the laboratory it is much easier to control pressure than volume. On the other hand, although Eq. (3.37d) indicates that we may also use $S$ as independent variable, in practice it is not all at convenient to use $S$ as independent variable or to control it, since there does not exist a measurement device for $S$, and it is not known how to keep entropy constant during the change of some other variable.

For a constant $T$, Eq. (3.28) yields:

$$\mathrm{d}G = \mathrm{d}H - T\mathrm{d}S \tag{3.38}$$

The enthalpy $H$ represents the heat content of the system, and the quantity $T\mathrm{d}S$ represents the ability to do the work needed for the reaction to take place. Equation (3.38) is very important since the value of $\mathrm{d}G$ determines whether a certain reaction can take place "*spontaneously*" or needs external energy to occur. If $\mathrm{d}G < 0$, i.e., if the change $\mathrm{d}H$ of the enthalpy (heat content) is less than $T\mathrm{d}S$, the reaction will take place spontaneously. Otherwise ($\mathrm{d}G > 0$), the reaction needs at least $\mathrm{d}G$ worth of energy to force it to occur.

Actually, the change of Gibbs free energy, $\Delta G$, in a reaction represents the maximum amount of work that can be obtained from it. For example, $\Delta G$ in the oxidation of glucose is $\Delta G$ = 686 kcal = 2870 kJ. This is the energy that sustains the life of living cells.

### 3.5.2 The Statistical Concept of Entropy

The statistical concept of entropy was developed by Maxwell, Boltzmann, and Gibbs extending the work of classical thermodynamics, via the "*molecular theory*" of gases, into the domain of *statistical mechanics*. Boltzmann defined entropy as a measure of the number of possible microstates (*micro*scopic *states*) of a system in thermodynamic equilibrium, consistent with its *macroscopic* thermodynamic properties (or macrostate). For example, the macroscopic property temperature of a system defines a *macrostate* variable, whereas the kinetic energy of each molecule in the system specifies a microstate. Actually, the macrostate thermodynamic variable temperature expresses the average of the microstate variables, i.e., the average kinetic energy of the system. Consequently, when the molecules of a gas have higher velocities (i.e., higher kinetic energies), the temperature of the gas increases. It is obvious that the macrostate of a system can be described by a small number of variables (like, $U$, $T$, $V$, etc.) only if the system is in thermodynamic equilibrium. Boltzmann's definition of entropy is given by:

$$S = k_{\mathrm{B}} \ln N \tag{3.39}$$

where $N$ is the total number of microstates consistent with the given macrostate, and $k_{\mathrm{B}}$ is a physical constant called the *Boltzmann constant*, which (like the entropy) has units of heat capacity. The term ln$N$ is dimensionless. It is remarked that "N" is not the total number of particles in the system, but rather the overall number of "microstates" where the particles will be, under the condition that all such microstate populations would lead to the same macrostate.

A simple demonstration of the Eq. (3.39) is the following [46]. We start with the classical definition of entropy $S = Q/T$ and calculate the heat $Q$ needed by an ideal gas to expand from volume $V_1$ to volume $V_2$ at temperature $T$ and unchanged pressure $p$. The energy of $Q$ is equal to:

$$Q = p \int_{V_1}^{V_2} \mathrm{d}V$$

where $p$ and $V$ are related by the ideal gas state equation:

$$pV = nRT = k_{\mathrm{B}} MT \tag{3.40}$$

where $R$ is the universal gas constant ($R$ = 8.314 J/mol. K), $n$ is the number of moles, $M$ is the number of molecules, and $k_{\mathrm{B}}$ is Boltzmann's constant given by:

$$\begin{aligned} k_{\mathrm{B}} &= R(n/M) = R/(M/n) \\ &= R/(\text{Avogadro number } 6.02 \times 10^{23}) \\ &= 1.38^{06} \times 10^{-23} [\mathrm{J/K}] \end{aligned}$$

Thus, the entropy increase is equal to:

$$S = \frac{Q}{T} = \frac{P}{T}\int_{V_1}^{V_2} \mathrm{d}V = k_{\mathrm{B}} M \int_{V_1}^{V_2} \frac{\mathrm{d}V}{V}$$
$$= k_{\mathrm{B}} M \ \ln(V_2/V_1) = k_B \ln(V_2/V_1)^M$$
$$= k_{\mathrm{B}} M \ \ln N$$

where $N = (V_2/V_1)^{\ M}$ is the new number of microstates. This is Boltzmann's formula (3.39). Just as an illustration of the meaning of $N$, we consider a gas volume with 100 molecules and exactly 100 places for them. Suppose that the gas is heated so that the volume is doubled, i.e., $(V_2/V_1) = 2$ and the number of places is doubled. Now each molecule can be located in either of two places and thus we have $N = 2^{100}$ microstates. As a result, we have an increase of entropy equal to $\mathrm{kBln}2^{100} = 69.3$ kB. In general, $N$ is the number of possible microstates consistent with the given macrostate, i.e., the number of non-observable configurations (or ways) in which the observable macrostate can be obtained through different positions and momentums of the various molecules. For an ideal gas with $M$ identical molecules and $M_i$ molecules at the $i$th microscopic configuration (range) of position and momentum, $N$ is calculated using the permutations formula

$$N = M!/M_0!M_1!M_2!\ldots \tag{3.41}$$

where "!" is the *factorial symbol* and $i$ ranges over all possible molecular configurations. The denominator stands for taking into account the fact that identical molecules in the same microstate (configuration, condition) cannot be discriminated.

Boltzmann's entropy given by Eq. (3.39) is applicable to microstates that are equally probable, which means that the system must be in a state of thermodynamic equilibrium. For systems not in thermal equilibrium, the microstate probabilities must be treated individually.

In this case, the entropy is given by the following generalization which is known as *Gibbs entropy*:

$$S = -k_{\mathrm{B}} \sum_i p_i \ln(p_i) \tag{3.42}$$

where $p_i$ is the probability that particle $i$ will be in a certain microstate, and all the $p_i s$ are computed for the same macrostate of the system. The negative sign is needed because all $p_i s$ belong to the interval $0 \leq p_i \leq 1$, and so $\ln(p_i) < 0$ for all $i$. Here, the summation extents over all the particles $i$. A verification of the fact that Eq. (3.42) is an expression equivalent to the known entropy relation (3.22) is the following [47, 48]. Consider a system in thermodynamic equilibrium with a heat

reservoir at a given temperature $T$. Then, the probability distribution (at equilibrium) over the energy eigenvalues $E_i$ are given by Boltzmann's distribution:

$$p_i = (1/K)\exp(-E_i/k_B T) \tag{3.43}$$

where $K$ is the normalization factor (known as partition function) that assures that all probabilities sum to 1: $\sum_i p_i = 1$. Now, the differential changes of $S$ due to changes in the external parameters, upon which the energy levels depend, is found from Eq. (3.42) to be:

$$dS = -k_B \sum_i \ln(p_i) dp_i \tag{3.44}$$

Introducing $p_i$ from Eqs. (3.33)–(3.34), and noting that $\sum_i dp_i = 0$, gives:

$$\begin{aligned} dS &= -k_B \sum_i (-\ln K - E_i/k_B T) dp_i \\ &= \frac{1}{T}\sum_i E_i dp_i = \frac{1}{T}\sum_i d(E_i P_i) - \frac{1}{T}\sum_i p_i dE_i \\ &= (dU - \delta W)/T = \delta Q/T \quad \text{(by (3.12))} \end{aligned}$$

which is the desired relation Eq. (3.22). In the above derivation, it was assumed that the external parameters exhibit slow changes which assure that the system does not change microstates, although the system's macrostate changes slowly and reversibly. In this case, the term $\sum_i p_i dE_i$ represents the (average) work done on the system via this reversible process.

### 3.5.3 *The Von Neumann Quantum-Mechanics Entropy Concept*

The term "*quantum*" was coined by *Max Planck* during his "black-body" radiation studies. He arrived at the conclusion that the radiation energy of black-bodies exist in discrete quantities (wave parcels), which he called "*quanta*". This holds in general for any electromagnetic radiation which actually consists of such wave parcels that are both "*particle*" and "*wave*". The extension of the classical entropy concept to the quantum-mechanics field is due to *John von Neumann* who defined his entropy as:

$$S(\rho) = -\mathrm{tr}(\rho \ln \rho) \tag{3.45}$$

where $\rho$ is the so-called (quantum) *density matrix* and "tr" is the conventional trace operator of a matrix:

$$\text{tr}\,\rho = \Sigma(\text{diagonal elements of}\,\rho)$$

The *density matrix* in quantum mechanics is a *self-adjoint* (or *Hermitian*) positive semidefinite matrix[1] of trace one that describes the statistical state of a quantum system. It is the quantum-mechanical analog of a phase-space-probability measure (probability of position and momentum) in classical statistical mechanics.

A quick explanation of the definition of $S(\rho)$ in Eq. (3.45) is the following. The state vector $|\Psi\rangle$ of a quantum system specifies fully the statistical performance of a measurement $B$. Consider a mixed quantum system consisting of the statistical combination of a finite set of pure states $|\Psi_i>$ with corresponding probabilities

$$p_i(0 \le p_i \le 1, \quad \sum\nolimits_i p_i = 1.$$

This means that the *preparation process*, i.e. the reproducible scheme used to generate one or more homogeneous ensembles to be studied for the system, ends in the state $|\Psi_i\rangle$ with probability $p_i$. The expectation $\langle B \rangle$ of the measurement $B$ is given by:

$$\langle B\rangle = \sum_i p_i \langle \psi_i | B | \psi_i \rangle \tag{3.46}$$

Now, let us define the "*density matrix*" (*operator*) of the quantum system as:

$$\rho = \sum_i p_i |\psi_i\rangle\langle\psi_i| \tag{3.47}$$

Then, it follows that $\langle B \rangle$ can be written as:

$$\langle B\rangle = \text{tr}[\rho B] \tag{3.48}$$

The density operator $\rho$ defined by Eq. (3.47) is a positive semi-positive operator, and has a spectral decomposition:

$$|\phi_i\rangle \tag{3.49}$$

where $|\varphi_i\rangle$ are orthonormal vectors, and the coefficients $\lambda_i$ are the eigenvalues of $\rho$.

The Neumann (quantum mechanics) entropy is defined as follows:

[1]†A complex-valued matrix $\mathbf{A}$ is called Hermitian (or self-adjoint) if $\mathbf{A}^{\mathrm{H}} = \mathbf{A}$, where $\mathbf{A}^{\mathrm{H}}$ is a matrix with elements the conjugate elements of the transpose matrix $\mathbf{A}^{\mathrm{T}}$.

$$\begin{aligned} S &= -\sum_i \lambda_i \ln \lambda_i \\ &= -\mathrm{tr}(\rho \ln \rho) = S(\rho) \end{aligned} \tag{3.50}$$

This is the definition given in Eq. (3.45). Multiplying Neumann's entropy by the Boltzmann constant $k_B$ tunes it with the statistical mechanics (Gibbs) entropy of Eq. (3.42).

The evolution in time of the density operator $\rho$ obeys the following equation, which is called the *von Neumann equation*:

$$i\hbar \partial\rho/\partial t = [H, \rho] \tag{3.51}$$

where $i$ is the imaginary unit, $\hbar$($h$-bar) is the *reduced Planck constant* ($\hbar = \mathrm{h}/2\pi$), $H$ is the *Hamiltonian* of the system[2], and the brackets symbolize a *commutator*:

$$[H, \rho] = H\rho - \rho H \tag{3.52}$$

This equation is valid if the density operator is considered to be in the *Schrödinger* picture. If the Hamiltonian is time independent, the solution to Eq. (3.51) has the form:

$$\begin{aligned} \rho(t) &= U(t)\rho(0)U^{\dagger}(\mathrm{t}) \\ U(t) &= \exp(-iHt/\hbar) \\ U^{\dagger}(t) &= \exp(iHt/\hbar) \end{aligned} \tag{3.53}$$

where $U^{\dagger}$ is the Hermitian conjugate of $U(t)$.

If $H$ depends explicitly on $t$, i.e., $H = H(t)$, then $U(t)$ is the solution of the following:

$$\mathrm{d}U(t, t_0)/\mathrm{d}t = -(i/\hbar)H(t)U(t, t_0) \tag{3.54}$$

It is recalled that the pure states (*key vectors*) $|\Psi_i\rangle$ are governed by the Schrödinger equation:

$$i\hbar \frac{\mathrm{d}}{\mathrm{d}t}|\Psi_i\rangle = H_i|\Psi_i\rangle \tag{3.55}$$

and so Eq. (3.51) is the statistical average of Schrödinger equations of the type (3.55)

Some basic properties of the density operator $\rho$ and the von Neumann entropy $S$ ($\rho$) are the following:

---

[2*]In the one-dimensional case, $H$ is given by $H = -(\hbar^2/2m)\partial^2/\partial x^2 + V(x)$ where, $V(x)$ is the time-independent potential energy of the particle at position $x$.

**P1**. The joint density operator of two systems $A$ and $B$ is denoted by $\rho_{AB}$. Then, the density of the subsystem $A$ is given by:

$$\rho_A = \mathrm{tr}_B \rho_{AB}$$

where $\mathrm{tr}_B$ is the so-called trace over the system $B$.
**P2**. If $A$ and $B$ are two distinct and independent systems, then: $\rho_{AB} = \rho_A \otimes \rho_B$ (product law).
**P3**. $S(\rho)$ is zero for pure states.
**P4**. $S(\rho)$ is invariant under changes in the basis of $\rho$.
**P5**. $S(\rho)$ is additive, i.e., $S(\rho_A \otimes \rho_B) = S(\rho_A) + S(\rho_B)$.
**P6**. $S(\rho)$ is concave, i.e., $S\left(\sum_{i=1}^{k} \mu_i \rho_i\right) \geq \sum_{i=1}^{k} \mu_i S(\rho_i)$ with $\mu_1 + \mu_2 + \cdots + \mu_k = 1$.
**P7**. $|S(\rho_A) - S(\rho_B)| \leq S(\rho_{AB}) \leq S(\rho_A) + S(\rho_B)$ (when $\rho_A$ and $\rho_B$ are the reduced density operators of $\rho_{AB}$).

### *3.5.4 The Non-statistical General Physics Entropy Concept*

Based on the non-statistical general physics concepts briefly presented in Sect. 3.2.9, Gyftopoulos and his coworkers presented two formulations for the entropy. The first is *purely thermodynamic* (without any probabilities) and the second is *purely quantum mechanical* (i.e., the probabilities are not mixtures of statistical and quantum probabilities). This second approach provides a unified quantum theory of mechanics and thermodynamics, based on the assumption of a wider set of quantum states than that postulated in classical quantum-mechanics. A brief exposition of them follows.

#### 3.5.4.1 Pure Thermodynamic Entropy Concept

The development of the entropy concept in this formulation is based on the properties of *Energy E* (see Sect. 3.2.8) and the *generalized available energy* $\Omega^R$, which is a generalization of Carnot's motive power and provides the limit on the optimum amount of energy that can be exchanged between a weight mg and a composite of system $A$ and reservoir $R$ (i.e., the limit on the optimum mechanical effect). The property $\Omega^R$ is additive like the energy $E$. Carnot assumed that $A$ is a reservoir, too.

Gyftopoulos and Berreta [10] showed that for an adiabatic process of system $A$, only the energy changes $E_1 - E_2$ of $A$ and the generalized available energy changes $\Omega_1^R - \Omega_2^R$ of the composite of $A$ and $R$ satisfy the relations:

$$E_1 - E_2 = \Omega_1^R - \Omega_2^R \quad \text{(for a reversible process)} \tag{3.56a}$$

$$E_1 - E_2 < \Omega_1^R - \Omega_2^R \quad \text{(for a non reversible process)} \tag{3.56b}$$

In a reversible process, both the system and its environment can be returned to their own initial states, whereas, in an irreversible process, this restoration cannot take place.

In terms of $E$ and $\Omega^R$, a property of $A$ at state $A_1$ can be determined which is called *entropy* $S$, and is evaluated via a reservoir $R$ and a reference state $A_0$ by the equation:

$$S_1 = S_0 + \frac{1}{c_R}[(E_1 - E_0) - (\Omega_1^R - \Omega_0^R)] \tag{3.57}$$

Here, $S_i$, $E_i$, and $\Omega_i^R$ are the values of $S$, $E$, and $\Omega^R$ at the states $A_i$ ($i$ = 0, 1), respectively.

The quantity $c_R$ is a well-defined positive constant that depends only on the reservoir R and is defined as:

$$c_R = c_{R_0}(\Delta E_{12}^R)_{\text{rev}}^A / (\Delta E_{12}^{R_o})_{\text{rev}}^A \tag{3.58}$$

where $(\Delta E_{12}^R)_{\text{rev}}^A$ is the energy change in a reversible weight process for the composite of $R$ and an auxiliary system $A$ in which $A$ changes from fixed states $A_1$ and $A_2$, and $(\Delta E_{12}^{R_0})_{\text{rev}}^A$ is the energy change of a reference reservoir $R_0$ under otherwise identical conditions. The quantity $c_{R_0}$ corresponds to the reference reservoir and can be assigned an arbitrary value. The value and the dimension of $c_R$ are equal to the value and the dimension of the temperature of every stable equilibrium state of a given reservoir $R$. Choosing $c_{R_0} = 273.16\,\text{K}$ assures that $c_R$ is measured in absolute temperature units (Kelvin).

The entropy $S$ is an *inherent property* of $A$ only, because it is shown to be independent of the reservoir. The entropy is also shown to have always non-negative values, with zero value for all states encountered in mechanics. Finally, from Eqs. (3.56a) and (3.56b), it follows that the entropy maintains a constant value during any reversible adiabatic process, and increases in any irreversible process of $A$, a result which is also true for spontaneous processes and zero-net-effect interactions.

### 3.5.4.2 Pure Quantum-Mechanical Entropy Concept

The quantum-mechanical entropy concept of von Neumann is based on the assumption that the probabilities associated with the measurement results of a system in a state $i$ are described by a wave-vector function $|\Psi_i\rangle$ or, equivalently, by a projector $|\Psi_i\rangle\langle\Psi_i| = \rho_i = \rho_i^2$, and that the density matrix $\rho > \rho^2$ is a statistical

average of projectors as given by Eq. (3.47) [19]. This means that each density operator $\rho$ represents a mixture of quantum-mechanical probabilities specified by projectors and non-quantum-mechanical (statistical) probabilities expressing our actual inability to model, handle, and control all the details of a preparation or of the interactions of the system with its environment, and so on.

Hatsopoulos and Gyftopoulos [28] found that there are quantum-mechanical situations that require a purely quantum-mechanical density operator which is not a statistical mixture of projectors [like that of Eq. (3.47)]. Such purely quantum-mechanical operators are identically prepared by an ensemble of identical systems, which is called a *homogeneous* or *unambiguous ensemble*.

Recall that two or more systems are identical if they are described by the same density operator. It is exactly this pure quantum-mechanical feature of density operators that allows the extension of quantum ideas to thermodynamics and vice versa. In [29, 30], eight conditions or criteria are given that must be satisfied (at a minimum) by any expression claimed to represent the entropy $S$. It was established [30] that, from the entropy expressions available in the literature, the only one that satisfies all those criteria (and so it is acceptable) is the von Neumann expression:

$$S = -k_{\mathrm{B}}\mathrm{tr}(\rho \ln \rho) \tag{3.59}$$

under the condition that $\rho$ is *purely quantum mechanical* (and not a mixture of quantum mechanical and statistical probabilities). The two criteria that are not conformed by all other entropy expressions of the literature are the following:

c1. The expression must be valid for every system (large or small) and every state (not only stable equilibrium states).
c2. The expression must, in general, be non-negative and be zero for all the states encountered in mechanics.

The other six properties that are possessed by the other entropy expressions are briefly the following: (i) invariance under reversible adiabatic conditions and an increase of value in irreversible adiabatic processes; (ii) additivity, uniqueness of maximum value of the expression; (iii) for given values of energy, parameters and amounts of constituents, only one state must correspond to the largest value of the expression; (iv) concavity and smoothness of the entropy versus energy plot in case of stable equilibrium states; (v) identical results for the thermodynamic potentials for all three systems $A$, $B$, $C$ when maximizing the composite $C$ of two systems $A$ and $B$; and (vi) reduction to relations established by experiment, for stable equilibrium states.

The Hatsopoulos and Gyftopoulos equation [28] that governs unitary evolutions of $\rho$ in time and is valid for both *isolated systems* ($H$ independent of time) and *non-isolated systems* ($H$ dependent on time), is postulated to have the same form as the von Neumann Eq. (3.51), but actually it is different because $\rho$ is not a statistical mixture of projectors, and so it cannot be produced by statistically averaging Schrödinger equations.

The general equation of motion of quantum thermodynamics that governs the time evolution of both reversible and irreversible processes is derived and fully studied in [10, 31–33].

### 3.5.5 *Rényi Entropy, Tsallis Entropy, and Other Entropy Types*

In 1961, Rényi [49] introduced a generalization of the standard Boltzmann-Gibbs-Shannon entropy that is dependent on a parameter $\alpha(a \geq 0)$ and is given by the formula:

$$H_\alpha(x) = -\log\left(\sum_{i=1}^{n} p_i^\alpha\right)/(\alpha - 1)$$

where $p_i$, $i$ = 1,2, …, n are the probabilities of $x_1$, $x_2$, …, $x_n$ and log is the logarithm in base 2. This is called the *Rényi entropy of order* α and has been used in statistics and ecology, where α indicates the index or degree of uncertainty diversity. Some special cases of $H_\alpha(x)$ are the following:

$$\begin{aligned}
&\alpha = 0, H_0(x) = \log n && \text{Hartley entropy}\\
&\alpha = 1, H_1(x) = -\sum_{i=1}^{n} p_i \log p_i && \text{Boltzmann entropy}\\
&\alpha = 2, H_2(x) = -\log \sum_{i=1}^{n} p_i^2 && \text{Collision entropy}\\
&\alpha \to \infty, H_\infty(x) = -\log \sup_{1 \leq i \leq n} (p_i) && \text{Min-entropy}
\end{aligned}$$

The name *Min-entropy* in the case $\alpha \to \infty$ is due to the validity of the following bounding relation:

$$H_\infty < H_2 < 2H_\infty$$

to indicate that $H_\infty$ is the smallest value of $H_2$, which very often is referred to simply as "*Rényi entropy*".

It is noted that $H_\alpha(x)$ is a lightly decreasing function of α, and $H_1(x) \geq H_2(x)$ because $\sum_{i=1}^{n} p_i \log p_i \leq \log \sum_{i=1}^{n} p_i^2$ (Jensen Inequality).

*Tsallis entropy* was introduced in 1988 [50] and depends on a parameter q as follows:

$$S_q(p) = \left(1 - \int_{-\infty}^{\infty} p^q(x)\mathrm{d}x\right)/(q-1)$$

in the continuous probability distribution case, and

$$S_q(p) = \left(1 - \sum_x p^q(x)/(q-1)\right)$$

in the discrete probability case, where $p(x)$ is the probability distribution of concern. When $q \to 1, S_q(p)$ reduces to the standard Boltzmann entropy Tsallis entropy does not add up from system to system, i.e., it is not based on the extensivity assumption, the parameter q being a measure of the degree to which the non-extensivity holds. In the non-extensive cases, the correlations between individual constituents of the system do not decay exponentially with distance, as they do in extensive situations. Actually, in non-extensive processes, the correlations die off as the distance is raised to some power (theoretically found or experimentally derived), which is the so-called "*power law*" (e.g., Richter's power law of earthquakes strength). Tsallis entropy has found application in the study of a large variety of physical processes and phenomena with power-scaling relationships of this kind (like tornadoes, chaotic systems, solid-state phenomena, anomalous diffusion in condensed matter, etc.)

For two independent systems $A_1$ and $A_2$ that obey the product probability law $p(A_1, A_2) = p(A_1)p(A_2)$, the Tsallis entropy has the property:

$$\begin{aligned} S_q(A_1, A_2) = S_q(A_1) + S_q(A_2) \\ + (1-q)S_q(A_1)S_q(A_2) \end{aligned}$$

which for $q = 1$ reduces to the *extensivity property*:

$$S_1(A_1, A_2) = S_1(A_1) + S_1(A_2)$$

with $S_1 = S$ (the standard entropy).

Clearly, the parameter $|1 - q|$ represents a measure of non-extensivity (departure from extensivity).

In [51], the entropy increase principle is extended to the case of Tsallis entropy in the framework of the *non-extensive statistical mechanics* (NSM) [52]. Specifically, an inequality for the change of Tsallis entropy is derived for the case where two non-extensive systems of different temperatures are brought into contact with each other and reach thermal equilibrium. When the two systems 1 and 2 are not in contact, i.e., when they are independent, the probability distribution $p_0$ and the Tsallis entropy $S_{q_0}$ of the total (composite) system are equal to $p_0 = p_1 p_2$ and $S_{q_0} = S_{q_1} + S_{q_2} + (1-q)S_{q_1}S_{q_2}$, respectively. The change of entropy after and before the contact (at thermal equilibrium) is equal to:

$$S_q - S_{q_0} = \frac{1}{1-q} \int_{-\infty}^{\infty} p^q [1 - (p_0/p)^q] \mathrm{d}x$$
$$= c_q \left\langle X_q(p_0, p) \right\rangle_q$$

where

$$c_q = \int_{-\infty}^{\infty} p^q \mathrm{d}x, X_q(p_0, p) = [1 - (p_0/p)^q]/(1-q)$$

and $\left\langle X_q(p_0, p) \right\rangle_q$ is the *q-expectation value* defined as:

$$\left\langle X_q \right\rangle_q = \int_{-\infty}^{\infty} p^q X_q \mathrm{d}x / \int_{-\infty}^{\infty} p^q \mathrm{d}x$$

For $(p_0/p) > 0$ and $q > 0$, the following inequality holds:

$$X_q(p_0, p) \geq q[1 - (p_0/p)]/(1-q)$$

where the equality is obtained if and only if $p = p_0$. Thus, for $q > 0$ we have $\left\langle X_q \right\rangle \geq 0$ where $\left\langle X_q \right\rangle = \int p X_q \mathrm{d}n$ is the conventional expectation. On the basis of the above, we get:

$$S_q(p_0, p) - S_{q_0} \geq 0$$

which states that the total Tsallis entropy cannot decrease, and it is unchanged if and only if $p = p_0$. The two systems in contact with each other can be treated as an isolated system.

In [53], a generalization of the Boltzmann-Gibbs entropy is provided, based on the Sharma–Mittal measure and the $q$-formalism of the generalized logarithm, which unifies Rényi and Tsallis entropies, thus showing that they actually are not generalizations of each other. In [54], it is shown that the Tsallis entropy matches exactly the previously defined Havrda–Charvat structural $\alpha$-entropy [55].

In [56], a general mathematical expression is provided which reduces to the expressions of the following entropy measures as particular limiting cases:

Aczel–Daroczy entropy, Varma entropy, Kapur entropy, Havrda–Charvat entropy, Arimoto entropy, Sharma–Mittal entropy, Taneja entropy, Sharma–Taneja entropy, Ferreri entropy, Sant'Anna–Taneja entropy, Belis–Guiasu entropy, Gil entropy, and Picard entropy.

## 3.6 The Second Law of Thermodynamics

### 3.6.1 General Formulations

The origin of the second law is attributed to the French physicist *Sadi Carnot* who in 1824 published his work entitled "*Reflections on the Motive Power of Fire*". In this publication, Carnot stated that work (motive power) is done by the flow of heat (caloric) from a hot to a cold body (working substance) [15]. Over the years, a large number of alternative formulations of the second law have been presented, which are all equivalent in the sense that each one can lead via logical arguments to every other form [14]. A list of 125 variations of the second law is provided in [34]. Here, only a few of them will be given both in terms of the concept of *entropy* and the concept of *exergy*. All of them actually demonstrate the nature's phenomenon of irreversibility, the so-called "*arrow of time*" [57, 58].

The formulation of Carnot states:

"*Heat travels only from hot to cold*"

The first formulation of Clausius is:

"*Heat generally cannot flow spontaneously from a material of lower temperature to a material of higher temperature.*"

Kelvin's formulation is the following:

"*It is impossible to convert heat completely into work in a cyclic process*".

Max Planck's formulation (1897) is:

"*It is impossible to construct an engine which, working in a complete cycle, will produce no effect other than raising of a weight and the cooling of a heat reservoir.*"

Caratheodory presented in 1909 the so-called "*axiomatic thermodynamics*", which is based on some interesting properties of Pfaffian differential equations [12]. He started mathematically with the definition of equilibrium states and thermodynamic coordinates. He then introduced a first axiom concerning the internal energy of a multiphase system and its variation that includes external work during an adiabatic process, i.e., $U_f - U_i - W = Q = 0$ (see Eq. 3.9). This axiom uses the term "*adiabatic accessibility*", coined by Caratheodory, and states that: "*In every neighborhood of any point* (*equilibrium state*) *A in thermodynamic phase, there are points adiabatically inaccessible from point A.*"

On the basis of this axiom, Caratheodory showed how to derive Kelvin's formulation of the second law and all other results of the classical thermodynamics developed in the nineteenth century. In the definitions and axioms of Caratheodory, there is no mention of temperature, heat, or entropy. Actually, heat is considered as a derived property that appears as soon as the adiabatic restriction is removed. Full discussions of the equivalence of Caratheodory's and Kelvin's formulations can be found in [59, 60].

Using their definitions of *system*, *property*, *state,* and *energy* outlined in Sect. 3.2.9, Gyftopoulos and Berreta formulated the second law as follows [10]:

> "*Among all the states of a system with given values of energy, amounts of constituents, and parameters there exists one and only one stable equilibrium state.*"

The stability in the above statement is, for each set of conditions, global (not local) [61]. In mechanics, it is found that for each set of conditions the only stable equilibrium state is that of the lowest energy. On the contrary, the second law assures that there exists a global equilibrium state for every value of energy. That is, the second law implies the existence of a large class of states, in addition to those encountered in mechanics.

Two consequences of the Gyftopoulos–Berreta form of the second law are the following:

1. In any system, starting from a stable equilibrium state, no energy is available to produce a mechanical effect as long as the values of the amounts of constituents, the internal forces, and the parameters experience no net changes. This consequence is known as the *impossibility of the PMM2* (perpetual motion machine of the second kind). Usually, PMM2 is erroneously considered as the statement of the second law, while in the above formulation it follows as a theorem based on the first and second laws.
2. Not all states of a system can be changed to a state of lower energy via a mechanical effect. This is a generalization of PMM2 and implies that there exists a property of a system in a given state that represents the optimum amount of energy that can be exchanged between the system and a weight process starting with system $A$ in a state $A'_1$ and ending in a state $A''_2$. This property, which is called "*generalized adiabatic availability*" (denoted by $\Psi$), is well-defined for all systems, but unlike energy is not additive. A property that has the features of $\Psi$ and at the same time is additive is the *generalized available energy* $\Omega^R$ discussed in Sect. 3.5.4.

### 3.6.2 Formulations Through Entropy

The second law, as formulated through the entropy $S$, states:

> "*The entropy of an isolated system not in thermal equilibrium tends to increase over time, approaching a maximum at equilibrium*"

The best known and most famous formulation of the second law was given in 1865 by Clausius and is the following:

> "*The entropy of the Universe tends to a maximum*"

The mathematical formulation of the second law for an isolated system, with the entropy $S$ considered a time-varying entity $S(t)$, is:

$$\frac{\mathrm{d}S(t)}{\mathrm{d}t} \geq 0 \tag{3.60}$$

where $t$ is time. In words, Eq. (3.60) states that:

"*Entropy in an isolated system can never decrease.*"

The "*non-decrease*" property in all natural processes that take place in isolated systems implies a particular direction of time that is known as the "*arrow of time*" (Sect. 3.12). As the time goes forward, the second law states that the entropy of isolated systems tends to increase or remains the same. It will not decrease. In other words, entropy measurements can be considered as a "*type of clock*".

Kelvin's and Max Planck's formulations given above mean that we cannot take energy from a high-temperature source and then convert all that energy into heat. Some of this energy has to be transferred to heat a lower temperature object. In other words, it is thermodynamically impossible to have a heat engine with a hundred-percent efficiency. Carnot has shown that the most efficient heat-engine cycle is a cycle (called now the *Carnot cycle*) that consists of two *isothermal* processes and two *adiabatic processes*. This cycle sets the upper physical limit on the value of the fraction of the heat that can be used for work. The Carnot cycle is an "*idealization*" since it assumes that the cycle is reversible and involves no change in entropy. This is because no real-engine processes are reversible, and in all real physical processes an increase of entropy occurs. On the other hand, the isothermal heat transfer is too slow to be of practical value and use.

To calculate the efficiency of the Carnot cycle, we consider an engine working as above between a *high-temperature* $T_\mathrm{h}$ and a *low-temperature* $T_\mathrm{c}$.

The Carnot cycle in *T-S* and *V-T* diagrams has the form of Fig. 3.6a, b.

The operation of the engine in this cycle is as follows:

1. **Hot slow isothermal expansion** (from point 1 to point 2). The hot gas receives heat $Q_h = T_h \Delta S$ and does work.
2. **Cooling adiabatic expansion** (from point 2 to point 3). The gas continues to do work without any heat exchange.
3. **Cold slow isothermal compression** (from point 3 to point 4). The cold gas receives work and gives out (wasted) heat equal to $Q_c = T_c \Delta S$.
4. **Heating adiabatic compression** (from point 4 back to point 1). The gas is compressed by outside work and returns to the original hot state 1.

According to Eq. (3.18), the total work done by the engine in this cycle is equal to the net amount of heat received, i.e.:

$$\text{Work done} = \oint \delta Q = Q_h - Q_c$$

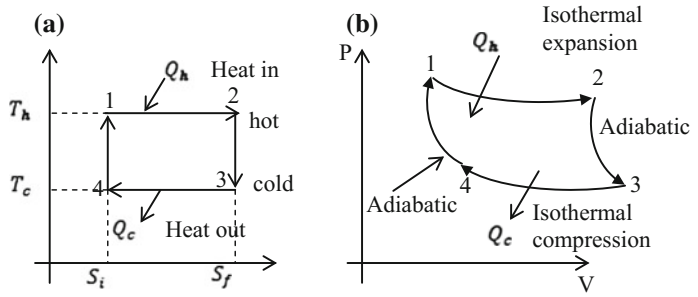



**Fig. 3.6** Representation of a Carnot cycle in the temperature–entropy and pressure–volume planes

Then, the engine's efficiency η is equal to:

$$\eta = \frac{\text{Work done}}{Q_h} = \frac{Q_h - Q_c}{Q_h} = \frac{T_h - T_c}{T_h} \tag{3.61}$$

or

$$\eta = (1 - T_c/T_h)100\% \tag{3.62}$$

From Eq. (3.61), it follows that for the Carnot cycle:

$$Q_h/T_h - Q_c/T_c = 0$$

which is a special case of the Clausius theorem:

$\oint \frac{\delta Q}{T} = 0$ Around a reversible cycle.

For any irreversible cycle, the efficiency is less than that of the Carnot cycle, i.e., there is less heat "*flow in*" ($Q_h$) and/or more heat "*flow out*" ($Q_c$) of the system. This implies that:

$$\oint \frac{\delta Q}{T} \leq 0 \tag{3.63}$$

This is well known as *Clausius inequality* which in words states that any real engine gives to the environment more entropy than that taken from it, thus leading to an overall increase in entropy.

Clausius first statement means that heat cannot flow from cold to hot without external work input. This, for example, is the case of a refrigerator, where heat flows from cold to hot with the aid of a compressor (i.e., not spontaneously) which takes electrical energy to operate, and returns heat to the environment. Overall, the entropy of the isolated system (consisting of the refrigerator plus the environment) is increased as required by the second law.

### 3.6.3 Formulation Through Exergy

*Available energy* is a concept first understood by Sadi Carnot for the case of heat engines and further developed by Helmholtz and Gibbs. This concept was applied to various processes and applications with many different names such as energy availability, available useful energy, energy quality, utilizable energy, available energy, and so on. But the name that was definitely adopted and stuck is *exergy*, coined by *Zorant Rant* in 1956. The term *exergy* comes from the Greek "εξ + έργου" (*ex* = from + *ergy* = work). Exergy is a combined property of a system and its environment. This is due to the fact that exergy, in contrast to energy, depends not only on the state of the system but also on the state of its environment (see Eq. 3.72). Some definitions of exergy proposed over the years are the following:

**Definition 1** Exergy of the system is the energy available for use.

**Definition 2** Exergy is the maximum work that can be done during a process that brings the system into equilibrium with a heat reservoir.

**Definition 3** Exergy of a system represents the totality of the free energies in the system.

**Definition 4** Exergy is the amount of work that can be obtained when some matter (system) is brought to a thermodynamic equilibrium state with the common components of the natural surroundings via a reversible process, involving interactions only with the above-mentioned components of nature.

According to the *first law,* the total amount of energy never changes even after many changes, movements, or events. But this law does not tell us "why energy cannot be reused over and over again." Answers to this question are given by the second law, when expressed through exergy, i.e.,

**Second Law via Exergy**
*In all energy movements and transfer-motions, exergy* (i.e., *energy quality, energy availability, utilizable energy*) *is consumed. In all energy movements and transfer-motions, exergy* (i.e., *energy quality, energy availability, utilizable energy*) *is consumed.*

This means that energy is not consumed, i.e., it may have changed from chemical to kinetic energy, or from electrical to thermal energy, but the total amount remains invariant. What is consumed is the energy's quality or availability or utilizability that represents the amount of action or work that it can do. Therefore, the second law of thermodynamics can be restated as follows:

> "*The amount of action/work that a certain amount of energy can do is reduced each time this energy is used*".

This implies that all natural processes are *irreversible*, because the available energy (exergy) driving them is reduced at all times, and so the quality of energy in

the Universe as a whole, is constantly diminishing. In other words, the second law of thermodynamics indicates that: "*All processes in the Universe move in the direction of decreasing exergy content of the Universe*".

This one-way direction of the Universe towards decreasing exergy implies that heat cannot flow (spontaneously) from a colder to a hotter body. That is, the second law says that the high quality (grade, concentration), useful (utilizable) energy always gets turned into a lower grade (less useful, less utilizable) energy, which is very difficult or impossible to be reused.

The decrease or destruction of exergy is proportional to the entropy increase in the "system and its environment". The consumed exergy is called *anergy*. For an isothermal process, exergy and energy are interchangeable concepts, and there is no anergy. From the above, it follows that:

"*The exergy of a system in equilibrium with its environment is zero*".

Actually, we can identify four components of exergy $Z$, namely [62]:

$$Z = Z_k + Z_p + Z_{\text{ph}} + Z_{\text{ch}} \tag{3.64}$$

where:

$$\begin{aligned} Z_n &= \text{kinetic exergy} \\ Z_p &= \text{potential exergy} \\ Z_{\text{ph}} &= \text{physical exergy} \\ Z_{ch} &= \text{chemical exergy} \end{aligned}$$

The kinetic and potential exergy components have the same meaning as the corresponding energy terms and can be neglected when analyzing most common industrial processes.

Physical exergy $Z_{\text{ph}}$ is the work that can be obtained by taking a substance via reversible physical processes from an initial state ($T$, $p$) to the state determined by the temperature and pressure of the environment $(T_e, p_e)$. Physical exergy is very important when optimizing thermal and mechanical processes including heat engines and power plants. But it plays a secondary or negligible role in large-scale systems, e.g., chemical or metallurgical processes at the industrial level. In these processes, chemical exergy is the exergy component that plays a dominant role for resource accounting or environmental analysis purposes. Chemical exergy is the work that can be produced by a substance having temperature $T_e$ and pressure $p_e$ to a state of thermodynamic equilibrium with the datum level components of the environment. Chemical exergy has a component related to the chemical reactions taking place in isolation, and a component related to the diffusion of the reaction products into the environment.

Now, let us formulate mathematically the concept of exergy and derive its relation to entropy. Consider an *isolated system*, called the *total system or universe,* composed of the *system* of interest and its *surrounding* (*environment*). The

surrounding is assumed sufficiently large so as to be regarded as a *heat reservoir* with constant temperature $T_e$ and pressure $p_e$. According to the second law, expressed in terms of the entropy $S_{\text{total}}$ of the total system, we have:

$$\mathrm{d}S_{\text{total}} = \mathrm{d}S + \mathrm{d}S_e \geq 0 \tag{3.65}$$

where $S$ and $S_e$ are the entropies of the system and its environment, respectively. By virtue of the first law (Eq. 3.12), the change $\mathrm{d}U$ of the system's internal energy is equal to:

$$\mathrm{d}U = \delta Q - \delta W + \mathrm{dE}_{\text{ch}} \tag{3.66}$$

where $\delta Q$ is the heat added to the system, $-\delta W$ is the work done by the system (which has negative sing), and $\mathrm{d}E_{\text{ch}}$ is the net chemical energy entering the system. The net chemical energy is given by

$$\mathrm{d}E_{\text{ch}} = \sum_i \mu_{i,R}\mathrm{dN}_i \tag{3.67}$$

where $\mu_{i,R}$ is the chemical potential, and $N_i$ is the number of modes of the component $i$. The heat that leaves the reservoir and enters the system is equal to:

$$\delta Q = T_e(-\mathrm{d}S_e) \leq T_e\mathrm{d}S \tag{3.68}$$

Introducing Eq. (3.68) into Eq. (3.66) gives:

$$\delta W \leq -\mathrm{d}U + T_e\mathrm{d}S + \mathrm{dE}_{\text{ch}} \tag{3.69}$$

Now the net work $\delta W$ done by the system can be split so:

$$\delta W = \delta W_u + P_e\mathrm{d}V \tag{3.70}$$

where $\delta W_u$ is the useful work that can be done by the system, and $P_e\mathrm{d}V$ is the work spent for the system's expansion against the environment. Combining the inequality Eq. (3.69) with Eq. (3.70), we get:

$$\delta W_u \leq -\mathrm{d}U + T_e\mathrm{d}S - p_e\mathrm{d}V + \mathrm{d}E_{\text{ch}} = -(U - T_eS + p_eV - E_{\text{ch}}) \tag{3.71}$$

The extensive quantity:

$$Z = U - T_eS + p_eV - E_{ch} \tag{3.72}$$

is exactly the thermodynamic entity called *exergy* of the system. Here $-T_eS$ represents the entropic (or heat) loss of the reservoir, $p_eV$ represents the available $pV$ work, and $E_{\text{ch}}$ is the available chemical energy. Combining Eq. (3.71) with Eq. (3.72) gives the inequality:

$$\mathrm{d}Z + \delta W_u \leq 0 \tag{3.73}$$

which states that the exergy change of the system, plus the change of the useful work done by the system, is non-positive. If no useful work is extracted by the system (i.e., if $\delta W_u = 0$), then Eq. (3.73) reduces to:

$$\mathrm{d}Z \leq 0 \tag{3.74}$$

which says that the change of exergy of the system is non-positive, i.e., the *exergy* (available work) of the system is *decreasing* (consumed) or *remains constant*. Exergy is decreasing if the process is irreversible, while it remains constant if the process is reversible (i.e., at equilibrium).

The above analysis shows that the *entropy law:*

$$\mathrm{d}S_{\text{total}} \geq 0 \tag{3.75a}$$

is equivalent to the *exergy law*:

$$Z + \delta W_u \leq 0 \tag{3.75b}$$

Therefore, at the macroscopic level we can use the second law in terms of exergy (as postulated at the beginning of this section) without considering or measuring directly entropy in a total isolated system (thermodynamic Universe). A rich bibliography on exergy and its use can be found in [63].

*Remarks*

1. If $E_{\text{ch}} = 0$ and the temperature $T$ of the system is always equal to $T_e$, then the exergy (3.72) is equal to:

$$Z = U - TS + p_e V + \text{ Constant}_1$$

Moreover, if $V$ is constant, then:

$$Z = U - TS + \text{ Constant}_2$$

The entity $U - TS$ is the *Helmholtz free energy* $A = U - TS$ defined in Eq. (3.27). Thus, under constant volume:

$$\mathrm{d}A < 0$$

if the process is to go forward, or $\mathrm{d}A = 0$ if the process is in equilibrium.

2. If $p$ is constant, we get:

$$Z = U - TS + pV + \text{Constant}_3$$

The quantity $G = U + pV - TS = H - TS$ is the *Gibbs free energy* defined in Eq. (3.28) with $H = U + pV$ being the *enthalpy* of the system. Thus, under constant pressure conditions, if $dG < 0$, then the process can occur (go forward) spontaneously, and, if $dG = 0$, then the process is at equilibrium. It is noted that the Gibbs free-energy conditions for a process to go forward spontaneously or be at equilibrium combine the first and second laws of thermodynamics, and for this reason is also known as the "combined law of thermodynamics".

## 3.7 The Third Law of Thermodynamics

The *third law* was first formulated by *Walther Nernst* in 1966 during his efforts to deduce equilibrium constants from thermal data [64]. The third law is often called *Nernst's theorem*. A first form of the third law states that:

> "*As temperature tends to absolute zero, the entropy of a system approaches a minimum well-defined constant*".

This means that entropy depends on temperature and leads to the formulation of the concept of absolute zero. That is, if all the thermal motion (kinetic energy) of particles (molecules) could be removed, a state called "absolute zero" would occur (Absolute zero = 0 K = −273.15 °C). It is remarked that the entropy's minimum value is not necessary zero. However, *Max Planck,* working on the statistical formulation of entropy, stated the third law in 1913 as follows [65].

> "*The entropy of each pure element of substance in a perfect crystalline state is zero at absolute zero*".

This is because, at absolute zero, only one way to arrange the molecules in the lattice is possible, and all molecules will be at their lowest energy state.

*Nernst* has found that the temperature dependence of the equilibrium constant depends only on enthalpy differences, and, by examining the temperature dependence of the free energy, he concluded that entropy differences, $\Delta S$, must become zero at absolute zero. To be consistent with Max Planck's result, he accepted that at absolute zero the absolute value of entropy of any system or process is zero, too, i.e.:

$$\lim_{T \to 0} \Delta S = 0 \quad \text{and} \quad \lim_{T \to 0} S = 0$$

The above postulate was later (1923) restated by G.N. Lewis and M. Randall as [66]:

"*If the entropy of each element in some perfect crystalline state be taken as zero at the absolute zero of temperature, every substance has a finite positive entropy; but at the absolute zero of temperature the entropy may become zero, and does so become in the case of perfect crystalline substances*."

However, today there are some rare exceptions to this postulate [66].
The third law stated in another way (verified experimentally) states [4]:

"*It is not possible to cool any substance to absolute zero by a finite number of cyclical operations, no matter how idealized they are*".

This means that a continually increasing, ultimately infinite, amount of work is needed to remove heat from a body as its temperature approaches absolute zero. It is recalled that, by virtue of the second law, reducing the entropy of a system implies the increase of the entropy of its environment. For his discovery of the third law, Nernst has was awarded the 1920 Nobel Prize in Chemistry [64].

## 3.8 The Fourth Law of Thermodynamics

To the present time, there is no principle or statement about energy, matter, time, or temperature globally accepted as being the fourth law of thermodynamics. Over the years, many physicists and thermodynamics scientists have formulated principles aiming to have them established as potential fourth or fifth laws of thermodynamics. Among them there are a few that are frequently cited as the *fourth law*, two of which are the "*Odum–Lotka*" *maximum energy flow and empower* "*principle*", and the principle expressed by the "*Onsager reciprocal relations*". A brief discussion of these two principles and a short list of some more "candidate fourth laws" follow. An extended list is provided in [67].

### *3.8.1 Lotka's Maximum Energy-Flux Principle*

This principle states:

"*Natural selection tends to make energy flux a maximum, so far as compatible with the constraints to which the system is subject*".

In this principle [68, 69], the constraints refer to the existence of an unutilized quantity of matter and available energy. Specifically Lotka states that: "*so long as there is present an unutilized residue of matter and available energy in every instance, considered, natural selection will so operate as to increase the total mass of the organic system, to increase the rate of circulation of matter through the system, and to increase the total energy flux through the system*".

Lotka extended this natural selection principle to *evolution*, stating that for evolution two sub-processes take place, namely: *selection of influences* and *generation of influences*. The first influences select and the second provide the material for selection. If the material provided for selection is limited (as, for example, in chemical reactions), the range of operation of the selective influences is accordingly limited. This is not so in the case of organic evolution where there is an element of uncertainty. He states that inorganic evolution it is at least a priori probable, i.e., among the large number of types presented for selection, the ones that will occur sooner or later will give the opportunity for selection to go in the direction of maximum energy flux as stated previously. Therefore, the law of selection becomes also the *law of evolution*, i.e.:

> "*Evolution, in these circumstances, proceeds in such direction as to make the total energy flux through the system a maximum compatible with the constraints*".

### 3.8.2 Odum's Maximum Rate of Useful-Energy-Transformation Principle

Lotka's principle of *maximum energy flux* was developed further by Odum, Tribus, and Pinkerton [70–74], combining it with the *maximum power theorem*, well known in the electrical systems field, and properly quantifying the law of natural selection and evolution. The word "*power*" in this setting was defined energetically, as the *rate of useful transformation* (hence the name of the principle in the title of the present section).

It was Ludwig Boltzmann who, for the first time, stated that the fundamental issue of contention in the life-struggle in the evolution of the organic world is "*available energy*". Lotka's principle is in agreement with Boltzmann's statement and expressed, in other words, declares that the life-preservation advantage goes to the organisms that have the most efficient capturing mechanisms for the available energy.

Odum's statement of the principle of *maximum power* (or rate of useful energy transformation) is based on the electrical engineering concept of "*impedance matching*" [75] and has been extended to biological and ecological systems. Actually, Odum regarded the world as an ecological-electronic-economic engine and pointed out that this principle "provides a potential guide to understanding the patterns of ecosystem development and sustainability". In several publications, Odum provided, as a practical example of the principle of maximum power, the *Atwood machine* [76].

The exact statement of the maximum power principle as given by Odum [72] is as follows.

> "*During self organization, system designs develop and prevail that maximize power intake, energy transformation, and those uses to reinforce production and efficiency*".

He further pointed out in 1994 that: "in surviving designs, a matching of high-quality energy with larger amounts of low-quality energy is likely to occur". In his 1995 publication [72], Odum provided a corollary of maximum power calling it "*maximum empower principle*", to clarify that higher level transformation processes are equally important as the low-level processes, and so excluded the possible misunderstanding that maximum power means that low-level processes have higher priority. For its generality, the combination of Lotka's and Odum's principles has been referred to by C. Giannantoni as the "*Fourth Law of Thermodynamics*" under the name "*The Maximum Empower Principle*" [77, 78]. The concepts and methods that are based on the maximum empower principle have been applied to many ecological and socioeconomic systems where self-organization and optimal exergy acquisition take place (see e.g., [79–83]).

### 3.8.3 Onsager Reciprocal Relations

Onsager's "*reciprocal relations*" (1931) are general and are valid for various pairs of thermodynamic forces and flows in a wide repertory of physical-chemical processes. These relations show that some particular relations between flows and forces in systems not in equilibrium are identical, subject to the condition that some local equilibrium concept exists.

The experiment has shown that temperature gradients (differences) in a system result in heat flow from the hotter to the colder parts of the system. In a similar way, mass flows from high-pressure to low-pressure areas. Furthermore, in cases where there are differences in both temperature and pressure in a system, heat flow can be caused by pressure differences and mass flow can be caused by temperature differences, and the ratio "*flow over force*" in each case has the same value, i.e.:

"Heat-flow/Pressure difference = Mass-density flow/Temperature difference"

Onsager has shown [84, 85] that this type of reciprocal relationship is a consequence of the principle of microscopic reversibility in statistical mechanics, and so they are valid in the *statistical ensemble* sense.

One way to formulate the Onsager reciprocal relations is to start with the formula of the entropy production density [86–88]:

$$s = \frac{1}{T}\sum_{k=1}^{n} \mathbf{J}_k[F_k - (\mathrm{grad}\mu_k)_T]$$

where $\mathbf{J}_k = \rho_k(\mathbf{v}_k - \mathbf{v})$ are the vectors of the diffusion flows with respect to any reference velocity $\mathbf{v}$, $\rho_k$, and $\mathbf{v}_k$ are the densities and velocities of the components, $F_k$ are the body forces per unit amount of each chemical species, $T$ is the absolute temperature, the scalars are the chemical potentials, and "grad" is the gradient operator. The quantity in the brackets on the right-hand side of the above equations is written in the form as follows:

$$F_i - (\text{grad}\mu_i)_T = \sum_{k=1}^{n} R_{ik}\mathbf{J}_k$$

where $\mathbf{R} = [R_{ik}]$ is a matrix of coefficients $R_{ik}$, called the *Onsager coefficients*. The second law of thermodynamics and the above form of *s*imply that the Onsager coefficient matrix is positive definite. Using the principle of microscopic reversibility Onsager has shown that $\mathbf{R}$ is also symmetric, i.e.,

$$R_{ik} = R_{ki} \quad \text{for all} \;\; i, k$$

These relations are exactly the so-called *Onsager reciprocal relations*. Another easy-to-follow presentation is provided in [89]. For the discovery and proof of his reciprocal relations, Onsager was awarded the 1968 Nobel Prize in Chemistry. Many workers have attempted to provide a phenomenological (non-statistical) proof of the Onsager reciprocal relations, but such a straightforward proof seems to remain unavailable. Jozsef Verhas in [88] showed that presuming the validity of Onsager reciprocal relations, exact proofs can be constructed by Newton's second and third laws, which means that a phenomenological proof of the reciprocal relations would be equivalent to a proof of the fundamental laws in physics.

### 3.8.4 Some Further Fourth-Law Statements

In the following, eight more principles or statements that are potential candidates for consideration as the fourth law are listed in chronological order. The details of their formulation can be found in the respective references.

1. **Harold Morowitz (1992)**
   The flow of energy from a source to a sink through an intermediate system orders that system [90, 91].
2. **Per Bak (1994)**
   Slowly driven systems naturally self-organize into a critical state [92].
3. **Michael Moore (1996)**
   In every contact of matter with matter, some matter will become unavailable for future use; thus, some matter is wasted ("some" is not to imply "minute", and waste means an irrevocable transfer that is beyond recycling) [93].
4. **Pierre Perrot (1998)**
   A fourth law can be any statement postulating the existence of an upper limit to the temperature scale (between $10^{11}$ and $10^{12}$ K) [3].
5. **Stuart Kauffman (2000)**
   Biospheres maximize the average secular construction of the diversity of autonomous agents and the ways those agents can make a living to propagate further. (An autonomous agent is a self-reproducing system able to perform at least one thermodynamic "work cycle") [94].

6. **R.E Morel and George Fleck (2006)**
   Systems increase entropy at the maximum rate available to them. (This implies that identical systems under identical conditions behave identically) [95].
7. **Murad Shibli (2007)**
   Considering time as a mechanical variable for a closed system with moving boundaries composed of homogeneous isotropic cosmic fluid, the system will have a negative pressure equal to the energy density that causes the system to expand at an accelerated rate. Moreover, the momentum associated with time is equal to the negative of the system's total energy [96].
8. **Philip Carr (proposal for fifth law, 2008)**
   Wherever possible, systems adapt to bring d$S$/d$t$ to a maximum over some variably sized window of visibility or, more explicitly, an open system containing a large mixture of similar automatons, placed in contact with a non-equilibrated environment, has a finite probability of supporting the spontaneous generation and growth of self- constructing machines of unlimited complexity. (This means that any system involving a large number of similar elements is potentially able to self-organize, using energy and matter taken from its environment) [97].

*Note: In the above principle number eight*:

- *Automaton* is a particle or physical object or machine that can interact with other particles, objects, or machines.
- *Machine* is a mechanism that can extract work from the conversion of high-energy reactants into low-grade waste products plus heat.
- A *non-equilibrated environment* is an environment containing materials that are able to react together in order to release energy plus waste products.
- *Unlimited complexity* implies that, within the constraints imposed by available energy, materials, possible pathways, and time, the system could evolve indefinitely to become a more and more effective mechanism for conversion of the materials found within the environment into waste products, waste energy, and machine structure.

## 3.9 Branches of Thermodynamics

The thermodynamics ancestors are the laws and inventions developed in the seventeenth century, such as the thermal argumentation law, the vacuum pump, the gas laws, and the steam engine. As a science, thermodynamics was developed in the seventeenth and eighteenth centuries, and in the nineteenth and twentieth centuries has embraced many theoretical- and applications-oriented areas of science, technology, and human life [2–4, 98]. The branches of thermodynamics may be divided into three main categories, namely: (*i*) *traditional branches,* (*ii*) *natural systems branches, and* (*iii*) *modern branches*.

### 3.9.1 Traditional Branches

The three core traditional branches, i.e., classical, statistical, and quantum thermodynamics, were discussed in previous sections. Here a short outline of the following additional traditional branches will be given:

- Chemical thermodynamics
- Engineering thermodynamics
- Biochemical thermodynamics
- Surface thermodynamics

**Chemical Thermodynamics**
Chemical thermodynamics is concerned with the study of energy, work, and entropy of physical state (phase) changes and chemical reactions (including equilibrium and non-equilibrium conditions) using the laws and tools of thermodynamics. The first publications on chemical thermodynamics are [99, 100]. Gibbs for his 1876 publication [101], and von Helmholtz for his 1882 publication [102] are recognized as the founders of the classical chemical thermodynamics, and Gilbert Lewis, Merle Randal, and E. Guggenheim, are considered as the fathers of modern chemical thermodynamics for their works published in 1923 and books [103, 104] in 1933, which have been integrated and presented in a unified way as a well-defined scientific branch in [105]. The chemical energy is due to the potential of chemical compounds to be transformed to other compounds via chemical reactions. This energy can be released or absorbed depending on the difference between the energy content of the reactants and the products. The state functions that are used in the study of chemical thermodynamics have already been presented and are internal energy ($U$), enthalpy ($H$), entropy ($S$), and Gibbs free energy ($G$). Chemical processes obey the laws of thermodynamics, as explained in previous sections, which lead to the Gibbs free-energy (or exergy) relation:

$$\mathrm{d}G = \mathrm{d}U - T\mathrm{d}S + p\mathrm{d}V \leq 0,$$

derived in Sect. 3.6.3. This relationship implies that any spontaneous reaction has negative free energy d$G$ and that at equilibrium d$G$ is zero. Basically, conventional chemical thermodynamics deals with processes at or near equilibrium. The processes that are far from equilibrium require special attention and methods and have been extensively studied by Ilya Prigogine, who discovered important new structures and phenomena of nonlinear and irreversible thermodynamics that find applications in a large repertory of fields (e.g., cancer grow, traffic flow, biological systems, etc.).

**Engineering Thermodynamics**
Engineering thermodynamics deals with the application of thermodynamics concepts, laws, and methods for solving engineering problems. Such problems include the calculation of fuel efficiency in engines, the energy and exergy calculation in

man-made systems (e.g., power installations, refineries, nuclear reactors, etc.), the energy conservation and economy, and the development of easy-to-use design plots/maps and tables from empirical data [106–108]. In particular, the *exergy-analysis* method provides the means for locating the cause and the real magnitude of energy resource waste and loss in the process at hand. This enables and facilitates the design of novel systems with higher energy efficiency or the performance improvements of existing systems. The integration of the exergy-analysis concepts with engineering economic concepts is called "*thermoeconomics*" and includes the identification of the real source of cost at the component level, the capital investment costs, and the operational and maintenance costs. In sum, thermoeconomics is concerned with the cost minimization in engineering systems on the basis of exergy considerations. A subclass of engineering thermodynamics is "*chemical-engineering thermodynamics*" which deals with applications of an exclusively chemical-engineering character, such as chemical reaction equilibrium, solution processes, osmotic processes, fluid mixing, etc. [109].

**Biochemical Thermodynamics**

Biochemical thermodynamics is primarily concerned with applications of thermodynamics in biochemistry, but very often is considered together (or synonymously) with biological thermodynamics or biotherm dynamics which refers to bioenergetics, i.e., to the study of energy movement and transformation in biological processes. Biothermodynamics deals with the quantitative analysis of the energy flows taking place inside and between living organisms, cells, structures, and communities. Notable references in this area include [110–116]. The energy available to do work during a chemical reaction is given quantitatively by the change in Gibbs free energy, which is equal to $\mathrm{d}G = \mathrm{d}H - T\mathrm{d}S$, where $H = U + pV$ is the enthalpy. The enthalpy, which is an extensive property, is a state function of a process (because it is defined via the state functions $U, p$, and $V$), and plays a fundamental role in biochemical-reaction thermodynamic calculations. It is remarked that enthalpy can actually be measured in practice. Energy flows in biological organisms originate from photosynthesis through which the sunlight energy is captured by the plants. All internal dynamic biochemical processes, such as ATP hydrolysis, DNA binding, etc., are included in the field of biothermodynamics.

**Surface Thermodynamics**

This branch, also called *adsorption thermodynamics,* is concerned with the behavior of atoms, molecules, bacteria, etc. (i.e., chemical species) near surfaces, and was founded by Gibbs. Adsorption is the process in which atoms or molecules, etc. (adsorbates) attach to a solid or liquid surface (adsorbent) originating from a bulk phase (solid, liquid or gas). Adsorption at the molecular level is caused by the attractive interactions between a surface and the species being adsorbed. Thermodynamically at the adsorption equilibrium, the Gibbs free energy is minimized. This means that adsorption (e.g., bacterial adhesion) takes place if, by itself, this leads to a decrease of the free energy (which is an isothermal-isobaric thermodynamic potential). Otherwise, it is not favored. Adsorption is directly applied to

filtration and detergent-action processes, and it is also very important in heterogeneous catalysis, electrochemistry, adhesion, and lubrication. Adsorption is typically expressed by the quantity of adsorbate on the adsorbent as a function of its pressure (if it is a gas) or concentration (if it is a liquid) at constant temperature (isothermal function). Two classical isothermal relations are the Freundlich–Kiister (1894) and the Langmuir (1916) equations [117, 118]. Important publications in the bacterial adhesion field include [119, 120]. A complete set of online surface science tutorials and lecture courses can be found in [121].

### *3.9.2 Natural Systems Branches*

This class includes the branches of thermodynamics that deal with large-scale natural phenomena. Here the branches related to the following natural and cosmological systems and phenomena will be reviewed:

- Meteorological systems
- Ecological systems
- Geological systems
- Earthquakes
- Cosmological phenomena

**Meteorological Systems**: The study of the phenomena involved in the energy transformation and movement that takes place in the atmosphere is called *atmospheric* (*or meteorogical*) *thermodynamics*. These phenomena which obey the laws of thermodynamics include among others, the following: insolation, terrestrial radiation-energy balance, tropospheric transport of surface heating and cooling, atmospheric tides, atmospheric moisture, evaporation and latent heat, heat advection, atmospheric stability, adiabatic processes, cloud, fog and precipitation, planetary-scale tropospheric phenomena, micrometeorological phenomena, atmospheric turbulence, mesoscale weather systems, atmospheric electricity and light phenomena, etc. Suitable references on atmospheric thermodynamics include [122–125].

**Ecological Systems**: The application of the principles and laws of thermodynamics in ecological systems is called "*ecological thermodynamics*". In particular, ecological thermodynamics is concerned with the study of the states and evolution of ecosystems through the use of energy-, entropy-, exergy-, and emergy-flow analysis. The founders of ecological thermodynamics are the Odum brothers (Howard Odum, ecologist, and Eugene Odum, zoologist). The first book in ecology was published in 1953 [126], but the full application of thermodynamics to ecology started in the 1970s [127–131]. The fundamental principle of ecological thermodynamics is that, while materials cycle in ecosystems, energy flows through them [132–135]. Flows of energy originate from plants to herbivores and goes to carnivores (and to decomposers all along the food chain). To measure the energy flow

through ecosystems, use is made of several measures of *efficiency* with which energy is converted into "*biomass*". The typical measures of energy efficiency in ecological anthropology are as follows: (i) output/input ratio = energy acquired/energy expended $= E_a/E_e$ and (ii) net return rate $= (E_a - E_E)$/time needed to acquire energy (= net energy/labor time). In the case of a population with limited energy, the increase in efficiency of energy acquisition is *adaptive* (since it increases the overall amount of energy that can be used), but, if the population's energy is unlimited, the total amount of energy captured is *non-adaptive* (because there is always sufficient "food supply").

**Geological Systems**: Chemical thermodynamics is the branch of thermodynamics used to study geological systems (minerals, magmas, rocks) and to understand the chemistry of geothermal systems (i.e., geothermal fluids that are flashed, cooled, or mixed). Chemical thermodynamics, when applied to geological systems, forms the branch that is called "*geochemical thermodynamics*" and enables us to predict what minerals will be produced under different conditions (the so-called *forward modeling*), and at the same time allows us to use mineral assemblages and mineral compositions to determine the conditions under which a rock was formed (thermobarometry). Usually, the calculations required are very complex; they need reliable thermodynamic data and in present days are usually performed by suitable computer programs (the so-called *thermodynamic modeling programs*). Similarly, a variety of computer programs are available that help in the thermobarometric calculations, i.e., the quantitative determination of the pressure and temperature at which a metamorphic or igneous rock reached chemical equilibrium. Basic publications on geochemical thermodynamics include [136–141].

**Earthquakes**: *Earthquake thermodynamics* is a branch of thermodynamics that examines the geophysical processes and phase transformations that form the triggering mechanisms of earthquakes in the Earth's interior. It is a very important area of research because thermodynamics has strong predictive power. Thus, modeling the earthquake and the Earth's interior using the concepts and laws of thermodynamics is expected to contribute much to successful earthquake prediction for the benefit of human safety and society. The processes involved in the Earth's and universe's evolution are irreversible, and the nonlinear interactions that take place lead to the formation of *fractal structures*. The interior boundaries have their origin in structural phase transformations. Actually, earthquake thermodynamics provides a microscopic model of earthquake sources. The precursory phenomena produced by the increasing stresses can be efficiently studied with the help of the physics of defects and the abnormalities in electric polarization and electromagnetic radiation that appear before the occurrence of the earthquake. A rich collection of studies in this area is provided in [142].

**Cosmological Phenomena**: It is globally accepted that cosmological phenomena obey the laws of physics and thermodynamics, but the existing theories of cosmic evolution suffer from the following "*paradoxon*". On the one hand, Einstein's equations are adiabatic and reversible, and so they cannot explain on their own the origin of cosmological entropy. On the other hand, the quantum nature of these equations make the results obtained very sensitive to quantum

subtleties in curved space-times. These difficulties have been studied by many cosmological and thermodynamics scientists, e.g., John Wheeler, Jacob Bekenstein, Stephen Hawking, Robert Wald, Ilya Prigogine, Benjamin Gal-Or, Richard Tolman et al. [143–150]. In sum, the study of the *thermodynamics of the universe or cosmological thermodynamics* depends on which form of energy considered as the dominant one, i.e., *radiation* (relativistic particles) or *matter* (non-relativistic particles). Very important is the reinterpretation of the matter–energy stress tensor in Einstein's equations provided by Prigogine et al. [145] which modifies the standard adiabatic conservation laws, thereby including irreversible-matter creation which corresponds to an irreversible energy flow from the gravitational field to the created matter constituents. As a result, the second law of thermodynamics implies that space–time transforms to matter, whereas the inverse is not allowed. Thus, the cosmological history starts from an instability of the vacuum rather than the usual initial singularity associated with the "Big Bang". In this theory, the instability at the origin of the universe is the result of fluctuations on the vacuum where black holes act as membranes that stabilize these fluctuations. The study of black holes in terms of radiation, entropy, energy, temperature, etc., is called "*black hole thermodynamics*" [143, 150].

### *3.9.3 Modern Branches*

Here, the following branches, most of which have a theoretical or abstract character, will be outlined:

- Non-equilibrium thermodynamics
- Economic-systems thermodynamics
- Business-systems thermodynamics
- Computer-systems thermodynamics
- Life thermodynamics

**Non-equilibrium thermodynamics**: This theoretical branch deals with systems that are far from thermodynamic equilibrium. The study of systems that are in equilibrium (the so-called equilibrium thermodynamics) belongs in its greater part to the "classical thermodynamics" outlined in some detail in previous sections of this book [151]. The study of systems that are near to equilibrium and are governed by Onsager reciprocal relations is called "*near-equilibrium thermodynamics*" or "*linear thermodynamics of irreversible processes*". Actually, the majority of systems are not isolated from their environment, thus exchanging energy and/or matter with other systems, and so they are not in thermodynamic equilibrium. Non-equilibrium thermodynamics is distinguished in *classical and extended* non-equilibrium thermodynamics. The former satisfies the requirement of *local thermodynamic equilibrium,* and the latter violates this requirement [152]. Local thermodynamic equilibrium of matter implies that the system at hand can be divided (in space and time) into small "cells" (of infinitesimal size) that satisfy to a

good degree the classical thermodynamic equilibrium conditions for matter. Important contributions in the field of non-equilibrium thermodynamics were made by Jou, Casas-Vaguez, Lemon, Onsager, Herwig, Prigogine, Mahulikar, Glansdorf, Meixner, Mazur, and de Groot [89, 152–157]. An important concept that had a strong influence on the development of the field is Prigogine's and Glansdorf's concept of "*dissipative structures*", a new form of dynamic states of matter, which lead to a spontaneous "*self-organization*" of the systems both from the "*space order*" and "*function*" points of view. More specifically, dissipative structures refer to the capability of complex systems to transfer their entropic changes and costs to other parts of the system, a property that is applicable to any type of physical or human-life system. Another influential concept in the area of non-equilibrium dynamics is the concept of "*maximum entropy production* (**MEP**)" originated by Paltridge [158] and fully developed by Swenson [159]. It is remarked here that Prigogine [160] presented a theorem of "*minimum entropy production*", which does not contradict Swenson's MEP law since it is applicable to the linear regime of a near stationary system in which the MEP principle does not actually apply. The MEP law is applicable to nonlinear systems and determines the most probable state path among all the possible ones [161].

**Economic-Systems Thermodynamics**: This is a branch that focuses on the interface between thermodynamics and economics. Actually, the laws of thermodynamics have a strong influence on economic activity and theory. For example, the implication of the conservation of energy and mass (first law) is that raw-material inputs to economic processes are not actually "consumed", because, after their extraction from the environment, they return finally to it as waste. The second law is also applicable here since the economic processes use "*low-entropy*" (high-"exergy") raw materials (e.g., high- grade metal ores and fossil fuels and discard "*high-entropy*" (low-exergy) wastes. The true economic importance of the second law is that exergy is not conserved and is a successful measure of resource quality and quantity of both materials and energy (mapped to "money" and "development"). Whenever exergy (available work) is consumed, a kind of entropy is generated. This "spent work" is expelled into the environment and can no more be used (thus called "anergy"). The first-known publication in economics that studies economic entropy is the 1971 book of Nicholas Georgescu-Roegen [162]. A class of applications of thermodynamics to economic systems focuses on the statistical-thermodynamics formulation and employs arbitrary probability distributions for selected conserved economic quantities. Here, like statistical mechanics, the macroeconomic variables are treated as the mean value of microeconomic variables and are determined by computing the partition function starting from an arbitrary function [163]. A collection of papers on the subject of economics and thermodynamics is provided in [164], and a good reference to the study of economic systems through the second law (exergy) where it is shown that exergy is no less a "factor of production" than "labor or capital" is [165]. The study of economics via thermodynamics concepts is also called *thermoeconomics*, a term coined by Myron Tribus [166, 167]. In general, the economic activity is studied by considering the economic system as a dissipative system (i.e., a non-equilibrium

thermodynamics-like system) where exergy is consumed and an exchange of resources, goods, and services takes place. This point of view has also been used in approaching the establishment and evolution of the alternative social/socioeconomic systems [168].

**Business-Systems Thermodynamics**: Here the energetic issues of the operations taking place in business systems are studied through the application of the thermodynamics concepts and laws. New concepts like the *corporate entropy and business thermodynamic efficiency* were developed and used [169–171]. Closely related to business thermodynamics is the field of *business chemistry* where a business is modeled as a chemical laboratory, test tube, etc., and people are regarded as chemical elements or human molecules that participate in the business operation [172, 173]. Two basic concepts of business thermodynamics are the *business-molecular organism* (**BMO**) concept and the *mental-flow rate* (**MFR**) concept, dealing with metabolism, efficiency, energy balance, and mass-balance metaphors. The *work output* includes the projects' and other deliverables which contribute to the increase of *return on investment* (**ROV**). But, according to Little [172, 173], because chemistry is concerned with the properties and changes of matter, any business can be performed more efficiently, if it includes in its personnel a "*chemistry professional*". To better satisfy this requirement, many universities around the world offer special courses with at least a "business–chemistry" major. The use and study of business-thermodynamic efficiency help in improving the overall business efficiency ratio and developing more successful "*business-operation-efficiency analysis*" (**BOEA**) [174]. In their effort to maximize their efficiency and face their complex problems, modern businesses use the so-called "*business-and-information-technology consultants*" who have special skills and know-how at several levels, namely, technological, industrial, market-level, etc. that are applicable to both short-term and long-term horizons [171].

**Computer-Systems Thermodynamics**: This branch deals with the application of thermodynamics concepts and laws for the analysis, design, and operation of computer systems. Computer systems are dynamic physical systems and therefore obey the laws of physics and thermodynamics. It is very important for their development and use to identify features and parameters that are time-independent and invariant (e.g., constants of motion). Present day computers are large and involve a large number of small-scale components that are working together and affect the overall performance of all. Therefore, the application of the concepts of statistical and quantum thermodynamics is a feasible way to study the macroscopic behavior of a computer as an aggregation of the statistical behavior of its small-scale building elements and components, which are now reaching the microscopic size. Examples of statistical and quantum analysis in computer systems are statistical mechanics of cellular automata, simulated annealing, quantum computation, quantum cryptography, and computational complexity [175–180].

**Life Thermodynamics**: This branch is concerned with the study of life and living organisms through the laws and methods of thermodynamics. In this framework, life is any animate matter [181] or molecular structure, which in contrast to inanimate is alive and is characterized by driving force, exchange force,

thermodynamic force, and induced movement. Historically, the first person to hypothesize that the processes of life are thermodynamic processes was Ludwig Boltzmann (1875). Working on the theories and results of Clausius and Thomson (Kelvin), he arrived at the conclusion that [182, 183]:

> '*The struggle for existence of animate beings is for* (*negative*) *entropy obtained from the sun and the earth*'.

In 1876, Richard Sears McCulloch, building on the Carnot cycle concept, arrived at the conclusion that [184]:

> "*An animal performing mechanical work must from the same quantity of food generate less heat than one abstaining from exertion, the difference being precisely the heat equivalent of that work*".

In 1944, Erwin Schrödinger published his famous book dealing with the basic question "what is life?" [185]. In this book, he connects the fields of "molecular biology", "biotechnology", and "biothermodynamics". His views have influenced modern research in these three branches, leading to major discoveries such as the chemical "codescript" existing in the *nucleic acids,* and the *helical structure* of the DNA (DeoxyriboNucleic Acid). In the thermodynamics field, Schrödinger revealed the ability of life to move towards "order", thus ensuring that nature not fall into thermodynamic uncertainty, randomness, and chaos. This is summarized in his overall conclusion that:

> '*Life feeds on negative entropy*'.

Today, the study of *life processes* (including human-life processes) utilizes the concepts of the *Gibbs free energy, enthalpy, exergy, and energy*, with which we are already familiar. Regarding the human being, it is globally accepted that the basic needs such as "eating", are thermodynamic processes, but there are many "debatable" theories about the human's higher needs and life performance [112, 184, 186–189].

A chronological representative (but non-exhaustive) list of views and statements about life and thermodynamics from the middle of the twentieth century to the present is the following

- **Alfred Ubbelohde** (**1947**): "There is reasonable expectation that more information about their (i.e., the laws of thermodynamics) bearing on life will be obtainable…as the result of measurements of the energy and entropy changes accompanying the activities of living organisms" [190].
- **Robert Lindsay** (**1959**): "Man's whole struggle …may be interpreted either as an instinctive or conscious and deliberate attempt to replace disorder with order, in other words, to consume entropy" [191].
- **Ilya Prigogine** and **Isabelle Stengers** (**1984**): "Far from equilibrium, the system may still evolve to some steady state, but in general this state can no longer be characterized in terms of some suitably chosen potential (such as entropy production for near equilibrium states)…. Dissipative structures actually correspond to a form of supramolecular organization" [192, 193].

- **Max Ferutz** (**1987**): "We live on free energy and there is no need to postulate negative entropy" [194].
- **Thomas Fararo** (**1992**): "Any special emergence of order-from-disorder phenomenon in our field (sociology) must be accounted for by the mechanism of social interaction not by a vague appeal to some thermodynamic situation" [195].
- **Ichiro Aoki** (**1994**): "The two-stage principle of entropy production (the early increase and the later decrease) is universally applied to aging over the life span of humans and other animals" [196].
- **Eric Schneider** and **James Kay** (**1995**): "Life emerges because thermodynamics mandate order from disorder whenever sufficient thermodynamic gradients and environmental conditions exist" [197].
- **Donald Haynie** (**2001**): "Any theory claiming to describe how organisms originate and continue to exist by natural causes must be compatible with the first and second laws of thermodynamics" [113].
- **Fritjof Capra** (**2002**): "Life is a far from thermodynamic equilibrium system that has emerged, due to the flow of matter and energy, past the bifurcation point" (i.e., life is a Prigoginian type thermodynamic system) [198].
- **Eric Schneider** and **Dorion Sagan** (**2005**): "Life is a terrible and beautiful process deeply tied to energy, a process that creates improbable structures as it destroys gradients. Yet as a scientific discipline, the thermodynamics of life—a sub-discipline of non-equilibrium thermodynamics—remains esoteric within science and virtually unknown to the public" [199].
- **Georgi Gladyshev** (**2009**): "The phenomenon of life is easily understood as a general consequence of the laws of universe; in particular the laws of thermodynamics, and a *principle of substance stability*….The definition of a life as the biochemical-physical phenomenon can be given on the basis of the exact sciences, without mention of numerous private attributes of a living substance and without physically baseless models of mathematical modeling such as Prigoginian thermodynamics" [200].

The work of Gladyshev [201–203], which started in 1978, is now recognized as a new branch of thermodynamics called *macrothermodynamics* or *hierarchical thermodynamics* or *structure thermodynamics* [206]. This is a general theory applicable to all systems that are specified by the functions of states. A hierarchical system in this theory consists of a set of subordinate subsystems belonging to a hierarchy (structural, spatial, or time). Hierarchical thermodynamics deals with near-equilibrium linear systems possessing time-varying state functions.

The above views about life, despite their differences in the details, show that it is generally recognized that indeed the laws of thermodynamics have contributed and will continue to contribute to a better and deeper understanding of what is life and how it is maintained on Earth. A subarea of life thermodynamics called *human thermodynamics* [187, 188] covers not only the "physiology" processes of the human body, but also higher level human processes in life and society spanning the most delicate issues of human existence and stability such as human relationships,

mind, intelligence, free will, purpose, love, religion, etc. As such, it has received strong criticism and opposition by many thinkers [189, 204–209]. But as *Ingo Müller* states in his 2007 book on history of thermodynamics [208] "Most socio-thermodynamics subjects are more for the future of thermodynamics rather than to its history". It should be remarked here that the study of the origin of life through the second law of thermodynamics is a much more difficult and complicated problem and an issue of strong scientific "debate" which is still continuing. The role and influence of energy and thermodynamics in human life and society will be discussed in some more detail in Part 3 of this book (Chap. 10).

Other branches of thermodynamics include the following:

- Neuro-thermodynamics
- Network thermodynamics
- Molecular thermodynamics
- Aerothermodynamics

Closing this section on the branches of thermodynamics, we point out Albert Einstein's opinion about thermodynamics:

> '*Thermodynamics is the only physical theory of a general nature of which I am convinced that it will never be overthrown*'.

## 3.10 Entropy Interpretations

The entropy concept has received a variety of interpretations that in many cases are contradictory and misleading. The issue of the interpretation of the Clausius entropy concept was of concern very early in the history of thermodynamics. In 1873, Willard Gibbs pointed out that "the different sense in which the word entropy has been used by different writers is liable to cause misunderstanding". The story of the variety of entropy's interpretations is still continuing. Here we will discuss the principal of these interpretations of the thermodynamic and statistical concept of entropy. The information concept of entropy will be examined in the next chapter. Specifically, the following entropy interpretations will be considered here:

- Entropy as unavailable energy
- Entropy as disorder
- Entropy as energy dispersal
- Entropy as opposite to potential.

### *3.10.1 Entropy Interpretation as Unavailable Energy*

In this interpretation, "entropy for an isolated system is the quantitative measure of the amount of thermal energy *not available* to do work", i.e., it is the opposite of

available energy [210]. The second law states that entropy in an isolated system can never decrease, and so interpreting entropy as unavailable energy, the second law could be paraphrased as: "in an isolated system, available energy can never increase". This is exactly the formulation of the second law via *exergy* as discussed in Sect. 3.6.3. However, this does not mean that the interpretation of entropy as "unavailable energy" is correct. First, physically entropy is not energy; it has the physical dimensions of [energy]/[temperature]. Secondly, although it makes sense to talk about "available" energy or exergy (see Eq. 3.72), it makes no (physical) sense to talk about "unavailable energy". How can unavailable energy be defined and measured? The interpretation of entropy as unavailable energy was firstly given by James Maxwell (1868) and later corrected by him in the fourth edition of his book: "*Theory of Heat*" (London, 1875) (see also [17]).

### *3.10.2 Entropy Interpretation as Disorder*

For a long time, entropy has been interpreted as the degree of disorder in a thermodynamic system. Qualitatively, this has been justified by the fact that entropy refers to changes in the *status quo* of a system and is a measure of "molecular disorder". This interpretation is due to Boltzmann who called the second law of thermodynamics the *law of disorder*, justified by the fact that disordered states produced by local stochastic molecular collisions are the "most probable". In other words, Boltzmann stated that entropy is a measure of the probability of a given macrostate, so that high entropy indicates a high probability state and, correspondingly, low entropy indicates a low probability state. According to Boltzmann's view, "molecules moving at the same speed and in the same direction (ordered behavior) is the most improbable case conceivable, an infinitely improbable configuration of energy [211]". On the basis of this interpretation, many researchers have tried to derive computational formulas of disorder (and order) in atomic or molecular assemblies [212–214].

But, as Daniel Styer has illustrated experimentally in many cases of liquid crystallization, increased order accompanies increased entropy [215]. The physical explanation of this is the following. As cooling of liquid proceeds, the formation of crystals inside it starts to take place. Although these crystals are more ordered than the liquid they are coming from, in order to be able to form, they must release a great amount of heat, called the *latent heat of fusion*. Flowing out of the system, this heat increases the entropy of the surrounding much more than the decrease of entropy occurring in the liquid during the formation of the crystals. Thus, the total entropy of the crystallized liquid and its surrounding increases despite the fact that the disorder of the crystallized liquid is reduced and its order increased.

There are numerous examples in which the second law is obeyed by an isolated system and, at the same time, the system produces order in some part of it, so long as there is a greater increase of disorder in some other part of the system. Examples of this kind are provided in [207] where it is stated that "entropic forces become

significant at scales of, roughly, a few tens of nanometers to a couple of microns". It is through these kinds of forces by which the nature of subatomic particles contributes to the ordering type of molecules that can form, a phenomenon known in biology as the "*macromolecular phenomenon*".

Also, gravity has the effect that the atmosphere remains attached to the Earth in an *orderly sphere,* rather than floating off randomly. This is because, although the gases are relatively ordered around the Earth, a large quantity of energy becomes disordered (due, for example, to the heat generated by the impact of these gas molecules striking each other and the Earth's surface, which radiates off into space). Thus, in sum, the total entropy is increasing and the second law is not violated.

We close our discussion on the disorder interpretation of entropy by mentioning Elias Gyftopoulos question [216]:

> '*Why do so many professionals continue to believe that thermodynamic equilibrium is a state of ultimate disorder despite the fact that both experimental and theoretical evidence indicates that such a state represents ultimate order?*' [215, 217].

### *3.10.3 Entropy Interpretation as Energy Dispersal*

According to this interpretation [218], entropy is a measure of "*energy dispersal or distribution*" at a specific temperature, and the second law is expressed as: "*energy spontaneously disperses from being localized to becoming spread out if it is not hindered from doing so*." Historically, the concept of energy dispersal was used for the first time by Kelvin in 1952 [219] and in the mid 1950s by Kenneth Denbigh by interpreting the changes of entropy as a mixing or spreading of the total energy of each ingredient of a system over the available quantized energy levels [220, 221]. It is noted that many authors, who previously had been presenting the "disorder" representation of entropy, subsequently, in new editions or new books, accept the "energy dispersal" representation. For example, Peter Atkins, who previously stated that entropy is energy dispersal leading to a disordered state [222], now has discarded the disorder representation and writes simply that "spontaneous changes are always accompanied by a dispersal of energy" [223]. Other authors that adopted the energy dispersal interpretation of entropy are Stanley Sandler [214], John Wrigglesworth [224], Gupta [225], Cecie Starr [226], and Andrew Scott [227].

The "energy dispersal" interpretation of entropy has been promoted over the last ten years by Frank Lambert, who has written relevant educational material for both instructors and students, and promotes it on several websites [228]. He includes simple examples in support of the interpretation (rock falling, hot frying pan, cooling down, iron rusting, air shooting out from a punctured high-pressure tire, ice melting in a warm room, etc.), how much energy is dispersed, and how widely spread out are the examples. The "energy dispersal" interpretation of entropy has been strongly questioned in the *Encyclopedia of Thermodynamics,* invoking a considerable number of arguments and excerpts of publications on this topic [229].

### 3.10.4 Entropy Interpretation as Opposite to Potential

This interpretation is based on the fact that nature is intrinsically active and that, for any distribution of energy (potential) which is not in equilibrium, a thermodynamic force (i.e., the gradient of the potential) exists, through which nature acts spontaneously to dissipate or minimize this potential, or, equivalently to maximize entropy. Clausius coined the term "entropy" to characterize the dissipated potential, in terms of which the second law states that "the world acts spontaneously to minimize potential". For example, if a glass of hot liquid is placed in a colder room, a potential occurs and a flow of heat is spontaneously generated from the glass to the room until it is minimized (or the entropy maximized). At this point, the temperatures are the same and the heat flow stops (thermodynamic equilibrium).

This fact has led *Rod Swenson* (1988) in the development of the following "*Law of Maximum Entropy Production*" (**MEP**) [230]: "A system will select the path or assemblage of paths that minimizes the potential or maximizes the entropy at the fastest rate given the constraints". This law is in contrast to Boltzmann's interpretation of the entropy law as a "law of disorder" and his view that "transformations from disorder to order are infinitely improbable". According to Swenson, such transformations characterize planetary evolution as a whole and occur regularly in the universe, predictably and ordinarily with "probability one" [231]. Thus the law of "*maximum entropy production*" is the physical selection principle that explains why the world is in the "*order-production business*". The MEP law complements the second law that says that entropy is maximized, while the law of maximum entropy production says that it is maximized (potentials minimized) at the fastest rate given the constraints. Actually, the MEP law does not contradict or replace the second law, but it is another law in addition to it.

As Swenson and Turvey indicate [232]: "If the nature chooses those dynamics that minimize potentials at the fastest rate (steepest descent) given the constraints (MEP law), and if ordered flow produces entropy faster than disordered flow (as required by the balance equation of the second law), then the world can be expected to produce order whenever it gets the change. Thus, the world behaves in an opportunistic manner for the production of dynamic order, because in this way the potentials are minimized at a higher rate. This means that the world is in the *order-production business* because ordered flow produces entropy faster than disordered flow".

For example, consider a well-sealed house in which heat flows to the outside environment only by conduction via the walls. If a door or window is opened (i.e., a constraint on the heat flow is removed), a new path of heat flow is provided, and the heat flow rate at which potential is minimized is increased. Of course, until the potential is minimized, heat flow will also continue through the walls, i.e., each path will drain all that it can, but the greater larger part of the potential will be reduced through the fastest way (i.e., the window or door). This means that the system will automatically select the assemblage of paths, from among the available ones, such that it goes to the equilibrium (final) state of minimum potential at the fastest rate, given the constraints as required by the MEP law.

Swenson has extensively and deeply studied the order-production business of the world and has shown that the MEP law has notable implications for evolutionary and culture theory, ecology theory, human development, and globalization [159, 233, 234]. Here, it is useful to mention the following law of Hatsopoulos and Keenan [235, 236] and Kestin [237] which subsumes the zeroth, first, and second laws, and was named the "*Law of Stable Equilibrium*" by Hatsopoulos and Keenan, and "*Unified Principle of Thermodynamics*" by Kestin [238]: "When an isolated system performs a process after the removal of a series of internal constraints, it will reach a unique state of equilibrium which is independent of the order in which the constraints are removed".

From this law, we can deduce both the classical laws of thermodynamics for systems near thermodynamic equilibrium and laws for systems far-from-equilibrium, i.e., systems in the nonlinear phase far from thermodynamic equilibrium (e.g., whirlpools, all living systems, and ecosystems) [238].

## 3.11 Maxwell's Demon

"*Maxwell's Demon*" is the name of an imaginary experiment proposed in 1871 by James Clerk Maxwell to contradict the second law of thermodynamics. This experiment (creature) was given the name "*demon*" by Kelvin in his paper: "*Kinetic Theory of the Dissipation of Energy*", Nature, 441–444, 1874, and further discussed in his 1879 lecture "*the Sorting Demon of Maxwell*" (Proc. Royal Institution, Vol. ix, p. 113, Feb. 28, 1879) which stated as follows: "The word '*demon*', which originally in Greek meant a supernatural being, has never been properly used as signifying a real or ideal personification of malignity. Clerk Maxwell's '*demon*' is a creature of imagination having certain perfectly well-defined powers of action, purely mechanical in their character, invented to help us to understand the '*Dissipation of Energy*' in nature". The lecture ended as follows: "The conception of the '*solving demon*' is merely mechanical, and is of great value in purely physical science. It was not invented to help us to deal with questions regarding the influence of life and of mind on the motions of matter, questions essentially beyond the range of mere dynamics".

The original description of the *demon* as given by Maxwell is as follows [17]. "If we conceive of a being whose faculties are so sharpened that he can follow every molecule in its course, such a being, whose attributes are as essentially finite as our own, would be able to do what is impossible to us. For we have seen that molecules in a vessel full of air at uniform temperature are moving with velocities by no means uniform, though the mean velocity of any great number of them, arbitrarily selected, is almost exactly uniform. Now let us suppose that such a vessel is divided into two portions, $A$ and $B$, by a division in which there is a small hole, and that a being, who can see the individual molecules, opens and closes this hole, so as to allow only the swifter molecules to pass from $A$ to $B$, and only the slower molecules to pass from $B$ to $A$. He will thus, without expenditure of work, raise the

temperature of B and lower that or *A*, in contradiction to the second law of thermodynamics".

Some other Maxwell-like demons were proposed by *Eric H. Neilsen* and can be found on the website: http://home.final.gov/~neilsen/publications/demon/node4.html#SE. The simpler of them is the so-called "*pressure demon*", i.e., a valve between two chambers of gas that opens or closes to permit molecules to pass one way but not the other. If the two chambers were originally at equilibrium, the demon working in this way creates a pressure difference that could be used to do work, thus violating the second law. The energy required for doing this work would be provided by the random motion of the molecules in the gas. Clearly, if such a demon could be realized, the thermal energy of our surroundings could be used as a source of energy.

Over the years since Maxwell's time, very many scientists have attempted to show that the "demon" cannot possibly exist because this violates the second law of thermodynamics. Most of these "*demon exorcism*" papers or talks state that the demon needs to collect proper information to "know" the whereabouts (positions and velocities) of the particles in the system, i.e., he must have the capability to *observe* the molecules (e.g., with the help of photons that enter into the system from the outside). This means that to operate, the demon needs external energy, and overall consumes more *negentropy* than created by operating the valve [3]. Other exorcisms are based on the argument that the thermal fluctuations of the gas would not allow any autonomous mechanism to operate successfully as a Maxwell demon. According to them, the Brownian motion of the gas molecules that collide with the door heats up the door and forces it to "oscillate", thus, ceasing to operate as a one-way valve [239].

A few examples of claims and statements about the demon are as follows:

**Carlton Caves**: "A Maxwell demon can, in certain circumstances, do the impossible: transfer energy from a cool body to a warmer one". (His argument is based on information theory) [240].

**I. Walker**: "Here Maxwell's Demon is interpreted in a general way as a *biological observer system* within (possibly closed) systems which can upset thermodynamic probabilities provided that the relative magnitudes between observer system and observed system are appropriate. Maxwell's Demon within Boltzmann's gas model appears only as a special case of inappropriate relative magnitude between the two systems" [239].

**J.** and **D.J. Earman**: "We show how these efforts [to restrict the extent of violations of the second law] mutated into Szilard's (1929) proposal that Maxwell's Demon is exorcised by proper attention to the entropy costs associated with the Demon's memory and information acquisition…. We argue through a simple dilemma that these attempted exorcisms [via information theoretic concepts] are ineffective, whether they follow Szilard in seeking a compensating entropy cost in information acquisition or Landauer in seeking that cost in memory erasure, in so far as the Demon is a thermodynamic system already governed by the Second Law, no further supposition about information and entropy is needed to save the Second Law" [241].

**Elias P. Gyftopoulos**: (1) "We show that Maxwell's demon is unable to accomplish his task not because of considerations related to irreversibility, acquisition of information, and computers and erasure of information, but because of limitations imposed by the properties of the system on which he is asked to perform his demonic manipulations" [242]; (2) "Based on the principles of a unified quantum theory of mechanics and thermodynamics, we prove that Maxwell's demon is unable to accomplish his task of sorting air molecules into swift and slow because in air in a thermodynamic equilibrium state there are no such molecules" [243].

**Elias P. Gyftopoulos** and **Gian Paolo Beretta**: "Several presentations address the question of the feasibility of Maxwell's demon. We are not going to comment on each presentation separately because all claim to exorcise the demon but in our view none confronts the problem by Maxwell. …Of the myriads of publications on the subject [Maxwell's demon], only two address and resolve the problem specified by Maxwell. One purely thermodynamic (see (1) above [242])…. The second… is quantum thermodynamic" (see (2) above [243]) without statistical probabilities [244]. *Note*: The above two approaches (purely thermodynamic and purely quantum thermodynamic) are the approaches discussed in Sect. 3.5.4, which are based on the system properties and state definitions given in Sect. 3.2.9. Gyftopoulos, Hatsopoulos and Beretta's non-statistical general physics approach to thermodynamics and entropy concept has been criticized by Čapek and Sheedan in [245], but Gyftopoulos and Beretta have indicated in [244] that none of the statements made in [245] are valid.

## 3.12 The Arrow of Time

It is well known that none of the fundamental equations (classical physics, quantum physics, relativistic physics) contain an arrow of time. All equations can run equally well forwards and backwards. This reversibility in time is a scientific fact that is based on a deep philosophical truth, namely, that unidirectional time is not fundamental in physics, but emerges from a more basic substrate of bidirectional time. The term "*arrow of time*" or "*time's arrow*" was introduced in the book "The Nature of the Physical World (1928)" by the English astronomer *Arthur Eddington* (called the father of time's arrow) as an independent direction in his four-dimensional world model (organizations of atoms, molecules, bodies, and time).

At the macroscopic level in everyday life, we see that there is a unidirectional flow of time from past to present to future. Whatever physical property or phenomenon possesses such time asymmetry is an arrow of time [246, 247]. For the mathematical physicists, the inconsistency between the *time symmetry* of the physical equations and the *time asymmetry* of real-life seems to be a "*paradox*" or an *illusion,* or even a "*mystery*".

Figure 3.7 illustrates the arrow of time in terms of the cosmological evolution in the universe, and chemical, biological, and cultural evolution on Earth.

*Albert Einstein* quoted: "For those of us who believe in physics, this separation between past, present, and future is only an illusion, although a persistent one".

*Ben Goertze said:* "Time is a mystery, perhaps the ultimate mystery. It is deeply wrapped up with the mind/body problem and the mystery of consciousness. Yet only very infrequently do we scientists take time out to reflect upon the nature of time" [248].

The physical cosmologist *John Wheeler* stated: "No phenomenon is a physical phenomenon until it is an observed phenomenon". (The Demon and the Quantum, R.J. Scully, 2007). He further stated: "Time is defined so that motion looks simple" (Gravitation, 1973).

Over the years, in physics, cosmology, and human life sciences, several *arrows of time* have been introduced, based on various one-way evolving (unidirectional) physical properties or phenomena. These are the following:

- Psychological arrow of time
- Thermodynamic arrow of time
- Cosmological arrow of time
- Quantum arrow of time
- Electromagnetic arrow of time
- Causal arrow of time
- Helical arrow of time

A brief review of them follows.

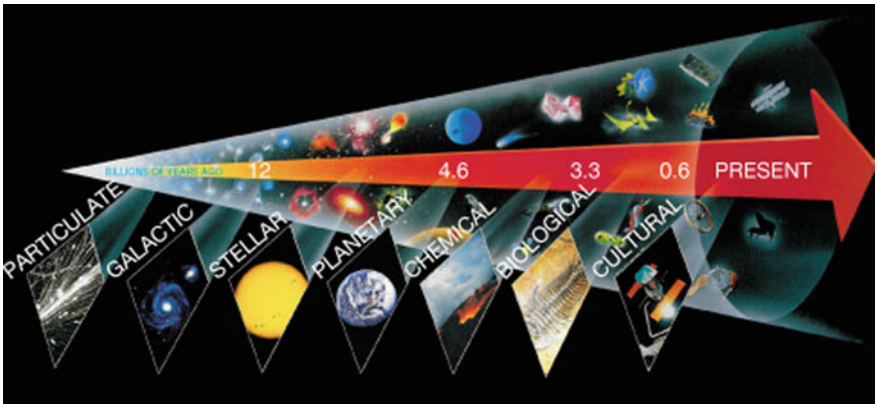



**Fig. 3.7** Illustration of the arrow of time. (http://astronomy.nju.edu.cn/~lixd/GA/AT4/AT428/IMAGES/AACHDLP0.JPG)

### 3.12.1 Psychological Arrow

This is our every day intuitive sense of time, i.e., the passage of time is an objective feature of reality. We are recording the continuously increasingly items of memory that we perceive. The present is always advancing into the future. People are born, age, and die. Old buildings are replaced by completely different new ones. Forests are burned to ash, etc. The present is always advancing into the future. The present is *what is real*. What we remember forms the past, whereas the events that are not yet stored in our memory and so we do nt remember them make the future. The "present" is a subjective concept, and the "*now*" depends on our point of view in exactly the same manner as that of "*here*" does [249] (Fig. 3.8).

The Scottish philosopher *Donald Mackay*, in his book "*Impossibility*", states that, if a human had access to information about her/his future actions and behavior, it would be impossible to predict that future. This is because information about future performance is exactly what a person will gain if this person is able to remember the future, and so the person could change the path of action with respect to how her/his memory of the future tells her/him to act. This appears to be a logical inconsistency, namely: If a person has the capability to remember the future, then those memories of the future instantly become unreliable. Thus, it appears to be "impossible to remember the future".

### 3.12.2 Thermodynamic Arrow

This arrow, also called the *entropy arrow*, is based on the second law of thermodynamics according to which entropy in an isolated system tends to increase over time, and on the irreversibility of the flow of heat from hot to cold. A problem that arises here is the above mentioned apparent paradox. The fundamental laws of the universe as we understand them are time symmetric, i.e., they do not possess this unidirectionality. What is then the relationship between the reversible behavior of atoms and molecules that make the objects and the irreversible behavior of the (macroscopic) objects that we see and touch? The explanation is the fact that here

**Fig. 3.8** Psychological arrow of time: the memories of the past (the future cannot be remembered)

we have phenomena on different scales that are hierarchically occurring and that these phenomena must be studied (to some extend) independently (e.g., to study ocean waves, we do not need to deal with atoms). The answer lies in Boltzmann's interpretation of entropy according to which, as time increases, a system becomes more disordered. For example, a cup of milk is a macroscopic system made up by a large number of microscopic particles that are ordered in one of several disordered states. If we bump this system (the cup of milk) out of its equilibrium, the milk is dashed to the floor, and it would be very difficult (practically impossible) to bring it back to the exact order it has previously before dashing.

According to Boltzmann, the probability of recreating that one orderly state is extremely low. This shows exactly the existence of the *arrow of time* from the past to the present to the future. Of course, here we have the complication that living organisms decrease locally their entropy by using energy at the expense of an entropy increase in their environment. And, as Lebowitz indicates, this arrow of time "Explains almost everything. What it does not explain is why we are here, when—if you look at all possible microscopic states—this is such an unlikely state?"

The cosmologist David Layzer has argued that in an expanding universe the entropy increases as required by the second law of thermodynamics, but the maximum possible entropy of the universe might increase faster than the actual increase of entropy. This implies that there may be some room for an increase of information (order) at the same time the entropy is increasing. He pointed out that if the equilibrium rate of the matter, i.e., the speed with which the matter redistributes itself randomly among all the possible states, was slower than the rate of expansion, then the "negative entropy" or "order" (defined as the difference between the maximum possible entropy and the actual entropy) would also increase. This negative entropy was interpreted by Claude Shannon as information (see Chap. 4), although visible structural information in the universe may be less than this "potential" for information [250] (Fig. 3.9).

### *3.12.3 Cosmological Arrow*

In cosmology, which is an empirical branch of science devoted to the study of the large-scale properties of the universe, the arrow of time, called the "*cosmological arrow of time*", is the direction of the expansion of the universe. This means that the "*radius*" of the universe is used to define the arrow of time, although, according to *Mario Castagnino, Olimpia Lombardi,* and *Luis Lara:* "There is no reason for privileging the universe radius for defining the arrow of time over other geometrical properties of the space-time. …Geometrical properties of space-time provide a more fundamental and less controversial way of defining an arrow of time for the universe as a whole…, if certain conditions are satisfied. The standard models of contemporary cosmology satisfy these conditions" [The Arrow of Time in Cosmology (http://philsci-archive.pitt.edu/archive/00000800)].

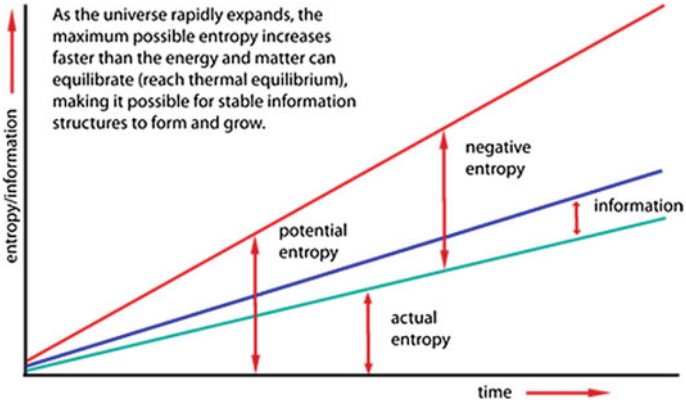



**Fig. 3.9** Graph of Layzer's explanation of negentropy/information (historical arrow) (http://www.informationphilosopher.com/problems/arrow_of_time/)

The "*radius-based arrow of time*" has been related to the thermodynamic arrow, the relation being considered to be a consequence of the *initial conditions* in the early universe. On this issue, *Joel Lebowitz* of Rutgers University says [249]: "The universe began in a state of very low entropy, a very ordered state; there was a uniform distribution of energy. It was no dumpy".

The cosmological arrow of time is illustrated by a schematic in Fig. 3.10.

*Peter Landsberg* states: "It has been suggested that thermodynamic irreversibility is due to cosmological expansion. …It seems an odd procedure to attempt to explain everyday occurrences, such as diffusion of milk into coffee, by means of theories of universe which are themselves less firmly established than the phenomena to be explained. Most people believe in explaining one set of things in terms of others about which they are more certain, and the explanation of normal irreversible phenomena in terms of the cosmological expansion is not in this category" (The Study of Time III, 117–118, 1973).

*Monsignor Georges Lemaitre* wrote: "The (universe) expansion thus took place in three phases: a first period of rapid expansion in which the atom-universe was broken into atomic stars, a period of slowing-down, followed by a third period of accelerated expansion. It is doubtless in this third period that we find ourselves today, and the acceleration of space which followed the period of slow expansion could well be responsible for the separation of stars into extra-galactic nebulae" (La Formation des Nebuleuses dans L'universe en Expansion, Comptes Rendus, 196, 903–904, 1933). Continuing on the cosmological arrow, *Joel Lebowitz* wrote: "Unlike the case with regular matter, where being disorganized, spread out, is a state of higher entropy with gravitation the state of increasing entropy is actually the state of clumping. That's why matter collects in planets, planets collect into solar systems, stars collected into galaxies, and galaxies collect into supergalaxies. As the universe has spread out, it has become highly irregular".

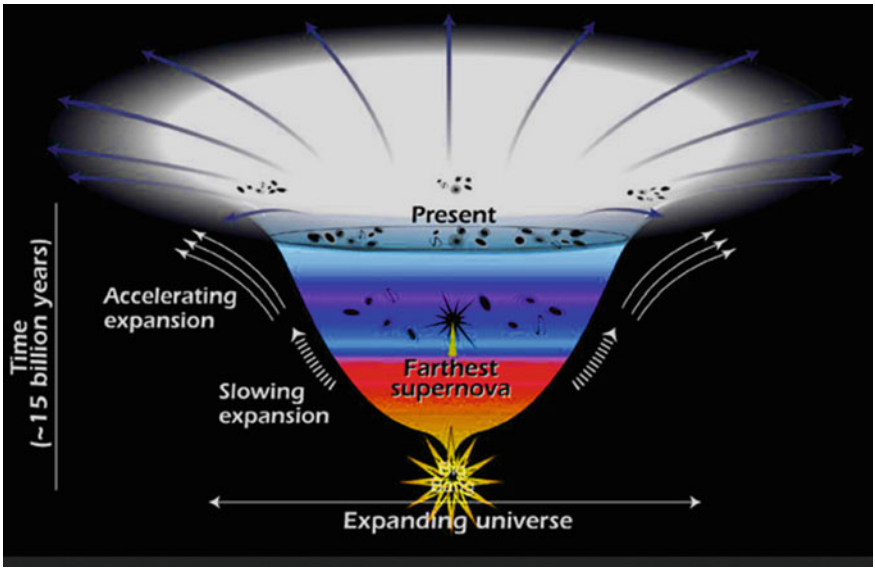



**Fig. 3.10** Pictorial representation of the cosmological arrow of time (direction of the universe's expansion) (https://patternizer.files.wordpress.com/2010/10/arrowoftime.jpg?w=620)

In other words, one can say that our place in the universe's evolution is a result of the cosmological arrow's reversal as gravity pulls everything back into a *Big Crunch*.

*Andreas Albrecht* wrote: "Still the ultimate origin of the thermodynamic arrow of time in the Standard Big Bang (SBB) cosmology lies in the very special highly homogeneous initial state, which is far removed from the collapsed states to which gravitating systems are attached. But now we have inflation which seems to tell the opposite story: Inflation tells us that the initial conditions of SBB, rather than being unusual, are the generic result of evolution toward an attractor… a kind of equilibrium. Can we really have it both ways? A careful analysis of the arrow of time gives a very interesting perspective on current issues in cosmology. This perspective helps us evaluate what role inflation, holography, or other ideas might ultimately have in explaining the special initial of the Big Bang, and how ideas such as inflation might fit into a global picture that includes the time before inflation and the full range of possible cosmic histories, as well as the cosmic acceleration observed today" (The Arrow of Time, Entropy and the Origin of the Universe, Science and Ultimate Reality Symp., In honor of John Archibald Wheeler, Princeton, N.J., USA, March 15–18, 2002; www.metanexus.net/archives.wheeler.html).

The thermodynamic arrow points toward the so-called *heat death* (*Big Chill*) as the amount of usable energy (exergy) tends to zero. Three quotes on heat death follow.

*William Thomson* (*Kelvin*): "Within a finite period of past, the earth must have been, and within a finite period of time to come, the earth must again be unfit for the habitation of man as at present constituted, unless operations have been, or are to be performed, which are impossible under the laws to which the known operations going on at present in the material world are subject." (On a Universal Tendency in Nature to the Dissipation of Mechanical Energy, Proceedings Royal Society Edinburgh, April, 1852).

*Oswald Spengler:* "Entropy signifies today the world's end as a completion of an inwardly necessary relation" (C. Smith and N. Wise, Energy and Empire).

*Charley Lineweaver* (ANU): "There is no doubt in any astrophysicist's mind that the sun will run out of fuel and expand into a red giant in about five billion years…. The universe appeared to be headed towards a state of 'heat death' when all of its stars, planets and other matter would reach exactly the same temperature. The question is, when it will end? And all you can say is we are closer to the heat death than we anticipated. If there is going to be a heat death, we're talking many, many times longer than the lifetime of the sun … it's a much longer time scale" (A Larger Estimate of the Entropy of the Universe, Astrophysical J., Quoted in the Daily Telegraph, 25 Jan 2010).

Figure 3.11 summarizes the definitions of the thermodynamic, psychological, and cosmological arrows.

### *3.12.4 Quantum Arrow*

Things become stranger when we move from classical thermodynamics to the quantum- physics world. Here the quantum evolution is governed by the time-symmetric equation of Schrödinger and by the wave function collapse which is asymmetric. The wave-function-collapse mechanism is very complicated and it is unclear what is the relationship between this arrow and the others. The interpretation of quantum physics becomes much simpler by assuming that time is bidirectional.

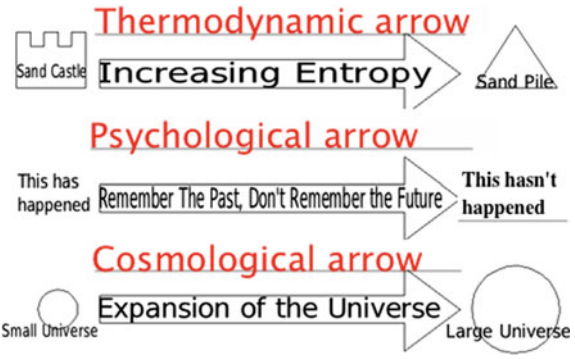



**Fig. 3.11** Interpretation of the three basic arrows of time. (http://1.bp.blogspot.com/-SZRoNS5Y8Pg/TwwOpwc6obI/AAAAAAAAABc/gQF3d0HraaQ/s400/arrows+of+time.png)

According to "*The Quantum Casino*", when one performs a measurement of a quantum observable, there is a "collapse of the wave function" in which a probability wave collapses to produce only a single observed value from a repertory of possible values. This process seems to be irreversible, i.e., to occur only in the forward direction. This is explained by the destruction of the coherent phase relationships of the interference terms when a particle interacts with the environment. The dissipation of these terms into the wider environment can be interpreted in terms of increasing entropy. Thus, quantum decoherence can be regarded and understood as a thermodynamic process. After decoherence, the process goes to thermodynamic irreversibility (www.ipod.org.uk/reality/reality_arrow_of_time.asp).

According to Gramer [251], every particle in the universe sends out waves in both time directions. Observed events appear when a forward wave collides with a suitable backward wave. This may be regarded as peculiar, but it is indeed fully consistent with all of our knowledge about the physical universe. All paradoxes of quantum reality disappear when one drops the assumption that events occur at particular points in time.

### *3.12.5 Electromagnetic Arrow*

Here, time is measured according to the movement of radiation (*radiative arrow*). All waves expand outward from their source, although the solutions of the wave equations may give both convergent waves and radiative ones. The electromagnetic arrow has been reversed in some special experiments in which convergent waves were created. Actually, a radiative wave increases entropy, whereas a convergent wave decreases it, which under normal conditions contradicts the second law.

A noticeable theory of symmetric-in-time electrodynamics was developed in the 1940s by *Richard Feynman* and *John Wheeler*. The arrow of time appears in electromagnetic dynamics via the dominance of *retarded waves* over *advanced waves*. For example, a TV or radio transmitter sends waves outwards (i.e., into the future) from the antenna. According to the Feynman–Wheeler theory, an electromagnetic transmitter sends half of its signals into the future and half into past.

### *3.12.6 The Causal Arrow*

It is very difficult to make a clear statement about the meaning of "*cause and effect*". In epistemology, the use of "*causality*" as an arrow of time cannot easily be perceived; we can only perceive sequences of events. For example, it does seem evident that dropping a cup of milk is the cause, while the milk dashing to the floor is the effect. Clearly, this shows the relationship of the causal arrow and the thermodynamics arrow. The causal arrow of time is a special case of the

thermodynamic arrow of time. Actually, we can cause something to happen only in forward time, not backward.

*Michael Lockwood* writing about causality in his book "The Labyrinth of Time" says: "We regard the forward direction of time, in stark contrast to the backward direction, as the direction in which *causality* is permitted to operate. Cause, we assume, can precede their effects, but cannot follow them".

If we assume causality to be bidirectional (time symmetrical), then we may consider that our present states are being caused by time-reversed future events as much as by past events. Perhaps we do not actually see causality to happen backward in time because of a bias in our psychological systems. Some complexity mechanism in our brains makes our thought processes perform only in forward time.

### 3.12.7 The Helical Arrow

Closing our discussion of the time's arrow, it is useful to mention the work of *Ben Goertzel* on a new concept of time's arrow, namely the "*helical time's arrow*", which is not linear like all other arrows [248]. He develops the helical arrow of time from the perspective of both physics and phenomenology.

As *Ben Goertzel* observes: "There are good reasons to take a view of time as something *helical* rather than linear … Ultimately, I believe, all of dynamical systems theory—and for that matter, all of the science, may have to be rethought in terms of helical time". He defines the helical model of time as follows:

"Helical time is time which moves around and around in circles, occasionally, however, moving somewhere that it has not gone before, going off in a new direction. It is backwards-forwards-moving time, without the backward and forwards necessarily being in equal balance (Fig. 3.12). The concept of helical time gives us a vision of a time that is moving linearly in one direction, but only probabilistically. … The word '*helical*' is used metaphorically. I am not talking about a precise mathematical helix structure". He also states: "The helix of time does not grow around the emergent axis of time like a vine around a pole. Rather, the emergent axis of time is a kind of coarse-grained, averaged view of the motion of the helix of time. The linear flow of time is obtained from the underlying helical flow by the imposition of subjective ignorance. The linear flow of 'time's arrow' is only a statistical approximation to the true nature of physical reality".

**Fig. 3.12** The helical arrow: a sketch of the forward and backward motion of time. For illustrative simplicity, space is collapsed onto a single vertical axis [248] (http://www.goertzel.org/papers/timepap.html on the physics and phenomenology of time)

## 3.13 Conclusions and Quotes for Thermodynamics, Entropy, and Life

In this chapter, we have guided the reader to the principal and basic concepts and branches of thermodynamics, with emphasis on the laws of thermodynamics and the major concept of entropy. To summarize, the four fundamental laws are the following:

**0th Law**: Law of thermal equilibrium.
**1st Law** : Conservation of energy law.
**2nd Law** : Law of non-reducing entropy (non-increasing exergy).
**3rd Law** : The law of entropy at absolute zero.

Regarding the fourth Law, we have discussed *Lotka's* maximum-energy-flux principle, and *Onsager's* reciprocal relations principle which have received much attention in several fields and are frequently mentioned as the fourth law of thermodynamics.

All these laws are very important and have influenced considerably thinking on the evolution about life by scientists and philosophers. But the law which has generated the widest and strongest discussions about the universe and the life on Earth is the second law, the law of entropy. Thousands of scientific papers, hundreds of books, and many websites treat these issues from different, sometimes very contradictory, points of view. A full discussion or criticism (constructive or speculated) of them is beyond the scope of the present book. Therefore, we present here a small, but representative, set of conclusions, opinions or quotations expressed by older and newer experts and thinkers in the field that, we feel, will give a "good" impression of the developments made in the field and the insights gained since the "caloric and motive power of fire" times.Some further important references on thermodynamics are [252–260].

### *3.13.1 Thermodynamics General Quotes*

**Sadi Carnot**: "The production of motion in the steam engine always occurs in circumstances which it is necessary to recognize, namely when the equilibrium of caloric is restored, or (to express this differently) when caloric passes from the body at one temperature to another body at a lower temperature" [Reflexions sur la Puissance Motrice du Feu (1824), Translation by *Robert Fex*: Reflections on the Motive Power of Fire (1986)].

**Rudolf Clausius**: "Heat can never pass from a colder to a warmer body without some other change, connected therewith, occurring at the same time" [Modified Form of the second Fundamental Theorem, Philosophical Magazine, 12, 86 (1856)].

**William Thomson**: "There is at present in the material world a universal tendency to the dissipation of mechanical energy. Any restoration of mechanical energy, without more than an equivalent of dissipation is impossible in inanimate material processes, and is probably never effected by means of organized matter,

either endowed with vegetable life or subject to the will of an animated creature". (On a Universal Tendency in Nature to the Dissipation of Mechanical Energy, Proc. Royal Soc. of Edinburgh, (1852), Mathematical & Physical Papers, vol. 1, p. 511).

**James Clerk Maxwell**: "We define thermodynamics … as the investigation of the dynamical and thermal properties of bodies, deduced entirely from the first and second law of thermodynamics, without speculation as the molecular constitution" (The Scientific Papers of James Clerk Maxwell, 664–665, 2003).

**Willard Gibbs**: "Any method involving the notion of entropy, the very existence of which depends on the second law of thermodynamics, will doubtless seem to many far-fetched, and may repel beginners as obscure and difficult of comprehension" [Graphical Methods in the Thermodynamics of Fluids (1873)]. "If we say, in the words of Maxwell, some years ago (1878), that thermodynamics is a science with secure foundations, clear definitions, and distinct boundaries, and ask when these foundations were laid, those definitions fixed, and those boundaries traced, there can be but one answer: certainly not before the publication of Clausius memoir (1850)".

**Ludwig Eduard Boltzmann**: "Since a given system can never of its own accord go over into another equally probable state but into a more probable one, it is likewise impossible to construct a system of bodies that after traversing various states returns periodically to its original state, that is a perpetual motion machine" (Address to a meeting of the Imperial Academy of Science, 29 May 1886).

It is true that energy is the most basic pillar of life and human life and that the laws of thermodynamics so far known have explained many issues of human life, especially the physiological and materialistic issues. However, despite the fact that many scientists working in the field have claimed that thermodynamics also explains human higher level mental and behavior issues, their arguments are not fully convincing and still leave many questions unanswered. Perhaps new laws of nature and the physical world to be discovered in the future will explain more issues of human life and society and will improve our understanding of human life. Some basic aspects of this topic will be discussed in Chap. 10.

**Frank A. Greco**: An answer can readily be given to the question, "Has the second law of thermodynamics been circumvented?" "Not Yet" (American Laboratory, Vol. 14, 80–88, 1982).

**Seth Lloyd**: "Nothing in life is certain except death, taxes and the second law of thermodynamics" (Nature 430, 971, 26 August 2004).

**Harold Morowitz**: "The use of thermodynamics in biology has a long history rich in confusion" (Beginnings of Cellular Life: Metabolism Recapitulates Biogenesis, Yale University Press, Yale, 1992).

**G.P. Gladyshev**: "Hierarchical thermodynamics is a necessary key theory for all branches of science" (Hierarchical Thermodynamics—General Theory of Existence and Living World Development).

### 3.13.2 *Entropy Quotes*

**James Lovelock**: "Few physical concepts have caused as much confusion and misunderstanding as has that of entropy" [240].

**Greg Hill** and **Kerry Thornley**: "The tendency for entropy to increase in isolated systems is expressed in the second law of thermodynamics—perhaps the most pessimistic and amoral formulation in all human thought" (Principia Discordia, 1965; Wikipedia).

**J. Barbour**: "It is often said that (the approach to equilibrium) … is accompanied by an increase in entropy, and (is) a consequence of the second law. But this idea actually lacks a theoretical foundation. This aspect of time asymmetry, is woven much deeper in the theory" [239].

**Freeman Dyson**: "The total disorder in the Universe as measured by the quantity that physicists call entropy, increases steadily over time. The total order in the universe, as measured by the complexity and permanence of organized structures, increases steadily over time too." (Distinguished lecture, University of Maryland, March 2 1998).

**Eugen Wigner**: "Entropy is an anthropomorphic concept".

**E.F. Schumacher**: "If entropy is an intrinsic property of matter it should be expressible in terms of physical quantities. Instead, we find that quantum statistical mechanical rendering of entropy is characterized by the existence of logical terms that reminds us the presence of a human mind that is describing or modeling the system"… Born of the unnatural union of wish and reality, entropy is objective enough to be useful in dealing with the physical world, but subjective enough that a purely physical interpretation is not possible". (Small is Beautiful. Harper and Row, New York, 1973).

**Elias Gyftopoulos**: "Entropy (is) an inherent, non-statistical property of any system in any state" (thermodynamic equilibrium or not) [216].

**Frank Lambert**: "In the Gibbs equation, $\Delta G = \Delta H - T\Delta S$, each term describes an aspect of the energy that is dispersed because of a chemical reaction occurring in a system …. If each term is divided by $T$, the Gibbs equation can be seen as an '*all entropy*' *equation*. The Gibbs 'works' because it corresponds to the second law as we would state it: Energy spontaneously disperses, if it is not hindered. When it does so, entropy increases in the combination of system plus surroundings". (http://entropysite.oxy.edu/students_approach.html).

**Tim Thomson**: "Entropy is what the equations define it to be. Any prose explanation should be compatible with the equations. … These generalized forms of entropy (viz., Tsallis and Renyi entropies), of relatively recent origin, serve to show that *entropy* is not just an old friend that we know quite well, as in classical thermodynamics, but also a concept that is rich in new ideas and scientific directions" (www.tim-thompson.com/entropy1.html).

**Henri Salles**: "Heat and its transfer, motion and its transfer all involve at least two distinct phenomena that are cold and hot, slow and fast. They are compounded phenomena. A compound phenomenon is always an effect and as such cannot be a

fundamental or basic phenomenon. The entropy concept, based on heat transfer, should not, by way of consequences be thought as a fundamental tenet but as only the effect of a cause" (The Harmony of Reality, Gravimotion Approach, http://what-is-entropy.info).

### 3.13.3 *Life and Human Thermodynamics Quotes*

The branch of thermodynamics that has received the greatest attention is, as expected, life and human thermodynamics. A set of opinions on this branch that cover broadly diverse views follows.

**James Lovelock**: "Entropy is a shadow kind of concept difficult to grasp, but we may point out that the reader, who would extend the notion of mechanism into life, must grasp it" [250].

**Eric Schneider** and **James Kay**: "The thermodynamic principle which governs the behavior of systems is that, as they are moved away from equilibrium they will utilize all avenues available to counter the applied gradients. As the applied gradients increase, so does the system's ability to oppose further movement from equilibrium" (*Restated Second Law*, derived from Kestin Unified Principle of Thermodynamics). "Life exists on Earth as another means of dissipating the solar induced gradient and as such is a manifestation of the restated second law" [238].

**Albert Lehninger**: "Living organisms preserve their internal order by taking from their surroundings free energy, in the form of nutrients and sunlight, and returning to their surroundings an equal amount of energy as heat and entropy" (Principles of Biochemistry, Worth Publishers, 1993).

**Ilya Prigogine** and **Isabelle Stempers**: "We grow in direct proportion to the amount of *chaos* we can sustain and dissipate" [193].

**Thomas Fararo**: "Of course, the empirical social systems we treat in sociology are not immune to the laws of thermodynamics; each organism, for instance, is in one aspect an open physical system to which a thermodynamic characterization applies" [195].

**William Blake**: "Energy is the only life and is from the Body, and Reason is the bound or outward circumference of Energy. Energy is Eternal Delight" (M. Klonsky, William Blake: The Seer and His Visions, Orbis, 1977).

**George Gladyshev**: "Life is the phenomenon of existence of the energy dependent dynamic hierarchic structures, mandated by hierarchical thermodynamics" [261].

**Stuart Kauffman**: "All living things are highly ordered systems: they have intricate structures that are maintained and even duplicated through a precise ballet of chemical and behavioral activities …. We may have begun to understand evolution as the marriage of selection and self-organization. … Chaos, fascinating as it is, is only part of the behavior of complex systems. There is also a counterintuitive phenomenon that might be called *antichaos*: some very disordered systems spontaneously '*crystallize*' into high degree of order. Antichaos plays an important part

in biological development and evolution" (Antichaos and Adaptation, Scientific American, August 1991).

**A.E. Wilder-Smith**: "What is the difference then between a stick, which is dead, and an orchid which is alive? The difference is that the orchid has *teleonomy* in it (i.e., information stored in its genes involving the concepts of design and purpose). It is a machine which is capturing energy to increase order. Where you have life, you have teleonomy, and then the Sun's energy can be taken and make things grow-increasing its order".

**A.E. Wilder-Smith**: "The pure chemistry of a cell is not enough to explain the working of a cell, although the workings are chemical. The chemical workings of a cell are controlled by information which does not reside in the atoms and molecules".

**William Thomson (Kelvin)**: "Dead matter cannot become living without coming under the influence of matter previously alive. This seems to me as sure a teaching of science as the law of gravitation. … Our code of biological law is an expression of our ignorance as well as of our knowledge. … and I am ready to adopt, as an article of scientific faith, true through all space and through all time, that life proceeds from life, and from nothing but life" (On the Origin of Life, Popular Lectures and Addresses, Address to the BAAS, Edinburgh, August, 1871).

**M. Waldrop**: "A hurricane is a self-organizing system powered by the steady stream of energy coming in from the sun, which drives the winds and draws rainwater from the oceans. A living cell—although much too complicated to analyze mathematically—is a self-organizing system that survives by taking in energy in the form of food and excreting energy in the form of heat and waste".

**C.G. Darwin**: "Through determining some kind of laws of human thermodynamics, we shall be more successful in doing good in the world ….; of course they cannot be expected to have the hard outline of the laws of physical science, but still I think some of them can be given a fairly definite form".

**Karlis Ullis**: "Human thermodynamics (is) the science concerned with the relations between heat and work in human beings (regarded as) physiological engines".

**Libb Thims**: "Human thermodynamics (is) the chemical thermodynamic study of human molecular reaction life" (www.eoht.info/page/Human+molecule).

**Mihály Csikszenmihályi**: "Emotions refer to internal states of consciousness. Negative emotions like sadness, fear, anxiety, or boredom produce *psychic entropy* in the mind. … Positive emotions like happiness, strength, or alertness are states of *psychic negentropy*. … They focus *psychic energy,* establish priorities, and this creates order in the consciousness (Flow: The Psychology of Optimal Experience, Harper Perrenial, 1990).

**Jing Chen**: "All human activities, including mental activities, are governed by physical laws and are essentially thermodynamic processes".

**Jing Chen**: "From poem writing to money making, the pursuit of low entropy is the man drive of human behavior" (Natural Law and Universal Human Behavior, http://ssrn.com/abstract=303500).

**R. Sterner** and **J. Elser**: "The stoichiometric approach considers whole organisms as if they were single abstract molecules" [262] (Source: www.eoht.info/page/Human+molecule).

**C. Wieland**: "To all rational readers, the use of the chemical theory is nonsense and childish fooling around" [263] (Source: www.eoht.info/page/Human+molecule).

**Steve Fuller**: "I am not a molecule" [264].

## References


1. W.T. Kelvin, An account of Carnot's theory of the motive power of heat. Trans. Edimburgh R. Soc. **XVI** (Jan. 2) (1849)
2. D.S.L. Cardwell, *From Watt to Clausius: The Rise of Thermodynamics in the Early Industrial Age* (Heinemann, London, 1971)
3. P. Perrot, *A to Z of Thermodynamics* (Oxford University Press, Oxford, 1998)
4. P. Atkins, *Four Laws that Drive the Universe* (Oxford University Press, Oxford, 2007)
5. Y. Cengel, M. Boles, *Thermodynamics: An Engineering Approach* (Mc Graw Hill, New York, 2002)
6. J. Dunning-Davies, *Concise Thermodynamics: Principles and Applications* (Horwood Publishing, Chichester, 1997)
7. H.A. Buchdahl, *The Concepts of Classical Thermodynamics* (Cambridge University Press, London, 1966)
8. R. Giles, *Mathematical Foundations of Thermodynamics* (Pergamon, Oxford, 1964)
9. W.M. Haddad, V.S. Chellaboina, S.G. Nerserov, *Thermodynamics* (Princeton University Press, Princeton, 2005)
10. E.P. Gyfropoulos, G.P. Beretta, *Thermodynamics: Foundations and Applications* (Dover Publications, Mineola, 2005)
11. R. Clausius, *The Mechanical Theory of Heat with its Applications to Steam Engines and to Physical Properties of Bodies* (John Van Voorst, London, 1865)
12. C. Caratheodory, Investigation into the foundations of thermodynamics (English Translation from Math. Ann., Vol. 67, pp. 355–386, 1908), in *The Second Law of Thermodynamics* ed. by J. Kestin (Hutchinson and Ross, Dowden, 1976), pp. 229–256
13. G.N. Lewis, M. Randall, *Thermodynamics* (revised by K.S. Pitzer and L. Brewer) (Mc Graw Hill, New York, 1961)
14. E. Fermi, *Thermodynamics* (Dover Publications, New York, 1956)
15. E. Mendoza, *Reflections on the Motive Power of Fire—and Other Papers on the Second Law of Thermodynamics by E* (Clapeyron and R Clausius (Dover Publications), Mineola, 1998)
16. L. Boltzmann *Lectures on Gas Theory* (University of California Press, Berkeley, (1896), 1964) (English Translation from Vorlesungen über Gas theorie, Leipzig 1895/98 UB: 05262 by S.G. Brush (1964), Republished by Dover Publications, New York, 1995)
17. J.C. Maxwell, Theory of Heat, (Longmans, Green and Co., New York, (1888), Reprinted by Dover Publications, New York, 2001)
18. Boltzmann Equation http://scienceworld.wolfram.com/physics/BoltzmannEquation.html
19. J. Von Neumann, *Mathematical Foundations of Quantum Mechanics (Mathematische Grundlagen der Quantenmechanik)* (Springer, Berlin, 1955)
20. Encyclopedia of Human Thermodynamics (a) http://www.eoht.info/page.system (b) http://www.eoht.info/page/boundary (c) http://www.eoht.info/page/Universe (d) http://www.eoht.info/page/Branches+of+thermodynamics

21. About.com physics, http://physics.about.com/od/thermodynamics/f/thermoprocess.htm http://physics.about.com/od/glossary/g/isothermal.htm http://physics.about.com/od/glossary/g/isochoric.htm
22. Adiabatic Processes, http://hyperphysics.phy-astr.gsu.edu/Hbase/thermo/adiab.html
23. Entropy. http://srikant.org/core/node9.html
24. Some Tools of Thermodynamics, http://www.chem.arizona.edu/~salzmanr/480a/480ants/ageq&max/ageq&max.html
25. The State of a System, http://www.chem.arizona.edu/~salzmanr/480a/480ants/chemther.html
26. S.R. Berry, *Understanding Energy, Entropy, and Thermodynamics for Everyman* (World Scientific, Singapore, 1991)
27. Calorimetry, http://www.chem.ufl.edu/~itl/2045/lectures/lec_9.html http://www.science.uwaterloo.ca/~cchieh/cact/c120/calorimetry.html http://www.answers.com/topic/calorimetry
28. G.N. Hatsopoulos, E.P. Gyftopoulos, A unified quantum theory of mechanics and thermodynamics, Part I, postulates. Found. Physics **6**(1), 15–31, 1976; Part II A available energy, ibid, Vol. 6(2), 127–141, 1976; Part IIB Stable equilibrium states, ibid, Vol. 6(4), pp. 439–455, 1976; Part III, Irreducible quantal dispersions, ibid, **6**(5), 561–570 (1976)
29. E.P. Gyftopoulos, Thermodynamic definition and quantum—theoretic pictorial illustration of entropy. ASME Trans. J. Energy Resour. Tech. **120**, 154–160 (1998)
30. E.P. Gyftopoulos, E. Cubucku, Entropy: thermodynamic definition and quantum expression. *Phys., Rev. E.* **55**(4), 3851–3858 (1995)
31. G.P. Berreta, E.P. Gyftopoulos, J.L. Park, G.N. Hatsopoulos, Quantum thermodynamics: a new equation of motion for a single constituent of matter. Nuovo Cimento **82**(B), 69–191 (1984)
32. G.P. Berreta, E.P. Gyftopoulos, J.L. Park, Quantum thermodynamics: a new equation of motion for a general quantum system. Nuovo Cimento **87**(B), 77–97 (1985)
33. G.P. Berreta, Quantum thermodynamics of non-equilibrium: Onsager reciprocity and dispersion-dissipation relations. Found. Phys. **17**, 365–381 (1987)
34. Institute of Human Thermodynamics, http://www.humanthermodynamics.com/HT-polls.html#anchor_59
35. Entropy in the 20th Century Chemistry Texts, http://www.eoht.info/thread/3300562/Entropy+in+20th+Century+chemistry+Texts
36. Physical meaning of entropy, http://sci.tech-archive.net/Archive/sci.physics.research/2008-11/msg00007.html
37. K. Laider, *The World of Physical Chemistry* (Oxford University Press, Oxford, 1993)
38. J. Black, *Encyclopedia Britannica* (by R.G.W Anderson)
39. I. Muller, *A History of Thermodynamics-The Doctrine of Energy and Entropy* (Springer, New York, 2007)
40. Combined Law of Thermodynamics, http://www.calphad.com/combined_law_of_thermodynamics.html
41. E.N. Hiebert, *Historical Roots of the Principle of Conservation of Energy* (Ayer Co Publ, Madison, 1981)
42. The Energy Story—Chapter 13—Nuclear Energy—Fission and Fusion, http://www.energyquest.ca.gov/story/chapter13.html www.aip.org/history/einstein/voice1.htm
43. The Equivalence of Mass and Energy, http://library.thinkquest.org/3471/energy_mass_equivalence_body.html
44. Energy, Work and Heat: The First Law, http://www.chem.arizona.edu/~salzmanr/480a/480ants/enwohe1/enwohe1.html
45. H.S. Leff, A.F. Rex, *Maxwell's Demon: Entropy, Information, Computing* (Princeton University Press, Princeton, 1990)
46. Entropy-General Systemics, http://www.generalsystemics.com/en/Entropy
47. Wikibooks, http://en.wikibooks.org/wiki/Entropy_for_Beginners
48. Wikipedia, http://en.wilipedia.org/wiki/Entropy_(statistical_thermodynamics)

49. Renyi, On the measures of entropy and information. Proc. 4th Berkeley Symp. Math. Statist. Probl. **1**, 547–561 (1961). http://digitalassets.lib.berkeley.edu/math/ucb/text/math_s4_v1_article-27.pdf
50. C. Tsallis, Possible generalization of Boltzmann-Gibbs statistics. J. Stat. Phys. **52**, 479–487 (1988). http://www.csc.umich.edu/∼crshalizi/notabene/tsallis.html http://www.mlahanas.de/Greeks/new/Tsallis.htm
51. D. Jiulin, Property of Tsallis entropy and principle of entropy increase. Bull. Astr. Soc. India **35**, 691–696, 2007
52. S. Abe, Y. Okanoto, *Nonextensive Statistical Mechanics and its Applications*, Lecture Series Notes in Physics (Springer, Heidelberg, 2001)
53. M. Masi, A step beyond Tsallis and Renyi entropies. Phys. Lett. A **338**(3-5), 217–224 (2005)
54. P.N. Nathie, S. Da Silva, Shannon, Lévy, C. Tsallis: A note. Appl. Math. Sci. **2**(28), 1359–1363 (2008)
55. J. Havrda, F. Charvat, Quantification method of classification processes: concept of structural α-entropy. Kybernetika **3**, 30–35 (1967)
56. M.D. Esteban, D. Morales, A summary on entropy statistics. *Kybernetika* **31**(4), 337–346 (1195)
57. H. Reichenbach, *The Direction of Time*, 2nd edn. (Dover Publications, New York, 1991)
58. H.D. Zeh, *The Physical Basis of the Direction of Time*, 5th edn. (Springer, Berlin, 2007)
59. M.W. Zemansky, R.H. Dittman, *Heat and Thermodynamics* (Mc Graw Hill, Singapore, 1981)
60. P.T. Landsberg, *Thermodynamics with Quantum Statistical Illustrations* (Interscience, New York, 1961)
61. N.N. Krasovskii, *Stability of Motion,* Translated by J.L. Brenner (Stanford University Press, Stanford, Calif., 1963)
62. A. Masim, R.U. Ayres, L.W. Ayres, An application of exergy accounting in five basic metal industries. *Working Paper,* INSEAD, Fontainebleau, France, May 2001 (Revision of 96/65 EPS)
63. Bibliography on Exergy, http://www.exergy.se
64. W. Nernst, Studies in chemical thermodynamics. *Nobel Lecture,* 12 Dec 1921
65. Max Planck, *The Theory of Heat Radiation* (Springer, Berlin, 1915)
66. Third Law of Thermodynamics (Wikipedia). http://en.wikipedia/wiki/Third_Law_of_Thermodynamics
67. Io HT: 15 + variations of the fourth law of thermodynamics, http://www.humanthermodynamics.com/4th-Law-Variations.html
68. A.J. Lotka, Contribution to the energetics of evolution. Proc. N.A.S. Biol. **8**, 147–151 (1922). http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1085052/pdf/pnas01891%2D0031.pdf
69. A.J. Lotka, Natural selection as a physical principle. Proc. Natl. Acad. Sci. **8**, 15–154 (1922)
70. H.T. Odum, Limits of remote ecosystems containing man. Am. Biol. Teach. **25**(6), 429–443 (1963)
71. H.T. Odum, *Ecological and General Systems: An Introduction to Systems Ecology* (Colorado University Press, Niwot Colorado, 1983)
72. H.T. Odum, Self-organization and maximum empower, in *Maximum Power: The Ideas and Applications of H.T. Odum,* ed. by C.A.S. Hall, (Colorado University Press, Colorado, 1995)
73. M. Tribus, *Generalized Treatment of Linear Systems Used for Power Production Thermostatics and Thermodynamics* (Van Nostrand, New York, 1961)
74. H.T. Odum, R.C. Pinkerton, Times speed regulator: the optimum efficiency for maximum output in physical and biological systems. Am. Sci. **43**, 331–345 (1955)
75. Impedance Matching, http://www.maxim-ic.com/appnotes.cfm/appnote_number/1849
76. Atwood Machine, http://hyperphysics.phy-astr.gsu.edu/hbase/Atwd.html http://demonstrations.wolfram.com/AtwoodsMachine

77. C. Giannantoni, Mathematics for generative processes: living and non-living systems. J. Comput. Appl. Maths. **189**(1–2), 324–340 (2006)
78. Maximum Power Principle: State Master Encyclopedia, http://www.statemaster.com/encyclopedia/Maximum-power-principle
79. T.T. Cai, T.W. Olsen, D.E. Campbell, Maximum (Em) power: a foundational principle linking man and nature. Ecol. Model. **178**(1–2), 115–119 (2004)
80. H.T. Odum, B. Odum, Concepts and methods of ecological engineering. Ecol. Eng. **20**, 339–361 (2003). Available on line at www.sciencedirect.com
81. H.T. Odum, Scales of ecological engineering. Ecol. Eng. **6**(1–3), 7–19 (1996)
82. H.T. Odum, *Environment, Power, and Society for the Twenty—First Century: The Hierarchy of Energy* (Columbia University Press, New York, 2007)
83. H.T. Odum, *Environmental Accounting, Emergy and Decision Making* (Wiley, NY, 1996)
84. L. Onsager, Reciprocal relations in irreversible processes I. Phys. Rev. **37**(4), 405–426 (1931). (http://prola.aps.org/abstract/PR/v37/i4/p405_1931)
85. L. Onsager, Reciprocal relations in irreversible processes II. Phys. Rev. **38**(12), 2265–2279 (1931). (http://prola.aps.org/abstract/PR.v38/i12/p2265_1931)
86. S.R. De Groot, P. Mazur, *Non-Equilibrium Thermodynamics* (North-Holland Publ. Co., Amsterdam, 1962)
87. P. Rysselberghe, *Thermodynamics of Irreversible Processes* (Herman Paris and Blaisdell Publ. Co., New York, 1963)
88. J. Verhas, Onsager's reciprocal relations and some basic laws. J. Comp. Appl. Mech. **5**(1), 157–163 (2004)
89. Onsager Reciprocal Relations—Definition from Answers.com http://www.answers.com/topic/onsager-reciprocal-relations
90. H.J. Morowitz, *Energy Flow in Biology* (Academic Press, New York, 1968)
91. H.J. Morowitz, *The Facts of Life* (Oxford University Press, New York/Oxford, 1992)
92. The Fourth Law of Thermodynamics. http://www.madsci.org (1994)
93. What is the Fourth Law of Thermodynamics, http://madsci.org/posts/archives/2000-11/973893769.Ch.r.html (1996)
94. S. Kauffmann, *Investigations* (Oxford University Press, New York/Oxford, 2000)
95. R.E. Morel, G. Fleck, A fourth law of thermodynamics. Chemistry **15**(4), 305–310 (2006)
96. M. Shibi, *The foundation of the fourth law of thermodynamics: Universe dark energy and its nature _ Can dark energy be generated?* Proceedings of International Conference on Renewable Energies and Power Quality (Seville, Spain, 2007)
97. P. Carr, A proposed fifth law of thermodynamics. http://www.canadaconnects.ca/quantumphysics/100/8
98. W. Muschik, Survey of some branches of thermodynamics. J. Non-Equilibr. Thermodyn. **33** (2), 165–198 (2008)
99. A.F. Horstmann, Theorie der dissociation. *Liebig's Ann. Chemie & Pharmacie,* Bd. **170** (CLXX), 192–210 (1973)
100. F.H. Garrison, Josiah Willard Gibbs and his relation to modern science. Popular Sci., 470–484 (1909)
101. W. Gibbs, On the equilibrium of heterogeneous substances. Trans. Conn. Acad. III, pp. 108–248, Oct. 1875–May 1876, and pp. 34–524, May 1877–July 1878
102. H. von Helmholtz, The thermodynamics of chemical operations (Die Thermodynamik Chemischer Vorgänge), in *Wissenschaftiliche Abhandlungen von Hermann von Helmholtz* (three volumes), ed. by J.A. Barth, Leipzig, vol. 2, (1882–95), pp. 958–978
103. G.N. Lewis, M. Randall, *Thermodynamics and the Free Energy of Chemical Substances* (McGraw-Hill, New York, 1923)
104. E.A. Guggenheim, *Modern Thermodynamics by the Methods of Willard Gibbs*, vol 8 (Taylor and Francis, London, 1933–1938)
105. J. Boerio-Goates, B.J. Ott, *Chemical Thermodynamics: Principles and Applications* (Academic Press, New York, 2000)

106. J.R. Howell, R.O. Buckius, *Fundamentals of Engineering Thermodynamics* (Mc Graw-Hill, New York, 1992)
107. M.J. Moran, Engineering thermodynamics, in *Mechanical Engineering Handbook*, ed. by F. Kreith (CRC Press, Boca Raton, 1999)
108. G.W. Dixon, Teaching thermodynamics without tables—Isn't time? http://www.computer.org/portal.web/csd1/doi/10.1109/IFITA.2009
109. S.I. Sandler, *Chemical and Engineering Thermodynamics* (McGraw-Hill, New York, 1989)
110. R.A. Alberty, *Biochemical Thermodynamics: Applications of Mathematics* (Wiley, New York, 2006)
111. R.A. Alberty, *Thermodynamics of Biochemical Reactions* (Wiley, New Jersey, 2003)
112. D. Lehninger, M. Nelson, *Cox, Principles of Biochemistry* (Worth Publishers, New York, 1993)
113. D. Haynie, *Biological Thermodynamics* (Cambridge University Press, Cambridge, 2001)
114. Biochemical Thermodynamics. http://www.chem.uwec.edu/Chem400_F06/Pages/lecture_notes/lect03/Atkins-Ch1.pdf
115. Bioenergetics. http://www.bmb.leeds.ac.uk/illingworth/oxphos/physchem.htm
116. R.A. Alberty, *Biochemical Thermodynamics* (Wiley Interscience, Published Online, 2006). http://www.interscience.wiley.com
117. J. Solids, Langmuir, the constitution and fundamental properties of solids and liquids, Part I. Amer. Chem. Soc. **38**, 2221–2296 (1916)
118. M. Jaroniec, A. Derylo, A. Marczewski, The Langmuir-Freundlich equation in adsorption from dilute solutions in solids. Monatshefte für Chemie **114**(4), 393–397 (2004). http://www.rpi.edu/dept/chem-eng/Biotech-Environ/Adsorb/equation.htm
119. M. Fletcher, *Bacterial Adhesion: Molecular and Ecological Diversity* (J. Wiley/IEEE, Hoboken, 1996)
120. J.B. Fein, Quantifying the effects of bacteria on adsorption reactions in matter-rock systems. Chem. Geol. **169**(3–4), 265–280 (2000)
121. Online Surface Science Tutorials and Lecture Courses. http://www.uksaf.org/tutorials.html
122. W. Zdunkowski, A. Bott, *Thermodynamics of the Atmosphere—A Course in Theoretical Meteorology* (Cambridge University Press, Cambridge, 2004)
123. J.V. Iribarne, W.L. Godson, *Atmospheric Thermodynamics* (Kluwer, Boston, 1981) (available from Springer)
124. Thermodynamics of the atmosphere, http://www.auf.asn.au/meteorology/section1a.html
125. A.A Tsonis, *An Introduction to Atmosphere Thermodynamics* (Cambridge University Press, Cambridge, 2002)
126. E.P. Odum, *Fundamentals of Ecology* (Saunders, Philadelphia, 1953)
127. H.T. Odum, E.C. Odum, *Energy Basis for Man and Nature* (McGraw-Hill, New York, 1976)
128. R.E. Vlanowicz, *Ecology the Ascendant Perspective* (Columbia University Press, New York, 1997)
129. S.E. Jorgensen, J.J. Kay, *Thermodynamics and Ecology* (Lewis Publishers, Boca Raton, 1999)
130. About Thermodynamics and Ecology. http://www.jameskay.ca/about/thermo.html
131. M. Ruth, *Integrating Economics, Ecology and Thermodynamics* (Kluwer, Boston, 1993) (available form Springer)
132. Energy Flow and Efficiency. http://courses.washington.edu/anth457/energy.htm
133. E. Cook, The flow of energy in industrial society. Sci. Am. **224**(3), 134–148 (1971)
134. M.A. Little, E.B. Morren Jr., *Ecology, Energetics and Human Variability* (W.C. Brown, Dubuque, Iowa, 1976)
135. R.A. Rappaport, The Flow of Energy in an Agricultural Society. Sci. Am. **224**(3), 116–132 (1971)
136. L. Cemic, *Thermodynamics in Mineral Sciences: An Introduction* (Springer, Berlin/Heidelberg, 2005)

137. B.J. Wood, *Elementary Thermodynamics for Geologists* (Oxford University Press, Oxford, 1977)
138. D.K. Nordstrom, *Geochemical Thermodynamics* (Blackburn Press, Caldwell, 1986)
139. Geochemical Thermodynamics. http://www.geokem.co.nz/capab_thermo.html
140. G.M. Anderson, *Thermodynamics of Natural Systems* (Cambridge University Press, Cambridge, 2005)
141. D. Derkins, A. Kozioi, *Teaching Phase Equilibria: Thermodynamics*. http://serc.carleton.edu/research_education/equilibria/thermodynamics.html
142. R. Teisseyre, E. Majewski, R. Dmowska, J.R. Holton, *Earthquake Thermodynamics, and Phase Transformation in the Earth's Interior* (Academic Press, New York, 2000)
143. R.M. Wald, *Quantum Field Theory in Curved Spacetime and Blackhole Thermodynamics* (The University of Chicago Press, Chicago, 1994)
144. E.J. Chaisson, *Cosmic Evolution: The Rise of Complexity in Nature* (Harvard University Press, Cambridge, 2001)
145. J. Prigogine, E. Geheniau, P. Gunzig, Nardone, thermodynamics of cosmological matter creation. JSTOR-Proc. Natl. Acad. Sci. USA **85**(20), 7428–7432 (1988)
146. G.W. Gibbons, S.W. Hawking, Cosmological event horizons, thermodynamics, and particle creation. PROLA: Phys. Rev. Online Arch. D **15**(10), 2738–2751 (1977)
147. B. Gal-Or, Cosmological origin of irreversibility, time, and time anisotropies (I). Found. Phys. **6**(4), 407–426 (1976)
148. R.C. Tolman, *Relativity, Thermodynamics and Cosmology* (The Clarendon Press, Oxford, 1934)
149. Thermodynamics of the Universe—Wikipedia. http://en.wikipedia.org/wiki/Thermodynamics_of_the_universe
150. Black Hole Thermodynamics. http://www.eoht.info/page/Black+hole+thermodynamics
151. C.J. Adkins, *Equilibrium Thermodynamics* (Cambridge University Press, Cambridge, 1983)
152. D. Jou, J. Casas-Vazquez, G. Lebon, *Extended Irreversible Thermodynamics* (Springer, Berlin, 1993)
153. P. Glansdorff, I. Prigogine, *Thermodynamic Theory of Structure, Stability and Fluctuations* (Wiley, London, 1971)
154. S.P. Mahulikar, H. Herwig, Conceptual investigation of the entropy production principle for identification of directives for creation, existence and total destruction of order. Phys. Scr. **70**, 212–221 (2004)
155. S.R. De Groot, P. Mazur, *Non-Equilibrium Thermodynamics* (Dover, New York, 1962)
156. Y. Demirel, *Non Equilibrium Thermodynamics* (Elsevier, Amsterdam, 2007)
157. C. Bustamante, Liphardt, F. Ritort, The nonequibrium thermodynamics of small systems. Phys. Today, pp. 43–48, (2005)
158. G.W. Paltridge, Climate and thermodynamic systems of maximum dissipation. Nature **279**, 530–631 (1979). http://adsabs.harvard.edu/abs/1979Natur.279.630P
159. R. Swenson, Autocatakinetics, evolution, and the law of maximum entropy production. Adv. Hum. Ecol. **6**, 1–46 (1977)
160. I. Prigogine, *Etude Thermodynamique des Phenomenes Irreversibles* (Desoer, Liege, 1947)
161. Wikipedia (a) Non-equilibrium thermodynamics, (b) Maximum-Entropy production
162. N. Georgescu-Roegen, *The Entropy Law and Economic Process* (Harvard University Press, Cambridge, 1971)
163. H. Quevedo, M.N. Quevedo, Statistical thermodynamics of economic systems. http://arxiv.org/abs/0903.4216
164. P. Burley, J. Foster (eds.), *Economics and Thermodynamics: New Perspectives on Economic Analysis* (Springer, New York, 1996)
165. R.U. Ayres, Eco-thermodynamics: economics and the second law. http://www.sciencedirect.com
166. S. Sieniutycz, P. Salamon, *Finite-Thermodynamics and Thermoeconomics* (Taylor and Francis, London, 1990)
167. M.E. Yehia, *The Thermoeconomics of Energy Conversions* (Pergamon, London, 2003)

168. W. Lepkowski, The social thermodynamics of Ilya Prigogine. Chem. Eng. News **57**, pp. 16, 16 Apr 1979
169. M.M. Callaway, *The Energetics of Business* (Lincoln Park Publications, Chicago, 2006)
170. L. Liss, Human thermodynamics and business efficiency. J. Hum. Thermodyn. **1**(6), 62–67 (2005). http://www.eoht.info/page/Business+thermodynamics
171. Journal of Human Thermodynamics. http://www.human+thermodynamics.com/JHT/Business-Efficiency.html
172. A.D. Little, The place of chemistry in business. Technol. Rev. **25**, 360–362 (1922)
173. http://www.eoht.info/page/Business+chemistry
174. Business Operation Efficiency Analysis. http://www.economy-point.org/b/business-operation-analysis
175. S. Wolfram, *Statistical Mechanics of Cellular Automata* (The Inst. For Advanced Studies, Princeton, N.J., 1983). http://www.ifsc.usp.br/~lattice/artigo-wolfram-cellular-autom.pdf
176. S., Wolfram, *Cellular Automata and Complexity: Collected Papers* (Addison—Wesley Pub., Reading, MA, 1994)
177. P. Salamon, P. Sibani, R. Frost, *Facts, Conjectures, and Improvements for Simulated Annealing* (Society for Industrial and Applied Mathematics Publ, Philadelphia, 2002)
178. M.A. Nielsen, I.L. Chuang, *Quantum Computation and Quantum Information* (Cambridge University Press, Cambridge, 2000)
179. H. Baker, Thermodynamics and Carbage collection. ACM Sigplan Not. **29**(4), 58–63 (1994). ftp://ftp.netcom.com/pub/hb/hbaker/ThermoGC.ps.Z
180. W. Peng, J. Li, Problem's thermodynamics energy analysis method. Proc. Intl. Forum Inf. Technol. Appl. (IFITA), **2**, 206–208, Chengdu, China, 15–17 May 2009
181. W. Sidis, *The Animate and the Inanimate* (R.G. Badger, Boston, 1925)
182. L. Boltzmann, The second law of thermodynamics, in Ludwig Boltzmann: *Theoretical Physics and Philosophical Problems—Select Writings*, ed. by B. McGuiness (D. Reider, Dordrecht, Netherlands (available from Springer, 1886)
183. J. Johnstone, *The Philosophy of Biology* (Cambridge University Press, Cambridge, 1914), p. 54
184. Entropy and Life: Wikipedia, http://en.wikipedia.org/wiki/Entropy_and_life
185. E. Schrödinger, *What is Life? Ch.6: Order, Disorder and Entropy* (Cambridge University Press, Cambridge, 1944)
186. J. Chen, *Universal Natural Law and Universal Human Behavior* (2002), http://papers.ssrn.com/s013/papers.cfm?abstract_id=303500
187. Human Thermodynamics: The science of Energy Transformation, http://www.humanthermodynamics.com
188. Encyclopedia of Human Thermodynamics (a) http://www.eocht/info (b) http://www.eocht.info/thread/3300562/Entropy+in+20th+Century+Chemistry
189. P.T. Landsberg, *Seeking Ultimates: An Intuitive Guide to Physics* (CRC Press, Taylor and Francis, London, 1999)
190. A.R. Ubbelohde, *Time and Thermodynamics, Chap. IX: Thermodynamics and Life* (Oxford University Press, Oxford, 1947)
191. R.B. Lindsay, Entropy consumption and values in physical science. Am. Sci. **47**(3), 376–385 (1959)
192. I. Prigogine, *Introduction to Thermodynamics of Irreversible Processes* (Interscience Publishers, New York, 1955), p. 12
193. I. Prigogine, I. Stengers, *Order out of Chaos: Man's New Dialogue with Nature,* Flamingo, (Mountain Man Graphics Web Publ. Australia 1984). http://www.mountainman.com/chaos_02.htm
194. M. Perutz, Erwin Schrödinger's what is Life and molecular biology? in *Schrödinger: Centenary celebration of a Polymath*, ed. by C.W. Kilmister (Cambridge University Press, Cambridge, 1987)
195. T. Fararo, *The Meaning of General Theoretical Sociology: Tradition and Formalization* (Cambridge University Press, Cambridge, 1992), pp. 86–87

196. I. Aoki, Entropy production in human life span: a thermodynamic measure of aging. Age 1, 29–31 (1994)
197. E.D. Schneider, J.J. Kay, order from disorder: the thermodynamics of complexity in biology, in *"What is Life: The Next Fifty Years Reflections on* the *Future of Biology*, ed. by M. P. Murphy, L.A.J. O'Neil (Cambridge University Press, Cambridge, 1995), pp. 161–172
198. F. Capra, *The Hidden Connections: A Science for Sustainable Living* (Anchor Books, New York, 2002), pp. 13–32
199. E.D. Schneider, D., Sagan, *Into the Cool: Energy Flow*, *Thermodynamics and Life* (Part I—The Energetic, Part II—The Complex, Part III—The Living, Part IV—The Human) (The University of Chicago Press, Chicago, 2005)
200. G.P. Cladyshev, What is life? Bio-physical perspectives. Adv. Gerontol. **22**(2), 233–236 (2009). http://www.ncbi.nlm.nih.gov/pubmed/19947386
201. G.P. Gladyshev, On the thermodynamics of biological evolution. J. Theor. Biol. **75**(4), 425–441 (1978)
202. G. Gladyshev, *Thermodynamic Theory of the Evolution of Living Beings* (Nova Science Publishers, Commack, New York, 1997), p. 137
203. G. Gladyshev, Thermodynamic self-organization as a mechanism of hierarchical structure formation of biological matter. Prog. React. Kinet. Mech. **28**(2), 157–188 (2003)
204. G.L. Murphy, Time thermodynamics and theology, Zygon. J. Relig. Sci. **26**(3), 359–372 (1991)
205. E.W. Barnes, *Scientific Theory and Religion: The World Described by Science and Its Spiritual Interpretation* (The University Press, Aberdeen, 1933)
206. H. Morowitz, *The Emergence of Everything* (Oxford University Press, Oxford, 2002)
207. P. Weiss, Another Face of Entropy: Particles Self-Organize to Make Room for Randomness. Sci. News **154**, 108–109 (1998)
208. S. Kaufman, *At Home in the Universe: The Search for the Laws of Self-Organization and Complexity* (Oxford University Press, Oxford, 1996)
209. S. Kaufman, *The Origins of Order: Self-Organization and Selection in Evolution* (Oxford University Press, Oxford, 1993)
210. B. Klyce, The Second Law of Thermodynamics. http://www.panspermia.com/secondlaw.htm
211. L. Boltzmann, The Second Law of Thermodynamics (address to a Meeting of the Imperial Academy of Science, 29 May 1886), Reprinted in: L. Boltzmann *Theoretical Physics and Philosophical Problem* (S.G. Brush, translator) (D. Reidel, Boston, 1974)
212. H.B. Callen, *Thermodynamics and an Introduction to Thermostatics* (J. Wiley, New York, 2001)
213. P.T. Landsberg, Can entropy and order increase together? Phys. Lett. **102**-A, 171–173, 1984
214. P.T. Landsberg, Is equilibrium always an entropy maximum? J. Statist. Phys. **35**, 159–169 (1984)
215. D.F. Styer, Insight into entropy. Am. J. Phys. **68**(12), 1090–1096 (2000)
216. E.P. Gyftopoulos, Entropy: an inherent, nonstatistical property of any system in any state. Int. J. Thermodyn. **9**(3), 107–115 (2006)
217. E.P. Gyftopoulos, Entropies of statistical mechanics and disorder versus the entropy of thermodynamics and order. J. Energy Resour. Technol. **123**, 110–123 (2001)
218. F.L. Lambert, Disorder—a cracked crutch for supporting entropy discussions. J. Chem. Educ. R. Soc. Chem. **79**, 79 (2002). http://jchemed.chem.wisc.edu/HS/Journal/Issues/2002/Feb/abs187.html http://www.entropysite.com/cracked_crutch.htm
219. W. Thomson, On a universal tendency in nature to the dissipation of mechanical energy. Proc. R. Soc. Edinb. 19 Apr 1852. http://zapatopi.net/kelvin/papers/on_a_universal_tendency.html
220. Energy Dispersal – Wikipedia, http://en.wilipedia.org/wiki/Energy-dispersal
221. K. Denbigh, *Principles of Chemical Equilibrium* (Cambridge University Press, Cambridge, 1981)
222. P. Atkins, *The Second Law* (J. Sci. Am. Libr., New York, 1984)

223. P. Atkins, J. De Paula, *Physical Chemistry* (Oxford University Press, Oxford, 2006)
224. J. Wrigglesworth, *Energy and Life Modules in Life Sciences* (CRC, Boca Raton, Fl, 1997)
225. M.C. Gupta, *Statistical Thermodynamics* (New Age Publishers, Ringwood, 1999)
226. C. Starr, R. Taggart, *Biology: The Unity of Diversity of Life* (Wadsworth Publishers, Las Vegas, 1992)
227. Scott, *Key Ideas in Chemistry* (Teach Yourself Books, London, 2001)
228. http://entropysite.oxy.edu/entropy_isnot_disorder.htm http://entropysite.oxy.edu/students_approach.html http://entropysite.com/entropy_is_simple/index.html#microstate
229. Energy Dispersal and Entropy Misinterpretations—Encyclopedia of Human Thermodynamics http://www.eoht.info/page/Energy+dispersal http://www.eoht.info/page/Entropy+(misinterpretations)
230. R. Swenson, in *Emergence and the principle of maximum entropy production: multi-level system theory, evolution and non-equilibrium thermodynamics*. Proceedings of 32nd Annual Meeting of the International Society for General Systems Research, (1988), p. 32
231. R. Swenson, Emergent attractors and the law of maximum entropy production: foundations to a theory of general evolution. Syst. Res. **6**, 187–198 (1989)
232. R. Swenson, M.T. Turvey, Thermodynamic reasons for perception action cycles. Ecol. Psychol. **3**(4), 317–348 (1991)
233. R. Swenson, Spontaneous order, evolution and autocatakinetics: the nomological basis for the emergence of meaning, in *Evolutionary Systems*, ed. by G. van de Vijver et al. (Kluwer, Dordrecht, The Netherlands, 1998)
234. The Law of Maximum Entropy Production. http://www.lawofmaximumentropyproduction.com/ http://www.entropylaw.com
235. G.N. Hatsopoulos, J.H. Keenan, A single axiom for classical thermodynamics. ASME J. Appl. Mech. **29**, 193–199 (1962)
236. G.N. Hatsopoulos, J.H. Keenan, *Principles of General Thermodynamics (Law of Stable Equilibrium)* (J. Wiley, New York, 1965), p. 367
237. J. Kestin, *A Course in Thermodynamics* (Blaisdell Publ. Co., New York, 1966)
238. E.D. Schneider, J.J. Kay, Life as a manifestation of the second law of thermodynamics. Math. Comput. Model. **19**(6–8), 25–48 (1994)
239. I. Walker, Maxwell demon in biological systems. Acta Biotheor. **25**(2–3), 103–110 (1976)
240. W. Brown, Science: *A Demon Blow to the Second Law of Thermodynamics* (New Scientist, Issue 1725, 14 July 1990)
241. J. Earman, D.J., Exorcist XIV: The Wrath of Maxwell's Demon, Part I. From Maxwell to Szilard, Studies in History and Philosophy of Science, Part B, **29**(4), 435–471, 1998; Part II. From Szilard to Landauer and Beyond, *ibid*, **30**(1), 1–40 (1999), http://philopapers.org/rec/EAREXT-2
242. E.P. Gyftopoulos, Maxwell's Demon: (I) a thermodynamic exorcism. Phys. A **307**(3-4), 405–420 (2002)
243. E.P. Gyftopoulos, Maxwell's Demon: (II) a quantum theoretic exorcism, *ibid,* **307**, (3–4), 421–436 (2002)
244. E.P. Gyftopoulos, G.P. Beretta, *What is the Second Law of Thermodynamics and Are there any limits to its Validity?* Quant-ph/0507187 (Elsevier Science, 19 July 2005)
245. V. Capek, D.P. Sheehan, *Challenges to the Second Law of Thermodynamics* (Springer, Dordrecht, 2005)
246. R.S. Morris, *Time's Arrows: Scientific Attitudes toward Time* (Simon & Schuster, New York, 1985)
247. J.T. Fraser, *Time: The Familiar Stranger* (MIT Press, Boston, Mass, 1987)
248. B. Goertzel, *From Complexity to Creativity: Explorations in Evolutionary, Autopoietic Dynamics* (Plenum Press, New York, 1996). Also: http://www.goertzel.org/papers/timepap.html
249. D. Pacchidi, The Arrow of Time, The Pennsylvania State University. http://www.rps.edu/time/arrow.html

250. D. Layzer, The Arrow of Time, Scientific American, 1975. http://www.scientificamerican.com/media/pdf/2008-05-21_1975-carroll-story.pdf
251. J.-G. Cramer, An overview of the transactional interpretation of quantum mechanics. Intl. J. Theor. Phys. **27**(2), 227–236 (1988)
252. R.C. Tolman, *The Principles of Statistical Mechanics* (Dover Publications, New York, 1979)
253. W.E. Burns, *Science in the Enlightenment: An Encyclopedia* (ABC-CLIO, Santa BarBara, CA, 2003), pp. 109–111
254. T. Levere, *Affinity and Matter-Elements of Chemical Philosophy (1800–1865)* (Taylor and Francis, London/New York, 1993)
255. J.R. Partington, *A Short History of Chemistry* (MacMillan and Co., New York, 1957)
256. H. Devoe, *Thermodynamics and Chemistry* (Prentice Hall, Saddle River, 2000)
257. Scribd: Hierarchical thermodynamics: general theory of existence and living world development. http://www.scribd.com/doc/8543701/Hierarchical-thermodynamics-general-theory-of-existence-and-living-world-development http://www.endeav.org/?id=47
258. S. Sandler, *Chemical and Engineering Thermodynamics* (Wiley, New York, 1989)
259. G.N. Hatsopoulos, From Watt's Steam engine to the unified quantum theory of mechanics and thermodynamics. Int. J. Thermodyn. **9**(3), 97–105 (2006)
260. J.C. Maxwell, E. Garber, S.G. Brush, *Maxwell on Molecules and Gases* (MIT Press, Cambridge, 1986)
261. G.P. Gladyshev, The principle of substance stability applicable to all levels of organization of living matter. Int. J. Mol. Sci. **7**, 98–110 (2006)
262. R.W. Sterner, J.J. Elser, *Ecological stoichiometry: The Biology of Elements from Molecules to the Biosphere*, Chap. 1 (Princeton University Press, Princeton, 2002)
263. C.M. Wieland, Letter to Karl August Bottiger, Weimar Quoted from Tantillo, 16 July 2001, pp. 9–10
264. S. Fuller, "I am not a molecule". New Sci. Issue 2502 (2007)

# Chapter 4
# Information I: Communication, Transmission, and Information Theory

*The importance of information is directly proportional to its improbability.*

Jeremia Eugene Pournelle

*Intelligence is not information alone but also judgment, the manner in which information is collected and used.*

Carl Sagan

**Abstract** Information is a basic element of life and society involved in all areas of human, scientific, technological, economic, and developmental activity. Information storage, flow, and processing are inherent processes in nature and living organisms. Information transmission and communication/networking techniques contribute to the development of modern society, including social, economic, business, scientific, and technological operations and activities. This chapter covers at a conceptual level the following issues of information: definition, historical landmarks of its manifestations, communication models, modulation/demodulation, computer networks, multimedia, informatics/telematics, Shannon information entropy, source and channel coding/decoding, and theorems of information theory. The above sets of information/communication models, techniques, and technologies are affecting, and will continue to increasingly affect, the social, economic/business, and developmental activities of people in the short- and long-term future.

**Keywords** Information · Communication systems and models · Modulation Demodulation · Analog modulation/demodulation · Frequency modulation/demodulation · Phase modulation/demodulation · Pulse modulation/demodulation Information theory/theorems · Shannon's entropy · Coding/decoding Source coding · Channel coding · Hamming distance · Convolutional codes Error detecting and correcting codes

## 4.1 Introduction

The term "*information*" is very diverse and is used in almost all areas of human scientific, technical, economic, and societal activity. In particular, information covers communication theory, information theory, information technology,



S. G. Tzafestas, *Energy, Information, Feedback, Adaptation, and Self-organization*, Intelligent Systems, Control and Automation: Science and Engineering 90, https://doi.org/10.1007/978-3-319-66999-1_4

informatics, documentation science, and information systems that are the result of the interactions between technology and human processes.

Very broadly, the development of human communication and information dissemination has spanned three main periods:

- Speech and language development era.
- Development of writing era.
- Information revolution era.

Speech and language enabled humans to exchange knowledge with others who were physically at the same place. Writing has enabled communication between people at different places and times and initiated the development of civilization. The information revolution is still on-going and expanding with advancements in computer science and engineering, telecommunication systems, knowledge science, and computer networks, the biggest and most powerful of which is the Internet.

The above shows that "*information*" is indeed one of the "*basic pillars*" of human life and society, as will be addressed in this book. In the present chapter, we will guide the reader to the concept of information, including a listing of the key historical landmarks of its particular manifestations and branches. The core of the chapter is the tour to communication and information transmission theory, namely, communication models, modulation/demodulation, Shannon's information entropy, source coding/channel coding, and theorems of information theory. The material in the chapter is not intended to constitute a complete treatment of the subject, but it is sufficient for the purposes of this book which principally aims at highlighting the impact of the "*information pillar*" on human life and society (Chap. 11). The fields of information documentation science, information science, information technology, and information systems will be discussed in the next chapter.

## 4.2 What Is Information?

According to the Oxford English Dictionary (1989), the two main contexts in which the term *information* is used are as follows:

- The act of molding the mind.
- The act of communicating knowledge.

The two processes, "*molding*" and "*communicating*", are intrinsically related and, in most cases, occur inseparably (although not in a clearly known way).

Other frequently used interpretations of the term information include the following (Dictionary.com):

- Knowledge communicated or received concerning a particular fact or circumstance.
- Knowledge gained through study, communication, research, instruction, experience, etc.
- The act or fact of informing.
- An indication of the number of possible choices or messages.

- Important or useful facts obtained as output from a computer via the processing input data with a program.
- Data at any stage of processing (input, storage, output, transmission, etc.).
- A numerical measure of the uncertainty of an experimental outcome.
- The result of applying data processing to data, giving it context and meaning. Information can then be further processed to yield knowledge.
- Directory assistance.

The concept of information as "*knowledge communicated*" has a dominant position in modern human life and society and has been fully developed after World War II with the advancement and use of information science, informatics, computer science, computer networks, information theory, information technology, and information systems. From an epistemological viewpoint, the information concept has also found extensive and important use in biology, physics, psychology, etc.

The term *information* has a Latin origin (*informatio*, *informo*) [1, 2] and is composed by the prefix "*in*" and the word "*formatio*" or "*formo*" which has the meaning of giving a form to something. The prefix "*in*" here is used to make stronger the act of this form giving. There are two primary contexts in which the term *information or informo* has been used, the first is a *tangible* (*corporaliter*) context and the second an *intangible* (*incorporaliter*) concept. The tangible context refers to low-level processes, and the intangible context to high-level processes (moral, pedagogical, spiritual, mental). Capurro [1] has studied the Greek origins of the term *informatio* and its subsequent evolution. The Greek origin is evident in the works of Cicero (106–43 B.C.) and Augustine (354–430 A.D.) [1, 2].

The Greek term for information is *pliroforia* (πληροφορία) composed by the two words "*pliro*" (πλήρης = complete) and "*foria*" (φέρω = carry/bring). The word pliroforia indicates that the common meaning (sense) of a simple or complex symbol consisting of two or more subjects is completely carried. Conceptually, the term pliroforia (information) signals the content that is complete and clear (in whatever form). In computer science, the information is reflected in the qualitative value of the "*bit* = **bi**nary digi**t**" 0 or 1 or the **qu**antum **bit** (*qubit*) in quantum computers. The computer processes the data (sequences of 0s, 1s) and provides processed data. The human gives *meaning* to these processed data and converts them into "information".

It must be remarked that today the word *information* is used in almost all scientific fields with various different meanings depending on the subject of each field and the processes studied therein (e.g., decrease in entropy, physical organization, a communication pattern, a form of feedback control, the meaning of a linguistic term, the probability of a signal transmitted over a communication channel, etc.). According to Bogdan [3]: "There seems to be no unique idea of information upon which these various concepts converge and hence no proprietary theory of information."

The issue of whether the concept of "information" should include necessarily a human knower or an interpretative component or exclude mental processes and user-oriented intentions and address only an objective magnitude or property of human beings is of continuous concern by scientists and philosophers [4]. Several approaches that belong between these two extremes, including the need for a unified theory of information, have been proposed [5].

In [6], *Machlup* and *Mansfield* present their view that: "Information is addressed to human minds and is received by human minds." All other senses are *metaphoric* and *anthropomorphic*. That is, the basic senses in information deal with "telling something or with the something that is being told." Therefore, according to [6], the term information is not appropriate for use in the context of signal transmission. In overall, information is a human phenomenon that involves individuals transmitting and receiving messages in the framework of their possible decisions and actions.

In [7], *Capurro* defines *information* as an *anthropological class* referring to the phenomenon of human messages with vertical and horizontal structures related to the Greek concept of *message* (αγγελία: aggelia or μήνυμα: menima) and the philosophic discourse (λόγος: logos). The debate about the unification and naturalization of the term *information* goes back to Boltzmann, Neumann, Nyquist, Wiener, and Hartley.

In [8], *R.V.L. Hartley* states: "*it is desirable to eliminate the psychological factors involved [in electrical transmission systems] and to establish a measure of information in terms of purely physical quantities*" (This is because electrical transmission systems have to do not with humans but with machines).

According to *C.S. Pierce* (1839–1914) and *C.W. Morris* (1901–1979), the information transmitted through a communication channel between an emitter and a receiver involves three levels:

- Syntactic level.
- Semantic level.
- Pragmatic level.

At the *syntactic level*, the information deals with the formal bonds that exist between the various elements that make up the information, the rules of the communication code, the capacity of the channels, and the system design and coding methods for the transmission, processing, and storage of the information.

The *semantic level* is concerned with the ways of expressing the meaning of the information (e.g., written or nonwritten, cultural, societal, or moral rules or traditions valid in the human group or society concerned).

At the *pragmatic level*, the information is related to its utility. This level is determined to a large extent by the background of the person(s) who receive the information (political, social, economic, psychological, and moral factors).

The above three levels have a hierarchical structure in which the information can be managed (transmitted, processed, stored) at the syntactic level without the need to necessarily know its conceptual content at the semantic level or its practical utility at the pragmatic level.

Indeed, the word *information* in *Shannon's* "Mathematical Theory of Communication" is used in a sense that must not be confused with *meaning* [9]. Shannon and Weaver state: "*the semantic aspects of communication are irrelevant to the engineering aspects*" [10]. Of course, as *Capuro* and *Hjorland* note [2]: "this does not mean that the engineering aspects are necessarily irrelevant to the semantic aspects.

The philosophical discussion about the concept of information was originated by *Norbert Wiener* who stated that "*Information is information, not matter or energy. No materialism which does not admit this can survive at the present day*" [2]. In the twentieth century, the notions of information and communication were applied at higher levels of abstraction and not to the communication of human knowledge as expressed in the above Wiener quotation. According to Titze [11], information is not a metaphysical principle but refers to a natural tendency for order and evolution. According to Stonier [12], structural and kinetic information is an intrinsic element of the universe, which is independent of the existence or not of an intelligent agent that can perceive it or not. Information may be manifested as "*infons*" which are comparable to "*photons*" [13]. Stonier distinguishes and separates the syntactic from the semantic features of information, and adopts the emergence of a global brain called "*noosphere*" by *Teillard de Chardin* (from the Greek "νους" = noos = mind, and σφαίρα = sphere = domain/globe) [14].

## 4.3 Historical Landmarks

The concept of information is generic and today spans all branches of science, technology, and human society that deal with the generation, acquisition, organization, storage, retrieval, interpretation, transformation, processing transmission, and utilization of information. Therefore, the *history of information* includes the evolution of computers, (theoretical) computer science, computation, information theory, information technology information science, information systems, multimedia, and informatics. Clearly, a detailed exposition of the combined history of all these parallel and overlapping branches, which in one or the other way involve the concept of information, needs too much space, and so here only the main historical landmarks will be given.

On the basis of the principal technology employed for the input, processing, output, and communication processes, the history of *information technology and systems* can be divided into the following four periods [15–17]:

- Pre-mechanical period (3000 B.C.–1450 A.D.).
- Mechanical period (1450–1840 A.D.).
- Electromechanical period (1840–1940).
- Electronic period (1940–Present).

### *4.3.1 Pre-mechanical Period*

Humans have been attempting to facilitate calculation using mechanical devices and to find ways to communicate for thousands of years. The communication of people using pictures and symbols (alphabet) was started by the Phoenicians

(around 3500 B.C.). Around 2900 B.C., the Egyptians develop *hieroglyphic writing*. Around 2000 B.C., the Greeks enhance the Phoenician alphabet through the addition of vowels. The Egyptians were writing on papyrus and around 2600 B. C. Chinese make paper from rags on which the modern paper making is based. Around 600 B.C., religious and other books were used, to permanently store information, made by folding papyrus vertically into leaves. Egyptians developed the first number system 1, 2,…, 9 as vertical lines (with the number 10 as a U circle, etc.), a number system similar to the present was developed in India (around 10–200 A.D.), and the zero number was added in around 870 A.D. The ancient Greeks constructed some sophisticated analog computers such as the *Antikythera mechanism* (involving rusted metal gears and pointers) which has discovered in 1901 on the island of Antikythera [18]. Around 200 B.C., human messengers on foot or horseback started to be used in Egypt and China with messenger relay stations available. Fire signals were used on many occasions instead of human messengers.

### 4.3.2 Mechanical Period

Principal dates in this period are as follows:

- **10–20 A.D.**: Romans establish postal services and use mirrors for sending messages (heliographs).
- **100–200 A.D.**: First wooden printing presses used in China and the first bound books.
- **1450**: Movable metal-type printing process (Johannes Gutenberg, Germany).
- **1600**: Slide rule; an early form of analog computer (William Oughtred, England).
- **1641**: Blaise Pascal's adding machine
- **Late 1600s**: Gottfried Wilhelm Leibniz's adding machine.
- **1822**: Charles Babbage engines (difference engine, analytical engine).
- **1830–1840**: Parts and processes similar to modern computers (storage, punch card, binary logic, fixed program, real-time concept). It appears that the first programmer is *Ada Augusta Byron* (Countess of Lovelace), a friend of Babbage, who wrote a report on Babbage's machine. The name of the *Ada* programming language was chosen to honor her.

### 4.3.3 Electromechanical Period

This period was characterized by the conversion of information and knowledge into electric pulses and the rise of mathematics.

1. **The start-up of telecommunications**:
   - **Late eighteenth century**: Voltaic battery.
   - **Early 1800s**: Telegraph.
   - **1835**: Samuel Morse develops the telegraphic Morse code.
   - **1876**: Telephone (Alexander Graham Bell).
   - **1894**: Invention and development of radio (Guglielmo Marconi).
2. **Electromechanical computing**
   - **1880–1940**: Census tabulation machine (Herman Hollerith), early punch cards, punch card workers, IBM Mark 1.
3. **Some landmarks of mathematics**
   - **1928**: The German mathematician David Hilbert poses the following questions: (a) Is mathematics complete (i.e., can every mathematical statement be either proved or disproved?), (b) Is mathematics consistent (i.e., is it actually true that statements like 0 = 1 cannot be proven by a valid method?), and (c) Is mathematics decidable (i.e., does there exist a mechanical way that can confirm the validity or not of a mathematical statement?).
   - **1931**: The answers to two of Hilbert's questions were given by Kurt Gödel who proved that every sufficiently powerful formal system is either inconsistent or incomplete, and also that the consistence of a consistent axiom system cannot be proven within this system. The third question remains unanswered.
   - **1936**: Alan Turing (1912–1954) gave an answer to the third question of Hilbert, via his formal model of a computer, known as the *Turing machine*. He showed that his machine would not be able to solve any problem (e.g., the question: given a Pascal program, does it halt on all inputs?—the so-called halting problem).

### *4.3.4 Electronic Period*

The general-purpose electronic digital computer was developed during World War II. One of the major needs in this period was the automatic calculation of ballistic equations.

- **1940**: At Iowa State University, an electronic computer was built for solving systems of linear equations (John Vincent Atanasoff and Clifford Berry).
- **1945**: Development of EDVAC (Electronic Discrete Variable Computer).
- **1946**: Based on the ideas of Atanasoff and Berry, the ENIAC (Electronic Numerical Integrator and Computer) system was built, originally intended for artillery calculations. This was the first fully working high-speed general-purpose computer using vacuum tubes.

- **1948**: Construction of Manchester Mark I (the first stored-program computer).
- **1949**: Construction of EDSAC (Electronic Delay Storage Automatic Calculator).
- **Late 1940**: UNIVAC (Universal Automatic Computer). This is the first general-purpose computer for commercial use.

We now give a brief description of the development of the four generations of digital computing and information processing.

**1950–1958**: **First generation**. Logic circuits which use vacuum tubes, punch cards for storage of external data, internal storage of data and programs using magnetic drums, machine language, assembly language, and the need for a compiler.

**1959–1963**: **Second generation**. Logic circuits using transistors designed on semiconductors, magnetic disks and tapes, magnetic cores, and high-level languages such as FORTRAN and COBOL.

**1964–1979**: **Third generation**. Integrated circuits (ICs) instead of individual transistors, magnetic tapes and disks, metal–oxide–semiconductor (MOS) memories, operating systems, and advanced languages, e.g., BASIC.

**1980–Present**: **Fourth generation**. Large-scale integrated (LSI) and very large-scale integrated (VLSI) circuits, central processing units (CPUs) on a single chip leading to the development of personal computers (PCs), e.g., Apple II (1977), Apple Mac (1984), IBM PC (1981), Microsoft Disk Operating System (MS-DOS), graphical user interfaces (GUIs) for PCs (1980), and MS Windows (version 3, 1990)).

### *4.3.5 Information Theory Landmarks*

The field of *information theory* as we know it today was initiated by Shannon's celebrated paper: "A Mathematical Theory of Communication" [9]. Shannon realized and adopted the need to have a communication theory, in which the communication signals must be utilized separately from the meaning of the messages they transmit. He also realized that a signal can be transmitted arbitrarily close to the theoretical channel capacity, even if the signal is contaminated by noise. These initial ideas have inspired and guided information, communication and computer engineers ever since. Information theory overlaps considerably with communication theory, but it is principally concerned with the basic and theoretical constraints on the processing and communication of information and not with the design and operation of individual components and communication devices. The principal historical landmarks of information theory are the following [9, 19–23]:

- **1929**: The publication of Leó Szilard on the decrease in entropy caused by the intervention of intelligent agents [21].
- **1948**: The introduction by Shannon of the entropy concept in information theory as an expected value that expresses the "information content" in a message (in units, such as bits) [9].
- **1949**: Publication of the book of Robert Fano concerning the transmission of information [23].
- **1956**: The publication of the book of Leon Brillouin about information theory and his ambitious attempt to incorporate all of scientific endeavor within the framework of Shannon's information theory [22].
- **1957**: The paper of Edwin Jaynes on maximum entropy principle (MEP) [24].
- **1961**: Publication of the book of Myron Tribus in which he tried to formulate the laws of thermodynamics via information theory [25].
- **1988**: Publication of the work of George Saridis on the application of MEP methods [24, 26] to control [27, 28].
- **2006**: Publication of the book of George Klir [29] in which he develops a "generalized information theory" (GIT) by viewing uncertainty as a manifestation of some information deficiency, and information as the capacity to reduce uncertainty.
- **2007**: Publication of the book of Ben-Naim [30] on the benefits of the use of the concept (term) information or uncertainty, instead of the term entropy, in statistical mechanics and thermodynamics.
  More detailed historical landmarks of information theory, especially regarding the development of error-correcting codes and lossless data compression, can be found in Wikipedia [31].

### *4.3.6 Computer Networks, Multimedia, and Telematics Landmarks*

The development of computer networks and multimedia has, over the years, led to advanced and very accurate uses of transmitted information elements of any type for the benefit of present human society [32]:

**Computer networks** Before the widespread adoption of internetworking, which led to the Internet, the majority of communication networks were allowing communications only between the stations of the network. One of the dominating methods of computer networking was based on the central mainframe concept, in which its terminals were enabled to be connected via long-leased lines. This method was in use, for example, during the 1950s for research communication purposes between Pittsburgh (Pennsylvania) and Santa Monica (California). During the 1960s, a period in which the telephone network was the primary communication network in the world, many groups were working toward enhancing and implementing "packet switching", especially in the defense field. The origin of the

Internet was the development of the *Advanced Research Project Agency Network* (*ARPANET*), the first ARPANET link being realized between the University of California and Stanford (21 November 1969). ARPANET was enhanced with ideas drawn from the ALOHA net and grew rapidly. The ARPANET development was based on the Request for Comments (RFC) process, which continues to be used until this day.

Some of the primary landmarks in the history of computer networking are the following [33]:

**1971**: The *ALOHA net,* a type of TDMA transmission system and protocol for terrestrial and satellite random access communication (developed at the University of Hawaii) became operational in June 1971. ALOHA is a Hawaiian word (symbol) meaning hello, or goodbye or love and coming from "*Alo*" = presence/front/face and "*ha*" = breath.
**1974**: The *X.25 network* developed by the *International Telecommunication Union* (*ITU*) was used on the British *SERCnet* network of academic and research sites (later it became *JANET*). The first ITU standard on X.25 was approved in March 1976.
**1982**: The TCP/IP (Transmission Control Protocol and Internet Protocol) is established as the standard for ARANET.
**1986**: TCP/IP is offered on workstations and PCs.
**1989**: The number of hosts exceeds 100,000.
**1990**: Several search tools (ARCHIE, Gopher, WAIS, etc.) are starting to enter the market after the official shut down of ARPANET.
**1991**: Development of the *World Wide Web* (WWW) by Jim Berners-Lee at CERN (European Center for Nuclear Research).
**1992–Present**: The WWW and Internet explode into the world [34].

**Multimedia** is the branch of computer science and information technology which deals with the computer-controlled integration of text, graphics, drawing, static and moving images (video), audio animation, and any other media suitable for representing, storing, transmitting, and digitally processing every kind of information. A multimedia application is any application that employs a collection of multiple media sources (e.g., texts, graphics, images, audio, animation, and/or video). Hypermedia is one of the multimedia applications [35].

The *multimedia* term is used in contrast to the term *media,* which indicates that only conventional types of printed or hand-produced material are used. The term *multimedia* was introduced by Bob Goldstein at Southampton, Long Island, in July 1966. Some landmark events in the area of multimedia are [36]:

- **1945**: Vannevar Bush publishes the first landmark paper that describes what amounts to a hypermedia system called "*Memex*".
- **1960**: Ted Nelson coined the concept of "*hypertext*".
- **1969**: Development of the hypertext editor at Brown (Nelson and Van Dam).
- **1985**: Negroponte and Wiesner formed the MIT Media Laboratory.
- **1989**: Tim Berners-Lee proposal for the World Wide Web to CERN.

- **1990**: Kristine Hooper Woolsey headed the Apple Multimedia Laboratory (with more than 100 scientists).
- **1992**: *JPEG* was adopted as the international standard for digital image compression. Also, the first *M-Bone audio* multicast on the Net was made.
- **1993**: Production of the first full-fledged browser, *Mosaic* of NCSA (The Illinois National Center for Supercomputing Applications).
- **1994**: Development of *Netscape* by Jim Clark and Marc Andreessen.
- **1995**: Introduction of JAVA for platform-independent application development (The first applet created is called Duke).
- **1996**: Introduction of Microsoft's *Internet Explorer*.

Further historical information pertaining to multimedia can be found in [37].

**Telematics** Telematics is the synergy of telecommunications and informatics (*telematics* = **tele** communications + inform **matics**). Therefore, as an overall term, "*telematics*" refers to the long-distance transmission and computer processing of information. The term "telematics" was coined in French by Simon Nora and Alain Minc as "*telematique*" (**tele** communication et infor**matique**) in their 1978 report entitled "L' Informatisation de la Société" [38]. This report was requested by the president of France *Valery Giscard d' Estaing* in 1976 to explore how the computer and informatics applications could be extended to the organizational, economic, and developmental issues of modern society. In their report, Nora and Minc made the following remark: "Today, any consumer of electricity can instantly obtain the electric power needs without worrying about where it comes from or how much it costs. There is every reason to believe that the same will be true in the future of telematics." Today, computers become smaller and smaller with ever lower energy requirements, and computing devices gradually become mobile and can accompany us wherever we go. These developments have led to the so-called mobile computing, ubiquitous computing, or pervasive computing. The historic landmarks of *telematics* involve the parallel and combined landmarks of computers and telecommunications. On the computer side, the two major landmarks are as follows [39, 40]:

- **1971**: Development of the first microprocessor, the Intel 4004.
- **1981**: Release of Osborne 1, the first portable computer.

On the telecommunications side, we mention the following:

- **1860**: Invention of the telephone by Antonio Meucci.
- **1895**: Invention of the wireless telegraph by Guglielmo Marconi.
- **1946**: AT&T introduces the first mobile telephone.
- **1958**: Launching of the first communication satellite, SCORE.
- **1969**: The ARPANET goes online and links for the first time two computers (University of California and Stanford).
- **1992**: The WWW is released by CERN worldwide.
- **2000**: The first commercial UMTS network is introduced in Japan.

We now present a summarized description of the basic concepts and methods of various manifestations of information that have been discussed in this section from a historical point of view, namely,

- Communication systems.
- Information theory.
- Computers and computer science.
- Informatics.
- Telematics.
- Multimedia.
- Information systems.
- Information science.

## 4.4 Communication Systems

### *4.4.1 General Issues*

In communication systems theory, signals represent information. Information is used neatly packaged in analog or digital form. Communication systems are used for both manipulating and transmitting information. The following cases are the main types:

- *Point-to-point communication* (from a single place to another place).
- *Broadcast communication* (from one place to many other places).
- *Multipoint to multipoint communication* (teleconferencing, chat rooms).

Communication systems can be *analog* (e.g., via radio waves) *or digital* (e.g., computer networks). The question arising here is: Which is better, analog or digital communication? This question has been answered by Shannon, in his work on information theory [9], who suggests the use of digital representation of signals and the digital communication strategy. This means that all information-bearing signals are converted into digital ones (discrete-time, amplitude-quantized signals) applying the sampling theorem which shows the condition that must be met for this conversion to be an accurate one. As mentioned in Sect. 4.3.5. Shannon has shown that in digital form, a properly designed system can communicate digital information without error despite the presence of communication noise in all transmission channels. This result has a fundamental importance and value not only in communications, but also in all areas where digital information is handled, e.g., compact discs (CDs) and digital video disks (DVDs).

### 4.4.2 Shannon–Weaver Communication Model

One of the fundamental concepts in communication and information theory is the "communication model" of Shannon and Weaver [10], which is illustrated in Fig. 4.1b. Figure 4.1a shows the graphical representation (called *block diagram*) of a system operating on a receiving signal $x(t)$ and producing an overall output signal $y(t)$. In Fig. 4.1b, we have the following components (blocks):

- The *information source* that selects a desired signal (message) $s(t)$, out of a set of possible messages, which is to be sent to the destination (sink). The message can have several forms, e.g., speech, music, strings of letters as in telegraph or teletype, or characters typed on the keyboard of a PC, etc.
- The *transmitter* that receives the message $s(t)$ and produces a signal $x(t)$ suitable for transmission over the channel. The signal $x(t)$ is either modulated or encoded, depending on the message's physical nature. In telephony, the transmitter's operation consists of changing sound pressure into electrical current. In telegraphy, an encoding operation takes place that generates a sequence of dots, dashes, and spaces corresponding to the message. In a multiplexed, pulse-coded, modulation system, the various speech functions are sampled, compressed, quantized, and encoded, and finally properly interleaved to construct the signal $x(t)$.
- The *channel* is the medium over which the signal is transmitted to the receiver. In the channel, the transmitted signal is typically corrupted by noise or distorted or attenuated by various phenomena [giving the corrupted message $r(t)$].
- The *receiver* is a sort of inverse transmitter, changing the transmitted signal back into a received message $\hat{s}(t)$ that must resemble $s(t)$ as much as possible.
- The *destination* (or information sink) that is the person (or thing) for whom the message is intended (and who uses it for the desired purpose).

The "*noise*" contaminating the transmitted message $x(t)$ may be *internal* (i.e., coming from the receiver's attitudes, beliefs, or knowledge) or *external* (i.e., caused

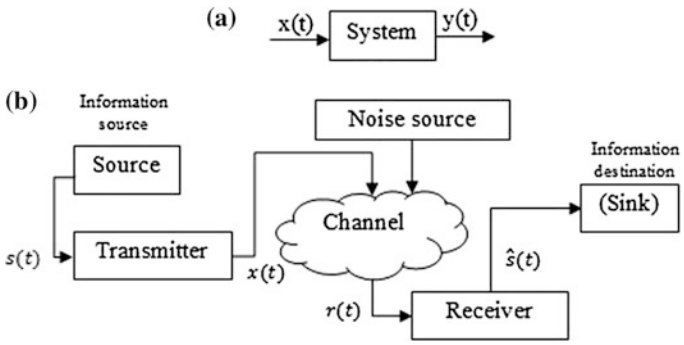



**Fig. 4.1** **a** Block diagram of system receiving an input signal $x(t)$ and producing an output signal $y(t)$, **b** Shannon–Weaver communication model

by other sources). This internal or external noise may either *strengthen* the intended outcome of the messages (if the information confirms the message) or *weaken* the intended outcome (the information in the noise disconfirms the original message).

The implementation of the *Shannon–Weaver communication model* needs the following:

- To understand signals and signal generation.
- To understand the nature of the information these signals represent.
- To know how information is transformed between analog and digital forms.
- To know how information can be processed by systems working on information-bearing signals.

These requirements are met through *electrical engineering* (which studies the ways signals are represented and manipulated electrically/electronically) and *signal engineering* (which studies the structure of signals, their information content, and the capabilities that this structure imposes upon communication systems), independently of what the signal sources are.

### 4.4.3 Other Communication Models

The Shannon–Weaver linear communication model is applicable to pure machine communication, i.e., it is not intended to match human communication. This model is focused on the technical problems associated with the selection and arrangement of discrete units of information and not with the semantic "content" of the messages. Weaver himself pointed out that: "*It is surprising but true that, from the present view point, two messages, one heavily loaded with meaning and the other pure nonsense, can be equivalent as regards information.*"

Two other communication models devised for human communication are the *Berlo model* of communication [43] and the *Schramm model* of communication [44], shown in Fig. 4.2a, b [45]:

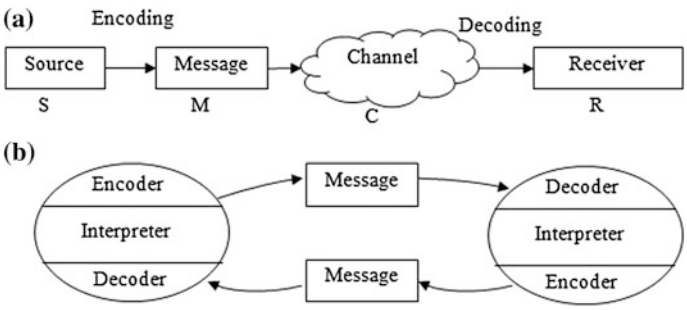



**Fig. 4.2** **a** Berlo's SMCR model: a source encodes a message for a channel to a receiver who decodes the message, **b** Schramm's communication model

Berlo's model is an enhanced adaptation of the Shannon–Weaver model so that, to cover the human communication, it provides a mechanism in the source for the following:

- Communication skills.
- Attitudes.
- Knowledge.
- Social and cultural issues.

In other words, the source is sufficiently flexible to include oral, written, electronic, or any other form of "symbolic" generator of messages. The "*message*" is considered to be the core element of the model, stressing the transmission of ideas, specifically content, elements, treatment, structure, and code. The receiver is recognized to be very important to communication, since it is actually the target. Like the source, the receiver takes into account and interprets communication skills, attitudes, knowledge, and social and cultural issues. The channel, i.e., the communication medium, involves all the human senses (hearing, seeing, touching, smelling, and tasting). The concepts of "encoding" and "decoding" emphasize the psychophysical problem every person has in translating his/her own thoughts in worlds or symbols and providing them in an understandable way to other people. Here, it is tacitly assumed that human communication is similar to machine communication (e.g., telephone signal, television signal, and computer information transmission, etc.). Clearly, the accuracy of human communication using this model depends on choosing the "proper" symbols, preventing interference, and sending efficient messages. However, even if the proper symbols are used, people may misunderstand each other. But all these issues are human-centered depending on agreements, beliefs, shared values, and attitudes.

The Schramm's model provides the additional features of "*feedback field expertise*" and "*psychological reference frame*" for the interacting humans. It is noted that Schramm's model is less linear than the Shannon–Weaver or the Berlo's model. But again, it is suitable only for bilateral communication between two partners (i.e., it does not cover the case of multilevel communication). Three nonlinear models of communication are the following:

- **Dance's helical spiral** (**1967**): This spiral depicts communication as a dynamic process, where the helix represents how communication evolves in a human from his/her birth up to the present [41].
- **Westley and MacLean's conceptual model** (**1957**): This model is based on the fact that the communication of a person begins when he/she starts to respond selectively to the immediate surroundings and not when he/she starts to talk [41].
- **Becker's mosaic model** (**1968**): Here, it is assumed that most communicative acts link message elements from several social situations (not just from one situation). These complex communicative events are linked to the activity of a receiver who moves through a permanently varying cube or mosaic of information [41]. This model adds a third dimension, but human communication

involves many more dimensions. Therefore, many researchers attempted to develop multidimensional communication models. Two such models are the following [41]:

- **Ruesch and Bateson functional model (1951)**: This involves four levels of analysis, i.e., *level 1* (intrapersonal process), *level 2* (interpersonal), *level 3* (group interaction), and *level 4* (cultural level).
- **Barnlund's transactional model (1970)**: A very systematic functional model where the key assumptions on which it is based are shown explicitly. Its most important property is that it includes no simple or linear directionality in the interplay between itself and the surrounding world.

### *4.4.4 Transmitter–Receiver Operations*

We now briefly discuss the operations of the *transmitter* and *receiver* in the Shannon–Weaver communication model (Fig. 4.1b).

**Modulation** The transmitter performs the *modulation* operation, namely, the superposition of the information (message) onto an electronic or optical carrier signal. Modulation can be applied to direct current (mainly by turning it on or off), alternating current, and optical signals. The *Morse code*, used in telegraphy and presently in amateur radio, uses a *binary* (i.e., two state) digital code similar to the code used in modern computers. In today's typical radio and telecommunication systems, the carrier is an alternating current (AC) sent over a given frequency band. Standard modulation methods are as follows:

- **AM** (**Amplitude modulation**), where the voltage superimposed on the carrier modulation signal is time varying.
- **FM** (**Frequency modulation**), where the frequency of the carrier waveform is varied (modulated) in suitably small amounts.
- **PM** (**Phase modulation**), where the natural flow of the alternating signal is delayed temporarily.

The above are called *continuous* (*analog*) *signal modulation* methods to discriminate them from **PCM** (**Pulse-Code Modulation**) which is employed to encode analog and digital signals in a binary (digital) way. In general, the modulation techniques can be classified as follows:

- Analog versus digital modulation.
- Baseband (low pass) versus bandpass (passband) modulation.
- Binary versus M-ary modulation.
- Linear versus nonlinear modulation.
- Memoryless modulation versus modulation with memory.
- Power-efficient versus bandwidth-efficient modulation.
- Constant envelope versus nonconstant envelope modulation.

- Equivalent digital modulation methods (amplitude-shift keying: ASK, frequency-shift keying: FSK, phase-shift keying: PSK). Two-way radios employ FM although some use *single sideband* (SSB). A combination of ASK and PSK is the *quadrature-amplitude modulation* (QAM).

**Demodulation** This is the inverse operation of modulation, which extracts the original information-bearing signal from the modulated carrier signal. An electronic device used to perform the recovery of the original message from the modulated carrier signal is called a *demodulator*. To each modulation technique, there corresponds an appropriate demodulation technique and an associated demodulator. For example, in an AM signal (which encodes the message into the carrier wave by varying its amplitude in proportion to the analog message sent), we have two kinds of demodulators. These are the *envelope detector* (which uses a *rectifier* allowing the current to flow only in one direction) and the *product detector*. The latter multiplies the incoming AM-modulated signal by the signal of a local oscillator with the same frequency and phase as the carrier used. After filtering, the original message is obtained. For FM, many typical forms of demodulators exist. For example, a *quadrature detector* phase shifts the signal 90° and multiplies it with the unshifted version. Among the terms produced by this operation is the original message which is selected and amplified. Another FM demodulator employs two AM demodulators, one tuned to the high end of the frequency band and the other to the lower end. The two outputs are then passed to a difference amplifier.

A pair of *transmitter* (coder, modulator) and *receiver* (decoder, demodulator) is called a *transceiver*. The general structure of modern communication systems involving a *CODEC* (Coder/Decoder) and *MODEM* (Modulator/Demodulator) has the form shown in Fig. 4.3.

**Multiplexing** To convey more information in a given amount of time, the bandwidth of a signal carrier is divided such that more than one modulated messages can be sent simultaneously on the same carrier. This is called *multiplexing*, where the carrier is sometimes called the *channel* and its separate message signal carried is referred to as *subchannel*. The multiplexer is the device that puts the individual signals onto the carrier and takes off received transmissions—separately. Typical forms of multiplexing are as follows:

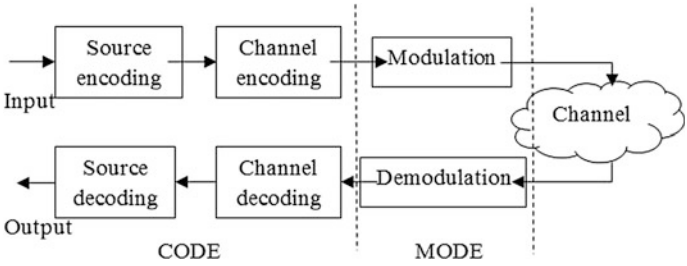



**Fig. 4.3** General structure of a digital communication system

**FDM**: Frequency-division multiplexing.
**TDM**: Time-division multiplexing.

FDM which divides the principal frequency of the carrier into separate sub-channels is commonly employed in analog communication systems, whereas TDM, which divides the principal signal into time slots, each one carrying a separate signal, is used in digital communication systems. A digital cellular phone technology based on *TDMA (time-division multiple access),* which was developed in the 1980s and is predominant in Europe, is the *Global System for Mobile (GSM)* communications. GSM is also used worldwide. GSM operates in the 900 MHz and 1.8 GHz bands in Europe, and the 1.9 GHz PCS band in the United States. It is based on a circuit-switched system, which divides each 200 kHz channel into eight 25 kHz time slots.

## *4.4.5 Analysis of Analog Modulation–Demodulation*

Here, we outline the mathematical analysis of analog modulation. The carrier wave is a sinusoid of the following form:

$$c(t) = A\cos(\omega_c + \phi)$$

where the *amplitude* A, *the carrier cyclic frequency* $\omega_c = 2\pi f_c$ ($f_c$ is the frequency), and the *phase* $\phi$ are the parameters that can be varied according to the message-bearing signal:

$$m(t) = M\cos(\omega_m t + \theta_0)$$

where $\theta_0$ is a constant and, without loss of generality, it can be assumed $\theta_0 = 0$.

### 4.4.5.1 Amplitude Modulation

In amplitude modulation (AM), the amplitude A of the carrier signal is modulated in proportion to the message-bearing (low frequency) signal m(t), to give

$$\begin{aligned} x(t) &= (A + M\cos(\omega_m t))\cos(\omega_c + \phi) \\ &= A(1 + \mu\cos\omega_m t)\cos(\omega_c t + \phi) \end{aligned}$$

where

$$\mu = M/A$$

is called the "*modulation index*" of AM, and for demodulation purposes (i.e., for the recovery of m(t) from the above modulated carrier signal) is typically selected less than one (i.e., M < A).

Using the well-known trigonometric identity:

$$\cos(\theta)\cos(\psi) = (1/2)[\cos(\theta+\psi)+\cos(\theta-\psi)],$$

the amplitude modulated signal $x(t)$ can be written as

$$x(t) = A\cos(\omega_c t+\phi) + \frac{A\mu}{2}[\cos(\omega_c+\omega_m)t+\phi] + \frac{A\mu}{2}\cos[(\omega_c-\omega_m)t+\phi]$$

This shows that $x(t)$ has three additive components with frequencies $\omega_c$, $\omega_c+\omega_m$ and $\omega_c-\omega_m$. The frequency $\omega_c+\omega_m$ is called the *upper side frequency*, and $\omega_c-\omega_m$ is called the *lower side frequency*. A typical AM signal with $\mu_0 = 0.8$ is shown in Fig. 4.4.

In the type of AM described above, both side frequencies $\omega_c-\omega_m$ and $\omega_c+\omega_m$ are transmitted. It is therefore called *Double Sideband AM* (DSB-AM). But for more efficiency, only one of the sidebands must be transmitted, in which case we have the so-called *Single Sideband AM* (SSB-AM). In particular, SSB-AM is called *Lower Sideband AM* (LSB-AM) or *Upper Sideband AM* (USB-AM), if only the lower frequency or the upper side frequency is transmitted, respectively. The case where the modulation index is μ > 1 is called overmodulation. Now, let us calculate the power of a modulated wave. The power is proportional to the square of the voltage (amplitude). Therefore, the power of the carrier wave is

$$\text{Carrier power} = KA^2$$

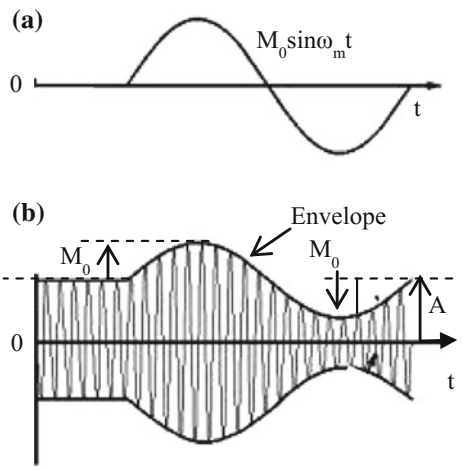



**Fig. 4.4** Typical amplitude signal with $\mu_0 = 0.8$

where the factor ½, required because $A$ is the amplitude (peak value) of $c(t)$, is included in the proportionality constant $K$.

The total sideband power is equal to

$$\text{Total side - band power} = KA^2\left(\frac{\mu}{2}\right)^2 + KA^2\left(\frac{\mu}{2}\right)^2 = \frac{\mu^2}{2} \times \text{Carrier power}$$

Therefore, the total power of the AM signal is equal to

$$P = KA^2\left(1 + \mu^2/2\right)$$

i.e., equal to the sum of the constant carrier power $KA^2$ and the power of the two sidebands, which depends on the value of $\mu$. For example, if $\mu = 1$, $P = KA^2 \times (3/2) = KA^2 \times 150\%$.

#### 4.4.5.2 Amplitude Demodulation

As mentioned in Sect. 4.4.4, for AM we have two kinds of demodulators (detectors): the *envelope demodulator* and the *product demodulator*. The signal

$$v(t) = A(1 + m(t))$$

is called the *envelope* of the AM signal. Clearly, if the envelope $v(t)$ is extracted, the transmitted message $m(t)$ can be recovered. A simple way to obtain the envelope is to use the "*envelope demodulator circuit*" shown in Fig. 4.5.

The envelope detector is of the so-called *noncoherent* (asynchronous) type. Better performance can be obtained by *coherent* (synchronous) detector in which both the frequency and the phase of the carrier are known at the demodulator. The amplitude of the carrier affects only the level of the demodulated signal, which can be changed by a simple amplifier. Therefore, the amplitude of the carrier is not important in the demodulation. The carrier signal is restored at the receiver by a circuit called the *carrier-recovery circuit*. In the *product detector*, the amplitude

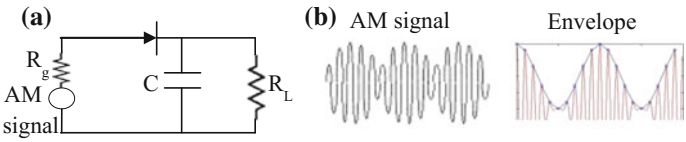



**Fig. 4.5** **a** Envelope AM demodulator. **b** The AM signal passing through the demodulator provides the envelope signal. This is done by the capacitor which removes the carrier frequency and leaves only the envelope

modulated signal is multiplied by the signal provided by a local oscillator. The simplest case is obtained if the local oscillator has the same frequency as the carrier, i.e., $\omega_c$. Then, the result of the multiplication is the sum of the original message and another AM signal at frequency $2\omega_c$. A low-pass filter is used to suppress this second component. The block diagram of the product detector is shown in Fig. 4.6.

In the above block diagram, we have

$$\begin{aligned} x(t)\,c(t) &= A(1+m(t))\cos(\omega_c t)\cos(\omega_c t) \\ &= \frac{1}{2}A(1+m(t))(1+\cos(2\omega_c t)) \end{aligned}$$

where, without loss of generality, the phase $\phi$ of the carrier was omitted and use was made of the same trigonometric identity as in the AM modulation. After low-pass filtering, the original message is recovered.

The carrier signal can be available to the product detector in two ways:

- Through transmission of the carrier.
- Through recovering of the carrier.

The carrier recovery can be achieved by using a transmitted pilot signal outside the passband of the modulated signal as shown in Fig. 4.7.

To recover the carrier signal, two techniques can be applied: (i) recovery by a bandpass filter (Fig. 4.8a), and (ii) recovery by a phase-locked loop—PLL (Fig. 4.8b).

It is noted that, with the product demodulator, we can also decode overmodulated AM, SSB, and AM with suppressed carrier, in addition to the standard DSB-AM.

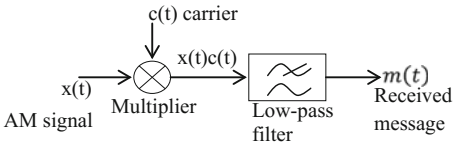



**Fig. 4.6** Block diagram of a product detector

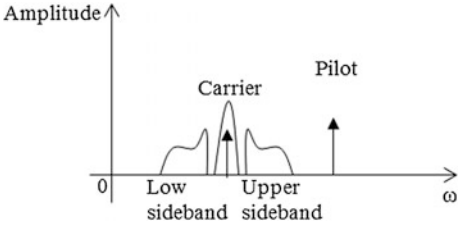



**Fig. 4.7** Pilot signal outside the AM signal

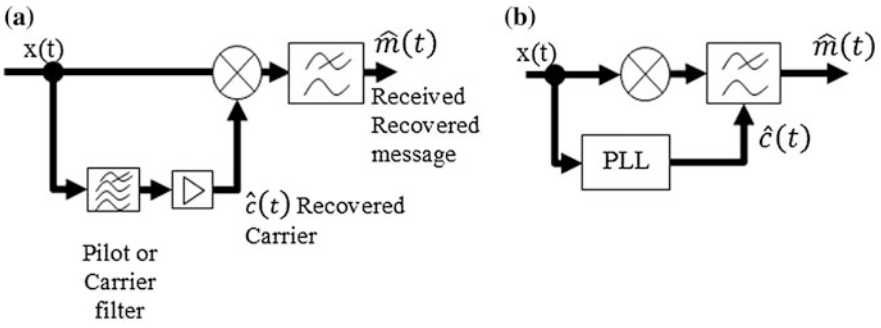



**Fig. 4.8** **a** Recovery of carrier by a bandpass filter. **b** Recovery of carrier by a phase-locked loop (PLL). Here, we use the symbols $\hat{c}(t)$ and $\hat{m}(t)$ to indicate approximate copies of $c(t)$ and $m(t)$

#### 4.4.5.3 Frequency Modulation

The principal advantages of *frequency modulation* (FM) over AM are as follows: (i) better signal-to-noise ratio, (ii) less interference effects between neighboring stations, and (iii) less radiated power. The drawbacks are as follows: (i) It requires much more bandwidth than AM (up to 20-times more), and (ii) The transmitter and receiver devices are more complex. In FM, the frequency of the carrier wave is changed in proportion to the message signal $m(t)$, i.e., we have

$$\omega(t) = \omega_c(1 + \mu\cos(\omega_m t))$$

where $f_c = \omega_c / 2\pi$ is called the *center frequency*, and $\mu$ is the degree of modulation. A frequency-modulated signal has the form shown in Fig. 4.9b.

Frequency modulation is popular in a variety of radio transmission applications from broadcasting to general point-to-point transmissions. It is also widely employed for broadcasting via very high frequency (VHF) because it provides a very good medium for high-quality transmissions. FM is also largely used in

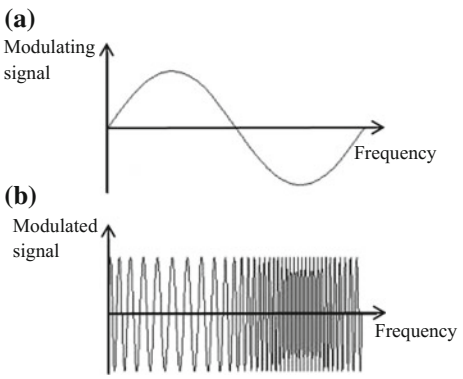



**Fig. 4.9** Frequency modulation. **a** The modulating signal. **b** The modulated signal

several mobile applications. Here, because $\omega$ is time varying, the FM signal is given by

$$A\cos\left[\int_0^t \omega(t)\,\mathrm{d}t\right] = A\cos\left[\int \omega_c(1+\mu\cos(\omega_m t))\mathrm{d}t\right]$$
$$= A\cos\left[\frac{\omega_c\mu}{\omega_m}\sin(\omega_m t)+\omega_c t+\theta_0\right]$$

where the constant of integration $\theta_0$ can be neglected as a constant angle. The quantity $\Delta\omega = \omega_c\mu$ is called the *cyclic frequency deviation*.

The parameter

$$\frac{\omega_c\mu}{\omega_m} = \frac{\Delta\omega}{\omega_m} = \mu_f$$

is called the *FM modulation index* (or *deviation ratio*) and has a value in radians that is different for each value of $\omega_m$. On the basis of this, the expression for an FM signal becomes

$$x(t) = A\cos\big(\omega_c t+\mu_f\sin(\omega_m t)\big)$$

Expanding this signal using and the identity for $\cos(\theta+\psi)$, we obtain

$$\begin{aligned} x(t) &= A\cos\big(\omega_c t+\mu_f\sin(\omega_m t)\big)\\ &= \cos(\omega_c t)\cos\big(\mu_f\sin(\omega_m t)\big) - \sin(\omega_c t)\sin\big(\mu_f\sin(\omega_m t)\big)\end{aligned}$$

The two complex functions $\cos\big(\mu_f\sin(\omega_m t)\big)$ and $\sin\big(\mu_f\sin(\omega_m t)\big)$ can be expressed as an infinite series of Bessel functions, namely [46, 47],

$$\begin{aligned}\cos\big(\mu_f\sin(\omega_m t)\big) &= J_o\big(\mu_f\big)+2J_2\big(\mu_f\big)\cos(2\omega_m t)\\ &\quad +2J_4\big(\mu_f\big)\cos(4\omega_m t)+2J_6\big(\mu_f\big)\cos(6\omega_m t)+\cdots\\ \sin\big(\mu_f\sin(\omega_m t)\big) &= 2J_1\big(\mu_f\big)\sin(\omega_m t)+2J_3\big(\mu_f\big)\sin(3\omega_m t)\\ &\quad +2J_5\big(\mu_f\big)\sin(5\omega_m t)+2J_7\big(\mu_f\big)\sin(7\omega_m t)+\cdots\end{aligned}$$

where the Bessel functions $J_k$ $(i=1,2,\ldots)$ of the first kind are defined as

$$J_k\big(\mu_f\big) = \sum_{q=0}^{\infty}\frac{(-1)^q}{q!(q+k)!}\left(\frac{\mu_f}{2}\right)^{k+2q}$$

and are provided in corresponding mathematical tables and computer-mathematics libraries. Using the trigonometric identities for $\cos(\theta)\cos(\psi)$ and $\sin(\theta)\sin(\psi)$, we get

$$\begin{aligned} x(t) = A\{&J_0(\mu_f)\cos(\omega_c t) - J_1(\mu_f)[\cos(\omega_c - w_m t) - \cos(\omega_c + \omega_m)t] \\ &+ J_2(\mu_f)[\cos((\omega_c - 2\omega_m)t) + \cos((\omega_c + 2\omega_m)t)] \\ &+ J_3(\mu_f)[\cos((\omega_c - 3\omega_m)t) - \cos((\omega_c + 3\omega_m)t)] \\ &+ J_4(\mu_f)[\cos((\omega_c - 4\omega_m)t) + \cos((\omega_c + 4\omega_m)t)] - \cdots \end{aligned}$$

We see that the FM signal involves the center frequency $\omega_c$ and an infinite number of side frequency pairs, each pair spaced by an amount equal to the modulating frequency $\omega_m$.

The amplitude of the successive side frequency pairs is determined by the Bessel function coefficients. Although theoretically the FM signal covers the entire frequency spectrum with sidebands, the J coefficients decrease relatively quickly and the series converges rapidly. So, the bandwidth actually required is finite. Because $\omega_m$ and $\mu_f$ are inversely proportional $(\mu_f = \Delta\omega / \omega_m)$, a small $\omega_m$ will result in more side frequencies of significant value, than those obtained in the case of a high $\omega_m$ as shown in Fig. 4.10.

It is noted that $J_0(\mu_f) = 0$ for $\mu_f = 2.405,\ 5.520,\ 8.654,\ 11.79,\ 14.93$, etc. Therefore, for these values of $\mu_f$, the center frequency component is zero (i.e., it does not exist).

A practical rule of thumb is *Carson's rule* which says that almost all ($\sim 98\%$) of the power of an FM signal is contained in a bandwidth $B_c$ equal to

$$B_c = 2(f_m + \Delta f)$$

where $\Delta f = \mu f_c$ is the frequency deviation $\Delta\omega / 2\pi$ from the central frequency.

#### 4.4.5.4 Frequency Demodulation

In FM, it happens that signals with a large frequency deviation are supporting higher quality transmissions at the expense of occupying a wider bandwidth. Thus,

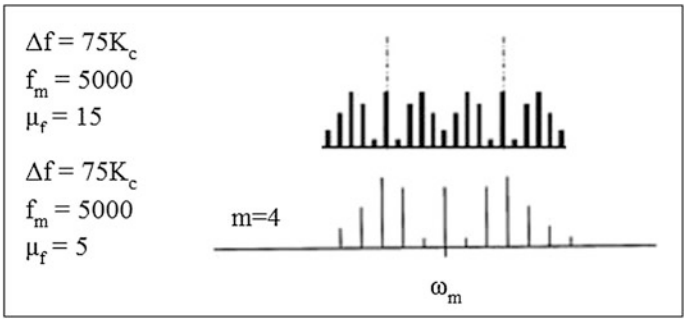



**Fig. 4.10** Frequency spectra of two FM signals with $\Delta f = 75\mathrm{K_c/s}$

in practice, several levels of frequency deviations (or modulation index values) are employed, according to the application that is used. Those with low deviation are known as *narrowband frequency modulation* (NBFM), with typical values ±3 kHz. NBFM is generally used for point-to-point communications. Higher frequency deviations are required for broadcasting and are called *wideband frequency modulation* (WBFM). Typical values of WBFM frequency deviation are ±75 kHz.

To receive FM, a receiver must be sensitive to the frequency variations of the incoming signal, which may be either NBFM or WBFM, and insensitive to amplitude variations. To assure that any amplitude variations are removed, the signals are amplified to such a degree that the amplifier goes into limiting. The demodulator must be frequency-dependent in order to be able to linearly convert frequency variations into voltage variations. This suggests that the FM demodulator must have an S-type voltage–frequency characteristic as shown in Fig. 4.11.

The principal types of FM demodulator circuits are as follows [48]:

- Slope detector.
- Ratio detector.
- Foster–Seeley detector.
- Phase-locked loop (PLL) detector.
- Quadrature detector.

**Slope detector** This is the simplest form of FM detector. It is essentially a tank circuit that is tuned to a frequency lightly higher or lower than the carrier frequency.

**Ratio detector** This detector is very popular because it provides a higher level of amplitude variation rejection. This has the result of a greater level of noise immunity (since most noise is amplitude noise), and a satisfactory operation with lower levels of limiting circuitry is required. The ratio detector consists of a frequency-sensitive, phase-shift circuit with a transformer and two diodes in series. The transformer performs the detection of the frequency variations of the incoming signal.

**Foster–Seeley detector** This detector, also called a *phase-shift detector*, was invented in 1936 by Dudley Foster and Stuart Seeley [49]. It employs a double-tuned, radio frequency transformer to convert frequency variations of its

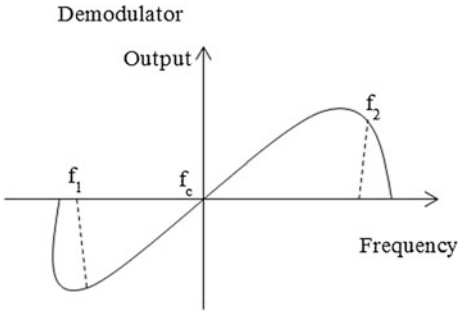



**Fig. 4.11** S-curve characteristic of an FM demodulator

input signal to amplitude variations. These amplitude variations are then rectified and filtered to give a dc output voltage that varies in both amplitude and polarity as the frequency of the input signal varies. The output voltage is zero when the input frequency is equal to the carrier frequency $f_c$. A typical response curve of the Foster–Seeley detector has the S-form shown in Fig. 4.11, where the frequency band of operation is $[f_1, f_2]$.

**Phase-locked-loop detector (PLL)** This detector produces a fixed (locked) relation to the phase of a "reference" input signal and responds to both the frequency and the phase of the input signals automatically, so that it matches the reference in both frequency and phase. Actually, a PLL detector is an example of a control system working under negative feedback. PLL FM detectors are very popular and used in several types of radio equipment ranging from broadcasting receivers to high-performance communications equipment. Their wide use started when integrated circuit technology had developed to allow the manufacture of radio frequency analog circuits. The basic structure of a PLL detector is shown in Fig. 4.12 [42].

The phase comparator compares the phase of the output signal with that of the input signal and sends the phase-error signal to the loop filter, where it is filtered and becomes the control signal $u(t)$ that drives the *voltage-controlled oscillator* (VCO). The key issue in designing a PLL detector is the loop filter, which must have sufficiently wide bandwidth to allow it to follow the anticipated variations of the frequency-modulated signal.

**Quadrature detector** The quadrature detector shifts the incoming signal, 90° and multiplies it with the unshifted signal. It does not need a center-tapped transformer and so it can easily be integrated into a single LSI chip, unlike the other detectors. The 90°-phase shift is achieved using a high-reactance capacitor and passes the phase-shifted signal to an LC circuit tuned at the carrier frequency. Then, the frequency changes produce an additional leading or lagging phase shift within the multiplier. The quadrature FM detector circuit is shown in Fig. 4.13. The operation of the quadrature FM detector is based on the property that multiplying two periodic signals with the same frequency generates a DC voltage that is directly proportional to the signal-phase difference.

At the resonance frequency, the phase of the LC circuit is zero, below resonance the phase is positive and above resonance the phase is negative, i.e., the phase changes with the variations of the frequency, and the same is also true for the output voltage.

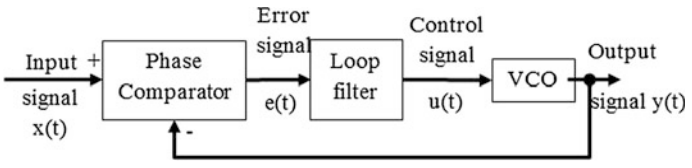



**Fig. 4.12** Basic structure of a PLL detector

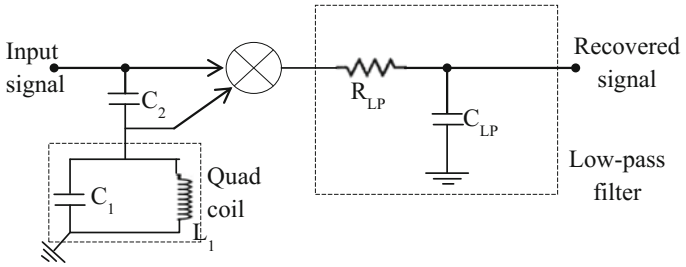



**Fig. 4.13** Quadrature detector circuit (The capacitor $C_2$ shifts the phase of the input signal 90°)

#### 4.4.5.5 Phase Modulation

For a carrier $c(t) = A\cos(\omega_c t + \phi_c)$ and a modulating signal $m(t) = \mu\cos(\omega_m t)$, the phase modulated (PM) signal $x(t)$ has the following form:

$$x(t) = A\cos(\omega_c t + \psi(t)),\ \psi(t) = m(t) + \phi_c$$

Here, we directly see that, as $m(t)$ increases or decreases over time, so does the phase shift of the modulated signal. The phase shift can also be viewed as a change of the carrier frequency. Therefore, phase modulation is equivalent to frequency modulation, and so it can be analyzed by the methods presented in Sect. 4.4.5.3. We note again that, for small amplitude-modulating signals, we have the undesired result of sidebands and poor efficiency. For very large, single, sinusoid modulating signals, the bandwidth of PM is approximately given by Carson's rule as in FM, i.e.,

$$B_C = 2(1 + \mu_{PM})f_m$$

where $\mu_{PM}$ is the *PM modulation index* defined as $\mu_{PM} = \Delta\phi$ with $\Delta\phi$ being the maximum phase shift. An example of PM signal is shown in Fig. 4.14.

#### 4.4.5.6 Phase Demodulation

Because of the equivalence of FM and PM (the carrier frequency is modulated by the time derivative of the phase shift), phase demodulation can be performed by any FM demodulator. Here, we will outline a phase demodulator of the PLL type, called a "*sinusoidal phase detector*", which is shown in Fig. 4.15 [42].

In the phase-locked loop detector of Fig. 4.13, the VCO tends to phase lock to the input in "*quadrature*", i.e., with 90°-phase difference $(\phi(t) \to \psi(t) + \pi/2)$. This means that we can define $\phi(t)$ as $\phi(t) = \theta(t) + \pi/2$ with $\theta(t) \to \psi(t)$ as $t \to \infty$.

In Fig. 4.18, the output $e(t)$ of the multiplier is equal to

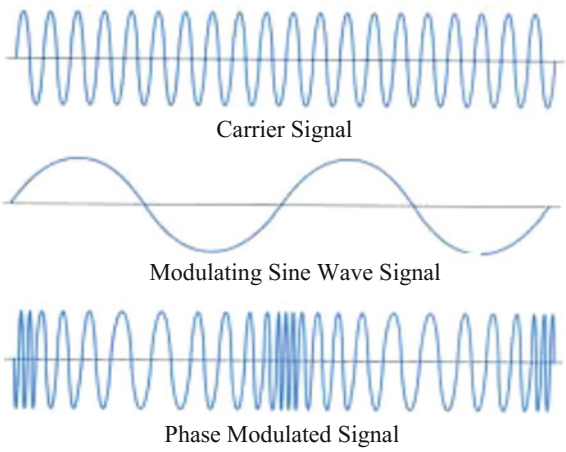



**Fig. 4.14** Typical PM signal

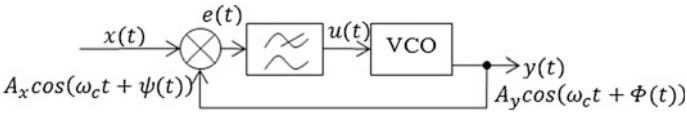



**Fig. 4.15** Sinusoidal phase detector

$$\begin{aligned} e(t) = x(t)y(t) &= \left(A_x A_y/2\right)[\cos(\psi(t) - \phi(t)) \\ &\quad + \cos(2\omega_c + \psi(t) + \phi(t))] \\ &= \left(A_x A_y/2\right)[\cos(\psi(t) - \theta(t) - \pi/2) \\ &\quad + \cos(2\omega_c + \psi(t) + \theta(t) + \pi/2)] \\ &= \left(A_x A_y/2\right)[\sin(\psi(t) - \theta(t)) - \sin(2\omega_c t + \psi(t) + \theta(t))] \end{aligned}$$

Now, assuming that the low-pass filter removes the second (high frequency) term, the output $u(t)$ of the filter is found to be

$$u(t) = \left(A_x A_y/2\right)\sin(\psi(t) - \theta(t))$$

where $\theta(t)$ is considered to be the output phase. Here, instead of a sawtooth function $F(\cdot)$, we have the sinusoid function:

$$F(p) = \left(A_x A_y/2\right)\sin(p(t)),\ p(t) = \psi(t) - \theta(t)$$

which has the plot shown in Fig. 4.16.

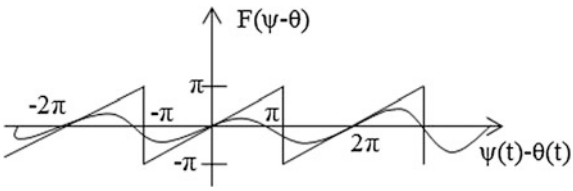



**Fig. 4.16** The sinusoidal gain function F(·). For comparison, the sawtooth function is also shown

Here, $\theta(t) \rightarrow \psi(t)$, and so the phase $\phi(t)$ of the detector tends to lock in quadrature with the input. Figure 4.17 shows in the same plot the AM, FM, and PM signals obtained using the same modulating signal and the same carrier in all cases. This plot can be generated by using the proper signal generation (SigGen) modules available in [50].

### 4.4.6 Pulse Modulation and Demodulation

**General Issues** *Analog modulation* (AM, FM, PM) is used for transferring an analog low-pass (baseband) signal, such as an audio signal or TV signal, over an analog passband channel, e.g., a restricted radio frequency band or a cable TV network channel. *Digital modulation* is used for transferring a digital bit stream over an analog passband channel, such as a limited radio frequency band or a public switched telephone network (where a bandpass filter restricts the frequency band to between 300 and 3400 Hz).

*Pulse modulation* is used to transfer a narrowband analog signal, such as a phone call, over a wideband baseband channel, or in some cases, as a bit stream over another digital transmission system. It is also used in neuron modeling and circuit design. Analog modulation/demodulation techniques were discussed in previous sections. Here, we will briefly discuss the basic pulse modulation schemes. Digital modulation schemes will be presented in the next section. Pulse modulation

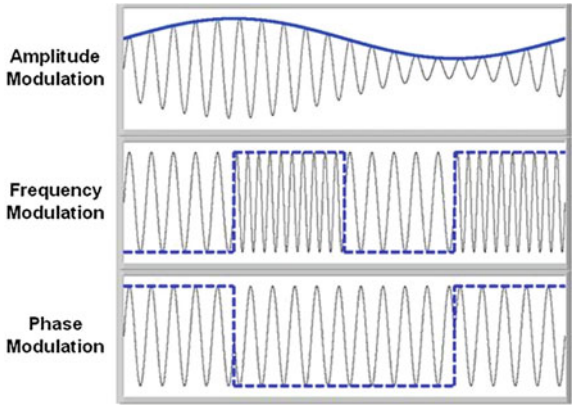



**Fig. 4.17** AM, FM, and PM signals compared. The carrier and modulating signals are shown superimposed on the top

schemes use as a carrier signal a pulse train. The form of the pulses can be selected from among several types that differ in terms of energy and spectral content consumption. Examples of pulse types are square pulses and raised-cosine pulses. The five basic pulse modulation methods are the following, according to the pulse train parameter, that is, modulated (amplitude, width/duration, frequency, and position of leading edge):

- Pulse-amplitude modulation (PAM).
- Pulse-width modulation (PWM).
- Pulse-frequency modulation (PFM).
- Pulse-position modulation (PPM).
- Pulse-code modulation (PCM).

**Pulse-Amplitude Modulation and Demodulation** In pulse-amplitude modulation (PAM), the amplitude (height) of individual pulses in the pulse train is varied from its normal value in accordance with the instantaneous amplitude of the modulating signal at sampling intervals. The width, frequency, and position of the pulses are kept constant. In other words, the information carried by the modulating signal is carried on a train of pulses being encoded in the amplitude of the pulses.

Figure 4.18 shows a typical PAM signal.

The PAM transmitter design is simple since it is a standard sampler with constant sampling period. Similarly, the receiver (demodulator) is simple.

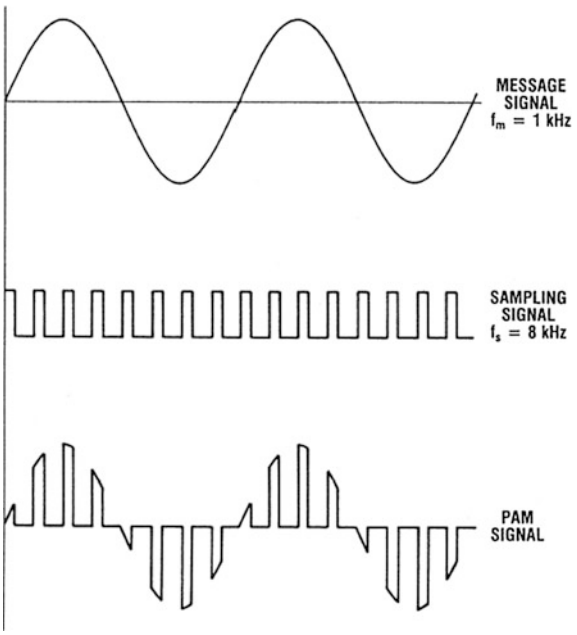



**Fig. 4.18** Example of PAM in which the amplitude of a square pulse carrier train is modulated by a sinusoid signal

**Pulse-Width Modulation and Demodulation** In *pulse-width modulation* (PWM) *or pulse-duration modulation* (PDM) or *pulse-length modulation* (PLM), the modulating signal changes the width of individual pulses from their normal values, in proportion to the change at each sampling interval. The amplitude, frequency, and position of the pulses are kept constant. PWM is extensively used in power electronics and power control applications, because it is a very good method of providing intermediate quantities of electric power between *fully on* and *fully off*. It is noted that a standard power switch provides full power when switched on, and zero power when switched off. In communication applications, the width of the pulses corresponds to specific data values encoded at one end and decoded at the other end. Pulses of various widths (lengths) that represent the transmitted information are sent at regular intervals determined by the carrier frequency. There is no need to use a clock signal, because the leading edge of the pulse train plays the role of a clock. But, to avoid a data value with a zero-width pulse, the addition of a small leading edge offset is required.

**Pulse-Frequency Modulation and Demodulation** In *pulse-frequency modulation* (PFM), the instantaneous amplitude of the modulating signal changes the frequency of the carrier-pulse train, leaving unaltered the amplitude and width of the pulses. PFM is analogous to PWM since the magnitude of the information-bearing (modulating) signal is encoded in the duty cycle of the square pulse train. Compared to PAM, PFM has the advantage of better immunity to noise interference. The drawback is that the design of transmitter and receiver is more complex. For this reason, in practice, PAM or PWM is more commonly used. An example of PFM is shown in Fig. 4.19.

**Pulse-Position Modulation** In *pulse-position modulation* (PPM), the variation of the instantaneously sampled values of the modulating signal changes the position of each pulse with respect to the position of a periodic reference pulse. The amplitude and width of the pulses are kept constant, and so the required transmitter power is constant. PPM has the drawback that it depends on the synchronization of the transmitter–receiver pair. PPM is widely used in optical communications. PPM and FSK (frequency-shift keying) modulation are two canonical forms of orthogonal modulation. PPM and FSK are actually dual modulation techniques in the sense that

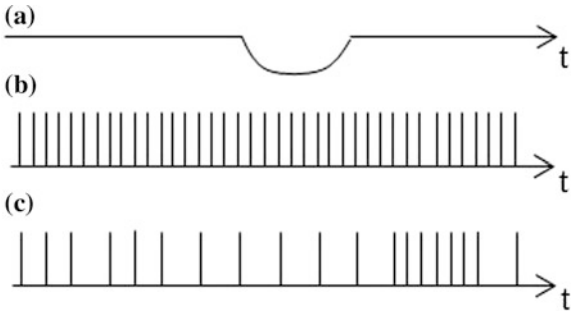



**Fig. 4.19** Examples of PFM: **a** modulating signal, **b** carrier periodic pulse train, and **c** PFM pulse train

in FSK each signal gets a different slice of the frequency band, whereas in PPM each signal gets a different slice of the signaling interval. A PPM signal can be produced by feeding a PWM signal into a differentiating circuit, which provides positive-and-negative-polarity pulses at the rising and falling edges of the PWM pulses. Passing these alternating polarity pulses to a rectification circuit, which cuts the positive fixed (non-modulated) pulses, we obtain the negative pulse train which is the desired PPM pulse signal.

**Pulse-Code Modulation and Demodulation** In *pulse-code modulation* (**PCM**), the amplitude of the analog modulating signal is sampled with a fixed sampling rate, and then it is quantized to a set of symbols, typically a binary code. The PCM principle is illustrated in Fig. 4.20.

The four-bit coded values of the sinusoid modulating signals at the sampling instants one up to 12 are as shown in Table 4.1.

The general structure of a PCM modulator has the form shown in Fig. 4.21.

This structure is implemented in several ways, using suitable, single-integrated circuits (known as analog-to-digital (A/D) converters).

*Pulse-coded demodulation* reproduces the analog input (message) from the digital output using a reverse sequence of operations. A PCM demodulator circuit is known as digital-to-analog (D/A) converter (see Fig. 4.22).

Pulse-code modulation has two important features:

- Noise contamination is almost completely eliminated when the pulse signals exceed the noise levels by at least 20 dB.
- The signal can be received and retransmitted as many times as desired with no distortion of the signal.

**Other Pulse Modulation Techniques** Besides the five types of pulse modulation just discussed, over the years, several other techniques with associated benefits and drawbacks have been developed. These are as follows [51]:

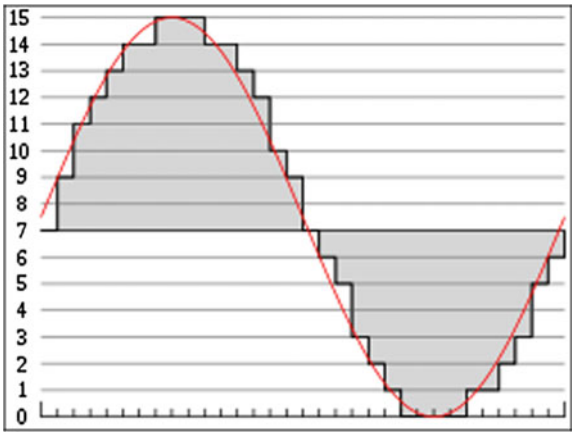



**Fig. 4.20** The principle of PCM. Sampling and binary coding using 16 quantized levels (i.e., four-bit quantization)

**Table 4.1** PCM of a sinusoid using the four-bit binary code 1-2-4-8

| Sampling instant | Sampled value | Binary coded value |
|---|---|---|
| 1 | 9 | 1001 |
| 2 | 11 | 1011 |
| 3 | 12 | 1100 |
| 4 | 13 | 1101 |
| 5 | 14 | 1110 |
| 6 | 14 | 1110 |
| 7 | 15 | 1111 |
| 8 | 15 | 1111 |
| 9 | 15 | 1111 |
| 10 | 14 | 1110 |
| 11 | 14 | 1110 |
| 12 | 13 | 1101 |

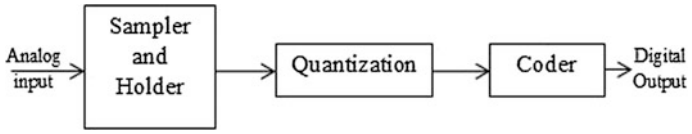



**Fig. 4.21** The pulse-code modulator performs the operations: sampling–holding, quantization, and binary (digital) coding

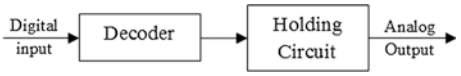



**Fig. 4.22** The PCM demodulator (or D/A converter) is performed by a binary decoder and an analog holder

- **ASK**: *Amplitude-Shift Keying*, where a finite number of amplitudes are used.
- **FSK**: *Frequency-Shift Keying*, where a finite number of frequencies are used.
- **PSK**: *Phase-Shift Keying*, where a finite number of phases are used.
- **BPSK**: *Binary-Phase Shift Keying*.
- **QPSK**: *Quadrature-Phase Shift Keying*.
- **QAM**: *Quadrature-Amplitude Modulation*.
- **ADPCM**: *Adaptive-Differential PCM*.
- **PDM**: *Pulse-Density Modulation*.

Some more specific schemes of quantized modulation (QM) are the following:

- **QAM**: *Quantized-Amplitude. Modulation*
- **QFM**: *Quantized-Frequency Modulation.*
- **QPAM**: *Quantized-Pulse-Amplitude Modulation.*
- **QPM**: *Quantized-Phase Modulation.*

- **QPPM**: *Quantized-Pulse-Position Modulation.*

For descriptions and details about them, the reader is referred to modern textbooks of communication systems [42, 52].

## 4.5 Information Theory

### *4.5.1 General Issues*

As we saw in Sect. 4.3.5, information theory was formally initiated by Claude Shannon (1948) who coined the concept of "*entropy*" as the average "*information content*" in a message, measured in bits (binary digits) [9]. The term "*information*" was originally coined by Ralph Hartley in his 1928 paper "*Transmission of Information*", where he treated "*information*" in a technical sense as a measurable quantity, reflecting the receiver's ability to recognize the sequence of symbols sent by the sender (without any concern about the meaning or semantic issues of these symbols). Hartley's formula for information is [8]

$$I = \log S^n = n \log S$$

where $S$ is the number of possible symbols and $n$ is the number of symbols in a transmission. The information I is measured in *decimal digits,* also called *Hartley units*.

Information theory is a mathematical theory principally concerned with coding–decoding. Its primary mathematical tools are probability theory and statistics. Shannon's formula for entropy is

$$H(X) = -\sum_{k=1}^{n} p_k \log_2 p_k$$

where $p_k = p(x_k)$ are discrete probabilities of the random process (message) $X$, with possible values $x_1, x_2, \ldots, x_n$, which express the probabilities that a particular message is transmitted. $H(X)$ is a measure of how much information is contained in the transmitted message.

Shannon gave the name "*entropy*" to the quantity $H(X)$ upon the suggestion of John von Neumann. Originally, he thought to call it "information" but since this word was overly used, he decided to call it "uncertainty". But, in a private conversation with von Neumann, he was advised to call it "*entropy*". "You should call it entropy, John von Neumann suggested, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, nobody knows what entropy really is, so in a debate you will always have the advantage."

In the following, we will present two derivations of the concept of entropy, $H(X)$, and the four fundamental theorems of information theory, namely,

- Nyquist–Shannon sampling theorem.
- Shannon's source coding theorem.
- Shannon's channel capacity theorem.
- Landau–Pollack bandwidth signal dimensionality theorem.

## 4.5.2 *Information Theory's Entropy*

The following two ways of introducing information theory's entropy will be discussed:

- Entropy as the information content of a measurement.
- Direct definition of entropy.

### 4.5.2.1 Entropy Derivation via the Information Content of a Measurement

The present derivation is as follows [53]. A real number $N$ is typically expressed by the infinite series:

$$N = d_{-n}10^n + \cdots + d_{-2}10^2 + d_{-1}10 + d_0 + d_1 10^{-1} + \cdots + d_n 10^{-n} + \cdots$$

where the $d_i$'s (digits) are integers between 0 and 9, and $d_{-n} \neq 0$ for $n > 0$. In shorthand, $N$ is written in the well-known decimal form:

$$N = d_{-n}d_{-n+1}\cdots d_{-1}d_0{\bullet}d_1 d_2 \cdots d_n$$

For each number, there exists a unique decimal expansion of this form, provided that we agree to exclude expansions like 2.36999… where nines repeat indefinitely, and express numbers like 2.5000… as 2.5 (omitting the infinitely repeating zeros).

Rational numbers are defined to be ratios of integers, i.e.,

$$a/b = d_n \cdots d_1 d_0{\bullet}d_1 d_2 \cdots d_n \quad (b > 0)$$

A decimal expression represents a fraction if and only if some sequence of digits in the expression repeats indefinitely. For example,

$$1/4 = 0.2500, \ldots, \; 1/6 = 0.166\ldots6\ldots$$
$$13/70 = 0.1857142857142857142857\ldots$$

With other numbers, the situation is more complex. For example, the decimal expressions for the numbers $\sqrt{2}$ and $\pi$ (the ratio of the circumference of a circle to its diameter) are as follows:

$$\sqrt{2} = 1.41421356\ldots,\ \pi = 3.14159265\ldots$$

where all digits must be calculated one by one. These classes of numbers differ in the amount of information they convey and in the effort required to obtain that information.

Although in some cases (like the ones mentioned above) we know explicitly all the digits of the expansion of a number, this is not so when we observe or measure a certain physical quantity. What one can do here is to obtain successive refined intervals that bound the actual value $m$ of the quantity $M$ at hand. Suppose that we get an estimate of the true value $m$ in the interval $\hat{m}_1 \in (a_1, b_1)$. If this is not satisfying, a second more precise measurement is performed which gives

$$\hat{m}_2 \in (a_2, b_2) \text{ with } a_1 < a_2 < b_2 < b_1$$

Clearly, the new bounding interval $(a_2, b_2)$ provides some additional number of digits in the decimal expansion of $m$. The number of decimal digits in the original decimal expansion is approximately given by $\log_{10}(b_1 - a_1)$, and the number of digits of the more accurate expansion is $\log_{10}(b_2 - a_2)$.

The gain $I_{10}$ in information when we go from the first interval $(a_1, b_1)$ to the second interval $(a_2, b_2)$ is equal to

$$\begin{aligned} I_{10} =& \log_{10}(b_1 - a_1) - \log_{10}(b_2 - a_2) \\ =& \log_{10}[(b_1 - a_1)/(b_2 - a_2)] \quad \text{decimal digits} \end{aligned}$$

If the number $N$ is represented in binary form: $N = c_{-n} \cdots c_{-1} c_0 \cdot c_1 c_2 \cdots c_n$ where $c_i = 0$ or 1 (for all $i$), the information gain is equal to

$$I_2 = \log_2[(b_1 - a_1)/(b_2 - a_2)] \text{ bits}$$

Since $b_1 - a_1$ is the length of the interval with limits $a_1$ and $b_1$, and $b_2 - a_2$ is the corresponding length of the second interval with limits $a_2$ and $b_2$, the ratio $(b_2 - a_2)/(b_1 - a_1)$ represents the probability $p$ that a random point in the interval $(a_1, b_1)$ lies in the interval $(a_2, b_2)$, i.e.,

$$p = (b_2 - a_2)/(b_1 - a_1)$$

Thus, the gain in information can be expressed as

$$I_2 = \log_2[(b_1 - a_1)/(b_2 - a_2)] = \log_2\left(\frac{1}{p}\right), \quad 0 \le p \le 1$$

Now, consider a system that can be in one of the finite sets of states $x_1, x_2, \ldots, x_n$ (arbitrarily ordered), and let $p_k$ be the probability of the system to be in state $x_k$. Clearly, the probabilities $p_k$ must satisfy the total probability condition:

$$\sum_{k=1}^{n} p_k = 1$$

Suppose that an observation of the system reveals that the system is in state $x_k$. To calculate how much information was gained after this observation, we assume that our observation system associates with each state a probability interval. For example, state $x_1$ with probability $p_1$ corresponds to the interval $0 < x < p_1$, state $x_2$ with the interval $p_1 < x < p_1 + p_2$, and so on. In general, the $k$th state $x_k$ is associated with the interval:

$$p_1 + p_2 + \cdots + p_{k-1} < x < p_1 + p_2 + \cdots + p_k$$

with limiting condition for the nth state of

$$p_1 + p_2 + \cdots + p_{n-1} < x < p_1 + p_2 + \cdots + p_n = 1$$

After observing that the system is in state $x_k$, the length of the measurement interval becomes $p_k$, whereas, at the beginning (without any measurement), we have $0 < x < 1$, i.e., a length interval equal to one. Therefore, the gain in information compared with the case before the observation is

$$\log_2(1/p_k)$$

The average information $\overline{I}$ gained when the system changes from one state to another, up to the state $x_n$, is equal to

$$\begin{aligned} \overline{I} &= p_1 \log_2(1/p_1) + p_2 \log_2(1/p_2) + \cdots + p_n \log_2(1/p_n) \\ &= \sum_{k=1}^{n} p_k \log_2(1/p_k) \\ &= -\sum_{k=1}^{n} p_k \log_2 p_k = H \end{aligned}$$

where $H$ is exactly the *entropy* introduced by Shannon in his 1948 paper [9]. Here, if $p_k = 0$ for some $k$, the value of the term $0 \log_2 0$ is taken to be 0, consistent with the limit $\lim_{p \to 0^+} p \log_2 p = 0$.

#### 4.5.2.2 Direct Definition of Entropy

This direct way of defining entropy is due to Shannon and can be found in most textbooks on information theory or communication systems. Consider a random variable $X$ that carries an infinite amount of information if it is continuous in amplitude range. Each realization (presentation) of $X$ can be considered to represent

a message. Actually, in a physical or biological system, it is not realistic to assume that the amplitude measurements can have infinite accuracy. Therefore, it seems to be realistic to assume that the value of $X$ can be uniformly *quantized* to a finite set of values, in which case $X$ is *discrete* random variable:

$$X = \{x_k; k = 0, \pm 1, \pm 2, \ldots, \pm n\}$$

where $2n + 1$ is the total number of discrete levels. Now, suppose that the event $X = x_k$ occurs with probability:

$$p_k = P(x_k) = P(X = x_k)$$

under the conditions:

$$0 \leq p_k \leq 1 \quad \text{and} \quad \sum_{k=1}^{n} p_k = 1$$

If for some $k$, the event $X = x_k$ occurs with probability $p_k = 1$, in which case all other $p_i$'s for $i \neq k$ are zero, the occurrence of the event $X = x_k$ does not add any "*information*" and does not incur any "*surprise*", since we know surely what the message is. However, if the various discrete levels occur with different probabilities, the occurrence of some event $X = x_k$ conveys some *information* (or surprise), which is higher when $p_k = p(x_k)$ is lower (i.e., the *uncertainty* about the occurrence of $X = x_k$ is higher). Thus, the terms "*uncertainty*", "*surprise*", and "*information*" can be used interchangeably in the information theory framework. In particular, the amount of information (or surprise) is related to the inverse of the probability of occurrence.

On the basis of this discussion, the amount of information obtained after the observation of the event $X = x_k$ with probability $p_k$ is defined as

$$I_B(x_k) = \log_B(1/p_k) = -\log_B p_k$$

where the base $B$ of the logarithm is arbitrary and when the base $B = 2$, $I(x_k)$ is measured in bits, when $B = 10$ $I(x_k)$ is measured in decimal digits (Hartley's), and when $B = e$ (natural logarithm) $I(x_k)$ is measured in natural units (nats).

The quantity $I(x_k)$ defined here has the following properties:

- $I_B(x_k) = 0$ for $p_k = 1$ (i.e., no information is gained if the occurrence of $X = x_k$ is absolutely certain).
- $I_B(x_k) \geq 0$ for $0 \leq p_k \leq 1$ (i.e., a loss of information never occurs).
- $I_B(x_k) > I(x_i)$ for $p_k < p_i$ $(k \neq i)$ (i.e., the less probable an event is, the more information is gained after its occurrence).

Clearly, the quantity $I_B(x_k)$ is a random variable with probability $p_k$. Therefore, the mean value $\overline{I}_B$ of $I_B(x_k)$ over the entire range of $2n + 1$ discrete values is equal to

$$\overline{I}_B(X) = E[I(x_k)] = \sum_{k=-n}^{n} p_k I(x_k)$$
$$= -\sum_{k=-n}^{n} p_k \log p_k = H(X)$$

and is called the *entropy* of the discrete random variable $X$, which takes a finite set of $(2n+1)$ values. The entropy $H(X)$ provides a measure of the average amount of information provided per message. The basic properties of the entropy $H(X)$ are as follows:

- $H(X)$ is continuous (a small change of the probabilities yield a small change in $H(X)$);
- $0 \leq H(X) \leq \log_B(2n+1)$;
- $H(X) = 0$ if and only if $p_k = 1$ for some $k$ (i.e., if there is no uncertainty);
- $H(X) = \log_B(2n+1)$ if and only if all discrete levels are equally probable ($p_k = 1/(2n+1)$ for all $k$);
- $H(X)$ is additive (i.e., the entropy is independent to how the process is divided into parts). This allows us to calculate the entropy of a system via the entropies of the subsystems, if we know the interactions between the subsystems. That is if the system is divided into $M$ blocks (subsystems) with $b_1, b_2, \ldots, b_M$ elements each, we have

$$H\left(\frac{1}{2n+1}, \ldots, \frac{1}{2n+1}\right) = H_M\left(\frac{b_1}{2n+1}, \ldots, \frac{b_M}{2n+1}\right) + \sum_{k=-n}^{n} \frac{b_k}{2n+1} H_{b_k}\left(\frac{1}{b_k}, \ldots, \frac{1}{b_k}\right);$$

- $H(X)$ remains unchanged if the events $X = x_k$ are re-ordered, i.e.,

$$H(x_n, \ldots, x_0, x_1, \ldots, x_n) = H(x_{-n+1}, x_n, \ldots, x_0, x_1, \ldots, x_n);$$

- For equal probability events, $H(X)$ increases with the number of observations, i.e.,

$$H\left(\frac{1}{2n+1}, \ldots, \frac{1}{2n+1}\right) < H\left(\frac{1}{2n+2}, \ldots, \frac{1}{2n+2}\right);$$

- The addition or removal of an event with probability zero does not change the value of the entropy:

$$H\left(\frac{1}{2n+1}, \ldots, \frac{1}{2n+1}, 0\right) = H\left(\frac{1}{2n+1}, \ldots, \frac{1}{2n+1}\right).$$

#### 4.5.2.3 Differential Entropy

The concept of entropy thus far discussed refers to discrete random variables. Here, we will introduce the corresponding concept for the case of a continuous random variable $X$ that has a probability density function $f(x)$. This concept is the following:

$$\begin{aligned} h(X) &= -\int_{-\infty}^{\infty} f(x)\log_B f(x)\,\mathrm{d}x \\ &= -E[\log_B f(x)] \end{aligned}$$

and is called the *differential entropy* of $X$ (in contrast to the absolute entropy $H(X)$). Of course, it is remarked right from now that $h(X)$ does not provide in any way a measure of the randomness of $X$. This expression for the differential entropy is justified by using the mean value theorem:

$$f(x_k)\delta x = \int_{k\delta x}^{(k+1)\delta x} f(x)\,dx$$

which leads to the Riemannian approximation:

$$\int_{-\infty}^{\infty} f(x)\,dx = \lim_{\delta x\to 0}\sum_{k=-\infty}^{\infty} f(x_k)\delta x$$

Define

$$\begin{aligned} H_\delta &= -\sum_{k=-\infty}^{\infty} f(x_k)\delta x\log_B[f(x_k)\delta x] \\ &= -\sum_{k=-\infty}^{\infty} f(x_k)\delta x\log_B f(x_k) - \sum_{k=-\infty}^{\infty} f(x_k)\delta x\log_B \delta x \end{aligned}$$

Then, as $\delta x \to 0$:

$$\lim_{\delta x\to 0} H_\delta = -\int_{-\infty}^{\infty} f(x)\log_B f(x)\,\mathrm{d}x - \lim_{\delta x\to 0}(\log_B \delta x)\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x$$

Or

$$h(X) = -\int_{-\infty}^{\infty} f(x)\log_B f(x)\,dx = \lim_{\delta x \to 0}(H_\delta + \log_B \delta x)$$

where the relation $\int_{-\infty}^{\infty} f(x)\,dx = 1$ was used.

Now, $\lim_{\delta x \to 0} \log_B \delta x = -\infty$, which implies that the entropy of a continuous variable is infinitely large. This is indeed justified by the fact that a continuous random variable may take any value in the interval $(-\infty, \infty)$, and so the uncertainty associated with the variable is infinite. Regarding the infinite offset, $\log_B 0 = -\infty$, as a *reference*, the quantity $h(X)$ is actually a *differential entropy*. Clearly, the differential entropy is not a limit of the Shannon entropy for $n \to \infty$, but differs from this limit by the infinite offset $\log \delta x \to -\infty$ as $\delta x \to 0$.

#### 4.5.2.4 Joint Entropy, Conditional Entropy, and Mutual Information

**Joint entropy** Consider two discrete random variables $X$ and $Y$. The entropy of their pairing $(X, Y)$ is called *joint entropy* $H(X, Y)$, i.e.,

$$\begin{aligned} H(X,Y) &= E_{(X,Y)}[-\log_B p(x,y)] \\ &= -\sum_{x,y} p(x,y)\log_B(x,y) \end{aligned}$$

**Conditional entropy** The *conditional entropy* of the random variable $X$ given the random variable $Y$ is defined as

$$\begin{aligned} H(X|Y) &= E_Y[H(X|Y)] \\ &= -\sum_{y\in Y} p(y)\sum_{x\in X} p(x|y)\log_B p(x|y) \\ &= -\sum_{x,y} p(x,y)\log_B[p(x,y)\,/\,p(y)] \end{aligned}$$

i.e., $H(X|Y)$ is the average conditional entropy $E_Y[H(X|Y)]$ over $Y$. It follows that

$$H(X|Y) = H(X,Y) - H(Y)$$

**Mutual information** The *mutual information* (or *transinformation*) is defined to be the amount of information that can be obtained about one random variable $X$ via the observation of another variable $Y$, i.e.,

$$I(X;Y) = E_{X,Y}[I_s(x,y)] = \sum_{x,y} p(x,y) \log_B \frac{p(x,y)}{p(x)\,p(y)}$$

where $I_s(x,y)$ is the point-wise (specific) mutual information. It follows that $I(X;Y)$ has the following property:

$$I(X;Y) = H(X) - H(X|Y)$$

This means that, if we know $Y$, we can have a saving of $I(X;Y)$ bits in encoding $X$, compared to the case in which we do not know $Y$. The mutual information has the following property:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(X) - [H(X,Y) - H(Y)] \\ &= H(X) + H(Y) - H(X,Y) \\ &= H(Y) - [H(X,Y) - H(X)] \\ &= H(Y) - H(Y|X) = I(Y;X) \end{aligned}$$

i.e., the mutual information is *symmetric*.

## 4.5.3 Coding Theory

### 4.5.3.1 General Issues

Coding theory is the branch of information theory that deals with the transmission of information across communication channels (noiseless, noisy) and the recovery of the messages. Coding theory aims to make messages easy to read, and should not be confused with *cryptography*, which is concerned with making messages hard to read. One of the principal goals of coding theory is to remove information redundancy (i.e., compress data) and correct the errors that may occur. Specifically, coding theory studies the properties of codes and their suitability for particular applications. Here, a brief outline of coding theory will be given. Full presentations can be found in the respective literature (e.g., [54–66]).

Coding is categorized into three categories:

- Source coding.
- Channel coding.
- Combined source and channel coding.

*Source encoding* (or *data compression*) performs data compression on the information sent by a source, so that the transmission is more efficient. *Channel encoding* adds extra bits to assure a more robust transmission regarding disturbances and noise present on the transmission channel.

#### 4.5.3.2 Source Coding

*Source coding* (*compression*) is the process of encoding information using fewer information-bearing units (e.g., bits) compared to a non-coded representation, through the employment of particular encoding schemes. Compression is very useful since it contributes to the reduction of the consumption of expensive resources. The development of source coding techniques and the design of equipment must be based on compromises among several factors, such as the degree of compression, the level of noise and distortion involved, and the computational resources needed for the coding and recovery of the data.

The two ways of data compression are as follows:

- Lossless data compression, where the data must be recovered precisely.
- Lossy data compression, where the bits required for recovering the data with the desired reliability, as measured by a distortion index, are added.

#### 4.5.3.3 Channel Coding

*Channel coding* is concerned with maximizing the information rate that the channel can convey reliably, i.e., with acceptable error probability. To this end, codes must be designed that allow fast transmission of data, involve many valid codewords, and can detect and correct various errors. This again requires a number of trade-offs between several conflicting issues. Although source coding tries to remove as much redundancy as possible, channel coding designs and investigates several error-correcting codes, which add sufficient redundancy (i.e., error correction) that assures efficient and faithful transmission of information across a noisy channel.

#### 4.5.3.4 Error Detecting and Correcting Codes

**Error detection** It deals with the detection of errors introduced by noise and other disturbances across the transmission channel from the transmitter to the receiver. The general way is to add some extra data bits (i.e., redundancy) to a message, which makes possible the detection of any errors in the conveyed message. The additional bits are called *check bits* and are determined by a suitable algorithm applied to the message bits. To detect any existing errors, the receiver applies the same algorithm to the received data bits and compares the output to the received check bits. If no matching of the values occurs, an error has taken place at some point of the transmission.

**Error correction** It deals with both the detection of errors and the recovery of the initial, error-free data. The three primary ways for design the channel code for error correction are as follows:

- **ARQ** (*Automatic Repeat reQuest*): The receiver requests the retransmission of the data that are assumed to contain errors.
- **FEC** (*Forward Error Correction*): The transmitter sends the encoded message with an error-correcting code. The receiver decodes the received message into the most likely data, without sending any request for retransmission.
- **Hybrid ARQ and FEC**: Here minor errors are restored without retransmission, and major errors are corrected by sending a request for retransmission.

Three ARQ schemes that are suitable for varying or unknown capacity communication channels (such as the Internet) are as follows:

- Stop-and-Wait ARQ.
- Go-back-N ARQ.
- Selective Repeat ARQ.

**FEC** is typically employed in lower layer communication and in storage devices (CDs, DVDs, Dynamic RAM) and are classified into two classes:

- *Block codes* which are processed in blocks. In this class, we have, among others, repetition codes, Hamming codes, and multidimensional parity-check codes.
- *Convolutional codes* which are processed on a bit-by-bit basis. Here, the so-called *Viterby decoder* performs optimal decoding.

#### 4.5.3.5 Block Codes

An $(n, k)$ block code, where $k$ is the number of input bits and $n$ is the number of output bits, is characterized by the following parameter:

$$\begin{aligned} &\textit{Code rate}: \ r = k/n \\ &\textit{Channel data rate}: \ R_0 = rR_n, \end{aligned}$$

where $R_n$ denotes the bit rate of the information source.

A binary block code $C$ of block length $n$ is a subset of the set of all binary n-tuples

$\bar{\mathbf{x}} = [x_0, x_1, \ldots, x_{n-1}]$, where $x_i = 0$ or 1 for $i = 0, 1, \ldots, n-1$. *Code vector* or *code word* is called an n-tuple belonging to the code.

- *Hamming weight* $w([x_o, x_1, \ldots, x_{n-1}])$ is the number of nonzero components of the n-type.
- *Hamming distance* $d(\mathbf{x}, \mathbf{x}')$ between two n-tuples $\mathbf{x}$ and $\mathbf{x}'$ is defined as the number of positions in which their components differ. From this definition, it follows that

$$d(\mathbf{x}, \mathbf{x}') = w(\mathbf{x} - \mathbf{x}'),$$

where
$\mathbf{x} - \mathbf{x}' = [x_0, x_1, \ldots, x_{n-1}] - [x'_0, x'_1, \ldots, x'_{n-1}] = [x_0 - x'_0, x_1 - x'_1, \ldots, x_{n-1} - x'_{n-1}]$

**Minimum Hamming distance**, $d_{\min}$, of the block code $C$ is the smallest Hamming distance between pairs of distinct code words.

Some theorems on the detection and correction ability of block codes are the following [61].

**Theorem 1** A code C can detect all patterns of s or fewer errors if and only if $d_{\min} > s$.

**Theorem 2** A code C can correct all patterns of s or fewer errors if and only if $d_{\min} > 2s$.

This theorem suggests the following implicit decoding rule: "*Decode* $\boldsymbol{c}$ (*the corrupted received codeword of an actual codeword* $\mathbf{a}$) *to the nearest codeword in terms of the Hamming distance, i.e.,* $d_{\min} = \min w(e) = \min w(\mathbf{c} - \mathbf{a}) = \min d(\mathbf{c}, \mathbf{a})$"

**Definition** A binary code C is linear if, for $\mathbf{a}$ and $\mathbf{a}'$ in C, the sum $\mathbf{a} + \mathbf{a}'$ is also in C. For example, the code $C = \{[0,0,0] + [1,1,0] + [1,0,1] + [0,1,1]\}$ is a linear code because $[1,1,0] + [101] = [011]$, and so on.

- The *minimum Hamming distance* of a linear block code is equal to the smallest weight of the nonzero code words in the code.

#### 4.5.3.6 Convolutional Codes

A *binary convolutional code* is characterized by a 3-tuple ($n$, $k$, $m$), where k is the number of input (message) bits, n is the number of generated (output) bits, and m is the number of memory registers (memory order). As in block codes, the parameter $r = k / n$ is called the *code rate*. Typically, k and n vary from 1 to 8, m from 2 to 10, and the code rate r from 1/8 to 7/8. In deep space applications, code rates r as low as 1/100 or smaller have been used. In many commercial cases, the convolutional codes are described as ($r$, $L$), where r is the code rate and L is the *constraint length* $L = k(m - 1)$, which represents the number of bits in the encoder memory that affect the generation of the n output bits.

A binary convolutional encoder is represented pictorially by a set of shift registers and modulo-2 adders, where the output bits are modulo-2 sums of selective shift register contents and present input bits. The diagram of a (2, 1, 2) convolutional code r is shown in Fig. 4.23.

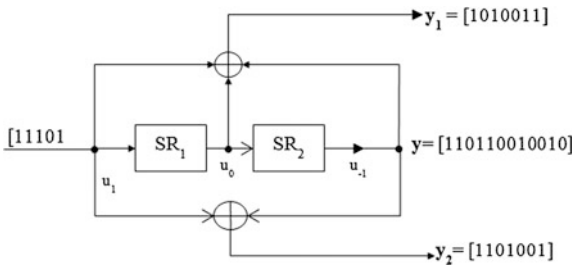



**Fig. 4.23** Binary (2, 1, 2) convolutional encoder ($SR_i = i$th shift register)

This is a rate $r = 1/2$ code. Each input bit is coded into two output bits. The constraint length of the code is $L = k(m-1) = 1 \times (2-1) = 1$.

The choice of the bits to be added for generating the output bit is determined by the so-called *generator polynomial* (g) for that output. In Fig. 4.23, the first output bit has a generator polynomial of 111. The second output has a generator polynomial of 101. The output bits of $\mathbf{y}_1$ *and* $\mathbf{y}_2$ are given by

$$y_{1,i} = \text{mod } 2[u_1 + u_0 + u_{-1}]_i = u_{1,i} \oplus u_{0,i} \oplus u_{-1,i}$$
$$y_{2,i} = \text{mod } 2[u_1 + u_{-1}]_i = u_{1,i} \oplus u_{-1,i}$$

The generator polynomial gives a unique error protection quality to the code. For example, one code (3, 2, 2) may have entirely different properties from another, depending on the selected polynomial. However, it is noted that not all polynomials lead to good error correction performance. A list of good polynomials for rate 1/2 codes is given in Table 4.2 [54].

The encoders of convolutional codes are represented by multivariable (**MV**) linear time-invariant (**LTI**) systems as shown in Fig. 4.24.

The $\mathbf{y}_j$ output in Fig. 4.21 is given by

$$\mathbf{y}_j = \sum_{i=1}^{k} \mathbf{x}_i * \mathbf{g}_j^{(i)}$$

**Table 4.2** Efficient generator polynomials for 1/2 codes

| Constraint length | $G_1$ | $G_2$ |
|---|---|---|
| 3 | 110 | 111 |
| 4 | 1101 | 1110 |
| 5 | 11010 | 11101 |
| 6 | 110101 | 111011 |
| 7 | 110101 | 110101 |
| 8 | 110111 | 1110011 |
| 9 | 110111 | 111001101 |
| 10 | 110111001 | 1110011001 |

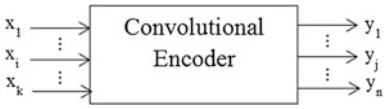



**Fig. 4.24** LTI representation of a MV convolutional encoder

where "*" denotes the *convolution operator,* and $\mathbf{g}_j^{(i)}$ is the *impulse response* of the *i*th input sequence with respect to the *j*th output. The *impulse responses* are the *generator polynomials* (*sequences*) of the encoder, previously discussed.

The impulse response for the binary (2, 1, 2) convolutional code of Fig. 4.20 is

$$\begin{aligned}\mathbf{g}_1 &= [1,1,1,0,\ldots] = [1,1,1]\\ \mathbf{g}_2 &= [1,0,1,0,\ldots] = [1,0,1]\end{aligned}$$

The outputs $\mathbf{y}_1$ and $\mathbf{y}_2$ corresponding to the input vector $\mathbf{x} = [1\ 1\ 1\ 0\ 1]$ are equal to

$$\begin{aligned}\mathbf{y}_1 &= [1\ 1\ 1\ 0\ 1] * [1\ 1\ 1] = [1\ 0\ 1\ 0\ 0\ 1\ 1]\\ \mathbf{y}_2 &= [1\ 1\ 1\ 0\ 1] * [1\ 0\ 1] = [1\ 1\ 0\ 1\ 0\ 0\ 1]\end{aligned}$$

As in block codes, the convolutional codes can be generated by a generator matrix multiplied by the input (information, message) vectors.

Let $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$ and $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\}$ be the input and output sequences. Arranging the input sequences as

$$\begin{aligned}\mathbf{x} &= \left[x_{1,0}, x_{2,0}, \ldots, x_{k0}; \ldots; x_{1,k}, x_{2,k}, \ldots, x_{k,k}\right]\\ &= \left[\mathbf{u}_0, \mathbf{u}_1, \ldots, \mathbf{u}_p, \ldots\right]\end{aligned}$$

and the output sequences as

$$\begin{aligned}\mathbf{y} &= \left[y_{1,0}, y_{2,0}, \ldots, y_{n,0}; \ldots; y_{n,0}, y_{n,1}, \ldots\right]\\ &= \left[\mathbf{z}_0, \mathbf{z}_1, \ldots, \mathbf{z}_p, \ldots\right]\end{aligned}$$

the *convolutional encoder* is described by the matrix-vector equation:

$$\mathbf{y} = \mathbf{x}\mathbf{A}$$

where $\mathbf{A}$ is the generator matrix of the code:

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_0 & \mathbf{A}_1 & \mathbf{A}_2 & \cdots & \mathbf{A}_m & & \\ & \mathbf{A}_0 & \mathbf{A}_1 & \cdots & \mathbf{A}_{m-1} & \mathbf{A}_m & \\ & & \mathbf{A}_0 & \cdots & \mathbf{A}_{m-2} & \mathbf{A}_{m-1} & \mathbf{A}_m \\ & & \ddots & & & & \end{bmatrix}$$

where the $k \times n$ blocks $\mathbf{A}_p$ are given by

$$\mathbf{A}_p = \begin{bmatrix} \mathbf{g}_{1,p}^{(1)} & \cdots & \mathbf{g}_{n,p}^{(1)} \\ \mathbf{g}_{1,p}^{(2)} & \cdots & \mathbf{g}_{n,p}^{(2)} \\ \cdots & \cdots & \cdots \\ \mathbf{g}_{1,p}^{(k)} & \cdots & \mathbf{g}_{n,p}^{(k)} \end{bmatrix}$$

and $g_{jp}^{(i)}$ the impulse response of the ith input with respect to the $j$th output:

$$\mathbf{g}_j^{(i)} = \left[ g_{j,0}^{(i)}, g_{j,1}^{(i)} \cdots g_{j,p}^{(i)} \cdots g_{j,m}^{(i)} \right]$$

The output vector $\mathbf{y}$ of the binary encoder (2, 1, 2) of Fig. 4.30 is given by

$$\mathbf{y} = \mathbf{x}\mathbf{A}$$

where $\mathbf{x} = [1\ 1\ 1\ 0\ 1]$, and

$$\mathbf{A} = \begin{bmatrix} 11 & 10 & 11 & & & & \\ & 11 & 10 & 11 & & & \\ & & 11 & 10 & 11 & & \\ & & & 11 & 10 & 11 & \\ & & & & 11 & 10 & 11 \end{bmatrix}$$

i.e.,

$$\mathbf{y} = [11,\ 01,\ 10,\ 01,\ 00,\ 10,\ 11]$$

as shown in Fig. 4.30.

A better understanding of the operation of an encoder can be obtained using one or more of the following graphical representations [59–64]:

- State diagram.
- Tree diagram.
- Trellis diagram.

The definition and the way these encoder diagrams can be drawn and used are given in the literature [54–57].

For the decoding of convolutional codes, there are available several different approaches that can be grouped into two basic classes [59–64]:

- Sequential decoding (*Fano* algorithm).
- Maximum-likelihood decoding (*Viterbi* algorithm).

Sequential decoding was one of the earliest techniques developed for decoding convolutionally coded bit streams. It was first coined by Wosencraft and refined by Fano. Maximum-likelihood decoding is a good alternative methodology that is best

implemented by the Viterbi algorithm. For full presentations of the above decoding techniques, the reader is referred to the bibliography [54–64].

### 4.5.4 Fundamental Theorems of Information Theory

Here, four fundamental theorems of information theory will be reviewed and their roles and usefulness in practice will be explained. These theorems are the following:

- Nyquist–Shannon sampling theorem.
- Shannon's source coding theorem.
- Shannon's noisy channel coding and capacity theorem.
- Landau–Pollack bandwidth signal dimensionality theorem.

#### 4.5.4.1 Nyquist–Shannon Sampling Theorem

This theorem, which was first formulated by Harry Nyquist in 1928 [67] and formally proved by Shannon in 1949 [68], states: "*To be able to reconstruct perfectly a sampled analog signal* $x(t)$ *from its sampled version* $x(kT),\ k = 0, 1, 2, \ldots$ *the sampling frequency* $1/T$ *Hz (where T is the sampling period) must be greater than twice the highest frequency W of* $x(t)$*.*"

If the sampling frequency is less than this limit, then frequencies in the original analog signal that are greater than half the sampling frequency will be "*aliased*" and will appear in the resulting signal as lower frequencies. If the sampling frequency is exactly twice the highest frequency of the analog input, then "*phase mismatching*", between the sampler and the signal, will distort the signal. Therefore, in practice, an analog low-pass filter must be used before the sampler to guarantee that no components with frequencies above the sampling frequency remain. This is called an "*anti-aliasing filter*" and must be very carefully designed, because a poor filter causes phase distortion and other effects. The minimum sampling frequency 2 W that permits exact reconstruction of the original signal is known as the *Nyquist frequency* (or *Nyquist rate*), and the time spacing between samples is known as "*Nyquist time interval*". To express the above concepts mathematically, let a signal $x(t)$ have a Fourier transform:

$$F[x(t)] = X(f) = 0 \quad for \quad |f| > W.$$

Then, $x(t)$ is completely determined by giving the value of the signal at a sequence of points, spaced $T = 1/2W$ apart. The values $x_k = x(k/2W) = x(kT)$, $T$ = sampling period, are called the samples of $x(t)$.

The sampled signal $x^*(t)$ is expressed as the amplitude modulation via $x(t)$ of a $\delta-$ pulse (Dirac) train $\sum_{k=-\infty}^{\infty} \delta(t - kT)$, i.e.,

$$x^*(t) = x(t) \sum_{k=-\infty}^{\infty} \delta(t - kT)$$

Since multiplication in the time domain is expressed by convolution "*" in the frequency domain, we have

$$\begin{aligned} X^*(f) &= X(f) * \left[ \frac{1}{T} \sum_{j=-\infty}^{\infty} \delta(f - j/T) \right. \\ &= \frac{1}{T} \int_{-\infty}^{\infty} X(s) \sum_{j=-\infty}^{\infty} \delta(f - s - j/T)\, \mathrm{d}s \\ &= \frac{1}{T} \sum_{j=-\infty}^{\infty} X(f - j/T) = X\left(\mathrm{e}^{j2\pi/T}\right) \end{aligned}$$

Figure 4.25 shows pictorially the Fourier transform $X^*(f) = X\left(\mathrm{e}^{j2\pi/T}\right)$ of the sampled version $x^*(t)$ of a signal $x(t)$ that has the Fourier transform $X(f)$ [42]. In this case, $X(f) \neq 0$, outside the frequency region determined by the Nyquist frequency $W = 1/2T$. Therefore, "aliasing distortion" appears, which is due to the overlap of the various periodically repeated sections of $X^*(f)$ in the frequency domain.

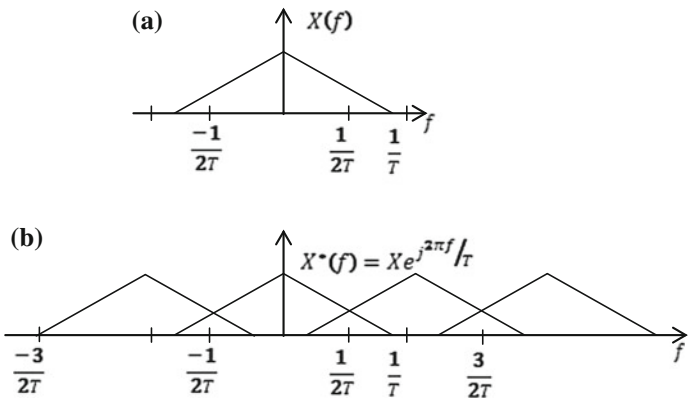



**Fig. 4.25** **a** Fourier transform of a continuous-time signal, **b** the corresponding Fourier transform of the sampled version with sampling frequency $1/T$. The overlaps indicate that in this case we have "aliasing effects"

#### 4.5.4.2 Shannon's Source Coding Theorem

Consider a source that assigns to each outcome (sample) $\mathbf{x}$ a binary word, which as we have seen is called the *code*. Here, the vector $\mathbf{x}$ is an outcome of a discrete random vector $\mathbf{X}$. The *asymptotic equipartition theorem* (which is the basis for the entropy's interpretation) states that for $\mathbf{x} \in \mathbf{X} = \{X_1, X_2, \ldots, X_n\}$, where $X_i$ are independent trials of $\mathbf{X}$, as $n \to \infty$ (i.e., asymptotically), there is a set of "*typical*" outcomes $S$ for which

$$P_X(\mathbf{x}) = 2^{-nH(\mathbf{X})}, \ \mathbf{x} \in S$$

and the total probability that the outcome is in $S$ is almost one. Since the "*typical*" outcomes are all equiprobable, this means that there must be approximately $2^{nH(\mathbf{X})}$ outcomes in $S$. As $n$ becomes larger, this approximation becomes more accurate [42].

In the information source, when $n$ is large, we can assign only the "*typical*" outcomes and omit the "*non-typical*" ones. By using $nH(\mathbf{X})$-bit code words, we are able to encode each of the $2^{nH(\mathbf{X})}$ typical outcomes with a unique binary word, for an average of $H(\mathbf{X})$ bits per component of the vector $\mathbf{x}$. Since $H(\mathbf{X})$ represents the average information obtained from the observation, each outcome of $\mathbf{X}$ needs an average of $H(\mathbf{X})$ bits. This is true only for an average of n components, not for an individual component.

The source coding theorem states that [42]:

> If a source can be modeled as repeated independent trials of a random variable $\mathbf{X}$ at r trials per second, then the source can be encoded by a source coder into a bit stream with bit rate less than $R + \varepsilon$, for any $\varepsilon > 0$, where $R = rH(\mathbf{X})$ is the so-called "rate of the source".

This source coding theorem establishes the limits of data compression. Stated in another way, Shannon's source coding theorem says that [69]:

> The minimum average number of bits, C, needed to encode n symbols (which are treated as n independent samples of a discrete random vector $\mathbf{X}$ with probability $p_{\mathbf{X}}(\mathbf{x})$ and entropy $H(\mathbf{X})$) satisfies the relation:
>
> $$H(\mathbf{X}) \leq C < H(\mathbf{X}) + 1/n.$$

In practice, $p_{\mathbf{X}}(\mathbf{x})$ is not exactly available, but we only have an estimate $q_{\mathbf{X}}(\mathbf{x})$ for use in the source coding process. In this case, the corresponding minimum $C_q$ satisfies

$$H(\mathbf{X}) + KL(p\|q) \leq C_q < H(\mathbf{X}) + KL(p\|q) + 1/n$$

where $KL(p\|q)$ is the relative entropy (or the *Kullback–Leibler*: KL) divergence, defined as

$$KL(p\|q) = \sum_{\mathbf{x}} p_{\mathbf{X}}(\mathbf{x}) \log[p_{\mathbf{X}}(\mathbf{x})/q_{\mathbf{X}}(\mathbf{x})] \geq 0$$

We note that $KL(p\|q)$ is the difference between the two probability distributions, $p_{\mathbf{X}}(\mathbf{x})$ and $q_{\mathbf{X}}(\mathbf{x})$. Clearly, $KL(p\|q) = 0$ if and only if $p_{\mathbf{X}}(\mathbf{x}) = q_{\mathbf{X}}(\mathbf{x})$.

Constructing practical codes that are close to $R$ is difficult. However, constructing good suboptimal codes is usually easy (see Sect. 4.5.3). In words, the source coding theorem (as expressed in the above two alternative ways) says that no source coding scheme can be better than the "*entropy of the source*", or bit stream rate less than the "*source rate*".

*Example* Let a binary random variable $\mathbf{X}$ with set of values (alphabet) $A = \{0, 1\}$. Its entropy is equal to

$$H(\mathbf{X}) = -a \log_2 a - (1 - a) \log_2 (1 - a)$$

where $a$ denotes the probability of taking the value 1, $a = p_{\mathbf{X}}(1)$. The entropy $H(\mathbf{X})$ is a function of $a$, i.e., $H = H(a)$ which has the graphical representation shown in Fig. 4.26 [42].

We observe that when $a = 1 - a = 1/2$ (equiprobable values 0 and 1), $H(\mathbf{X})$ takes a maximum, in agreement with the theory. The maximum of $H(\mathbf{X})$ is one bit. Also, $H(\mathbf{X})$ is zero when either $a = 0$ $(1 - a = 1)$ or $a = 1$, again in agreement with the theory. The fact that, for $a = 1 - a = 1/2$, the value of the entropy is $H(\mathbf{X}) = 1$ means that, to encode repeated samples of $\mathbf{X}$, we need on average one bit per sample (outcome). Here, this is also sufficient for each sample, not just on average, since the source is binary. A source coder that can achieve rate $R$ transmits samples of $\mathbf{X}$ unaltered (exactly). Now, suppose that $a = 0.1$, in which case

$$\mathrm{H(X)} = -0.1 \log_2(0.1) - 0.9 \log_2(0.9) = 0.47$$

This means that, to encode repeated samples of $\mathbf{X}$, we need 0.47 bits per outcome. Clearly, there are coding methods with average number of bits greater than 0.47 but less than unity.

In information theory, when dealing with a language, we speak about the entropy or rate of the language. For example, in the case of the English language, the *alphabet* consists of 26 letters (symbols) plus some additional symbols such as

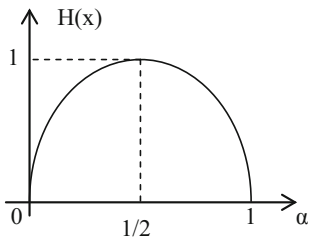



**Fig. 4.26** Graphical representation of the entropy $H(\mathbf{X})$ of a binary function in terms of the probability $a = p_{\mathbf{X}}(1)$

*space*, *comma*, *period*, etc. These symbols can be treated as n independent samples of a random variable **X** transmitted via the communication channel, each having a probability $p_{\mathbf{X}}(\mathbf{x})$, and entropy $H(\mathbf{X}) = -\sum_{\mathbf{x}\in\mathbf{X}} p_{\mathbf{X}}(\mathbf{x}) \log p_{\mathbf{X}}(\mathbf{x})$. For example, $p_{\mathbf{X}}(\mathbf{x} =$ "*s*"$)$ is much higher than $p_{\mathbf{X}}(\mathbf{x} =$ "*z*"$)$. To minimize the code length for the language, we assign shorter (more compressed) codes for symbols of higher probabilities (here, a shorter code for the letter "s" than that of the letter "z"). As length code for a symbol **x** with $p_{\mathbf{X}}(\mathbf{x})$, we can use its "*surprise*" $-\log p_{\mathbf{X}}(\mathbf{x})$.

#### 4.5.4.3 Shannon's Noisy Channel Coding and Capacity Theorem

Shannon's noisy channel coding and capacity theorem deal with the maximum possible efficiency of error-correcting codes (see Sect. 4.5.3) versus levels of data corruption and noise interference [10, 68]. This theorem establishes that a randomly designed error-correcting code is basically as good as the best possible code, and states [70]:

"The capacity of a discrete-memoryless channel is given by

$$C = \max_{p_{\mathbf{X}}(\mathbf{x})} \{I(X;Y)|p_{\mathbf{X}}(\mathbf{x})\} \quad \text{(bits/symbol)}$$

where $I(X;Y)$ is the *mutual information* between the channel input $X$ and the output $Y$:

$$I(X;Y) = H(X) - H(X|Y)$$

If the transmission rate $R$ is less than $C$, then, for any $\varepsilon > 0$, there exists a code with block length $n$ large enough whose error probability is less than $\varepsilon$. When $R > C$, the error probability of any code with any block length is bounded away from zero. If the channel is used $m$ times per second, then the channel capacity in bits per second is $C' = mC$ ".

*Example* Let a binary symmetric channel with cross-over probability 0.1. Then, $C \approx 0.5$ bits per transmission. Thus, we can send reliably through the channel 0.4 bits/per channel. This can be achieved by taking (for example) 400 input information bits and map them into a code of length 1000 bits. The whole code is then transmitted through the channel, in which case 400 information bits can be decoded correctly, but 100 bits may be detected incorrectly. Now, let us consider a continuous-time additive white-Gaussian channel with fully uncorrelated signal and noise. In this case, the channel capacity is found to be [70]:

$$C = W \log_2\left(1 + \frac{S}{N_0 W}\right) \frac{\text{bits}}{\text{s}}$$

where $S$ is the upper bound of the power of the input signal **x** (measured in Watt), $N_0/2$ is the Gaussian noise power spectral density, and $W$ is the channel bandwidth

in Hz. Clearly, $N_0W = N$ where $N$ is the total noise or interference power over the bandwidth $W$?

*Example* Using the above formula for the capacity of a Gaussian channel, we can compute the capacity of the voice band of a telephone channel. Typical values in this case are as follows:

$$W = 3000Xz,\ S/N = 1000 \text{ or } 10\log_{10}(1000) = 30\,\text{db}.$$

Thus, $C = 3000\log_2(1 + 1000) \approx 30$ kbits/s

This means that using this model one cannot design modems faster than 30 kbits/s. Because the signal-to-noise ratio is large, one can expect to be able to transmit $C/W = 10\,\text{bits/s/Hz}$ across the telephone channel.

*Remarks*

(a) If the noise and signal are not fully uncorrelated (as in the case of non-white additive noise), the signal-to-noise ratio S/N is not constant with frequency over the bandwidth. In this case, we can assume that the channel is modeled as a multitude of narrow independent Gaussian channels in parallel, and so C is given by

$$C = \int_0^W \log_2\left(1 + \frac{S(f)}{N(f)}\right) df$$

where $S(f)$ and $W(f)$ are the signal and noise power spectrums, respectively, which are functions of the frequency $f$.

(b) For large *S/N* ratio (i.e., $S/N \gg 1$), we get $C \approx 0.332\,W\ (S/N,\ \text{in db})$, where $S/N$ in db $= 10\log_{10}(S/N)$.

(c) For very small *S/N* (i.e., $S/N \ll 1$), we obtain $C \approx 1.44W(S/N) = 1.44W(S/N_0W) = 1.44(S/N_0)$, i.e., the channel capacity in this case is (approximately) independent of the noise bandwidth.

#### 4.5.4.4 Landau–Pollack Bandwidth Signal Dimensionality Theorem

This theorem is a consequence of the Nyquist–Shannon sampling theorem and states [42]: "*A signal cannot be both band-limited and time-limited.*"

Actually, a band-limited signal is not time-limited, because its energy cannot be entirely restricted to any finite interval of time. Similarly, a time-limited function is not band-limited because its energy cannot be totally confined to a finite band of frequencies. However, one can assume that a band-limited signal is approximately time-limited, and a time-limited signal is approximately band-limited.

Quantitatively, the Landau–Pollack theorem says: "The signal space of all finite-energy signals is infinite dimensional, but the subset of such signals that are band limited to $W$ Hz and approximately time limited to $[0, t_0]$, $t_0$ sufficiently large, is approximately finite dimensional with dimension $2Wt_0 + 1$."

This means that there exists a set of $2Wt_0 + 1$ orthonormal basis functions $\Psi_i(t)$, such that for any finite-energy signal $x(t)$ with energy $E_x$ that is band limited to $|f| < W$, for any constant $\varepsilon$ with $0 < \varepsilon < 1$, and for any $t_0$ sufficiently large, the following relations hold:

$$\int_0^{t_0} |x(t)|^2 dt > E(1 - \varepsilon)$$

$$\int_{-\infty}^{\infty} \left\{ x(t) - \sum_{i=0}^{2Wt_0} x_i \Psi_i(t) \right\}^2 dt < 12\varepsilon E_x$$

where $x_i$, $i = 0, 1, \ldots, 2Wt_0$ are the $2Wt_0 + 1$ expansion coefficients. In other words, if outside the interval $[0, t_0]$, the maximum fraction of the band-limited-signal's energy is $\varepsilon$, then this signal can be approximately expressed by a linear combination of a set of $2Wt_0 + 1$ orthonormal basis functions, with an energy error less than a fraction $12\varepsilon$ of the signal's energy. As $t_0$ gets larger, the fraction of energy outside the dimension $2Wt_0 + 1$ of this signal subspace becomes smaller.

To verify that the Landau–Pollack theorem is a consequence of the Nyquist–Shannon sampling theorem, we reason as follows [71]. Suppose that a signal is both band-limited and time-limited exists and that this signal is sampled at a frequency greater than the Nyquist frequency. These finitely many time-domain coefficients should represent the entire signal. In the same way, the entire spectrum of the band-limited signal should be represented through the finitely many time-domain coefficients resulting from the signal's sampling. Mathematically, this would require that a (trigonometric) polynomial can have infinitely many zeros, since the band-limited signal must be zero on an interval beyond a critical frequency that has infinitely many points. But we know from the fundamental theorem of algebra that a polynomial cannot have more zeros than its order. This contradiction is due to our incorrect assumption that a signal that is both band-limited and time-limited exists.

### 4.5.5 *Jayne's Maximum Entropy Principle*

The *maximum entropy principle* (**MEP**) was first formulated by Jaynes [26] and is actually a generic optimization problem, namely,

"Find a probability distribution $p_X(x)$, $x \in X$, that maximizes the Shannon's entropy $H(X)$ subject to a set of given constraints $c_1, c_2, \ldots, c_n$ which express

partial information about the probability distribution $p_X(x)$ sought, and also the typical axioms (constraints) of probability theory."

The most common constraints in practice are expected (mean) values of one or more random variables and/or random functions, or several marginal probability distributions of an unknown joint distribution. More specifically, this optimization problem is formulated as follows:

> Suppose we are given a random variable $X$ taking the values $\{x_1, x_2, \ldots, x_n\}$ where $n$ can be finite or infinite, and the mean (average) values of various functions $f_1(X), f_2(X), \ldots, f_m(X)$ where $m < n$. The problem is to determine the probability assignment $p_i = p(x_i)$ which satisfies the given data (constraints):
>
> $$\sum_{i=1}^{n} p_i = 1, \; p_i \geq 0$$
> $$\sum_{i=1}^{n} p_i f_k(x_i) = E[f_k(X)] = \Xi_K, \quad k = 1, 2, \ldots, m$$
>
> and maximizes Shannon's entropy:
>
> $$H(X) = -\sum_{i=1}^{n} p_i \log p_i$$

The solution to this mathematical problem can be determined using the well-known method of Lagrange multipliers, which however has the drawback that it does not make clear whether a true (global) maximum of $H(X)$ has been obtained. Here, without loss of generality, we will develop the solution for the case $f_k(X) = x_k$, in which the constraints about $f_k(X)$ are reduced to

$$\sum_{k=1}^{n} p_k x_k = E(X) = \xi$$

Also, for simplicity, the logarithm log in $H(X)$ is assumed to be the natural logarithm $(\log_e = \ln)$ [29]. We start the solution, by converting the above-constrained optimization problem into the equivalent unconstrained optimization problem, with respect to $p_k (k = 1, 2, \ldots, n)$ and the Lagrange multipliers $\lambda$ and $\mu$, of the Lagrange function:

$$L = -\sum_{k=1}^{n} p_k \ln p_k - \lambda \left( \sum_{k=1}^{n} p_k - 1 \right) - \mu \left( \sum_{k=1}^{n} p_k x_k - \xi \right)$$

Then, we write down the associated canonical (partial derivative) equations:

$$\partial L/\partial p_k = -\ln p_k - 1 - \lambda - \mu x_k = 0,\ k = 1, 2, \ldots, n$$
$$\partial L/\partial \lambda = 1 - \sum_{k=1}^{n} p_k = 0$$
$$\partial L/\partial \mu = \xi - \sum_{k=1}^{n} p_k x_k = 0$$

The first $n$ equations can be rewritten as

$$\begin{aligned} p_1 &= \mathrm{e}^{-1-\lambda-\mu x_1} = \mathrm{e}^{-(1+\lambda)}\mathrm{e}^{-\mu x_1} \\ p_2 &= \mathrm{e}^{-1-\lambda-\mu x_2} = \mathrm{e}^{-(1+\lambda)}\mathrm{e}^{-\mu x_2} \\ &- - - - - - - - - - - \\ &- - - - - - - - - - - \\ p_n &= \mathrm{e}^{-1-\lambda-\mu x_n} = \mathrm{e}^{-(1+\lambda)}\mathrm{e}^{-\mu x_n} \end{aligned}$$

which, if divided by the sum $p_1 + p_2 + \cdots + p_n = 1$, gives

$$p_k = \mathrm{e}^{-\mu x_k} / \sum_{i=1}^{n} \mathrm{e}^{-\mu x_k},\ k = 1, 2, \ldots, n$$

Now, multiplying both sides of the $p_k$ equation by $x_k$ and adding, we get

$$\xi = \sum_{k=1}^{n} x_k \, \mathrm{e}^{-\mu x_k} / \sum_{k=1}^{n} \mathrm{e}^{-\mu x_k}$$

from which we obtain

$$\sum_{k=1}^{n} x_k \, \mathrm{e}^{-\mu x_k} - \xi \sum_{k=1}^{n} \mathrm{e}^{-\mu x_k}$$

Finally, multiplying this equation by $\mathrm{e}^{\mu\xi}$ gives

$$\sum_{k=1}^{n} (x_k - \xi) \, \mathrm{e}^{-\mu(x_k - \xi)} = 0$$

This is a nonlinear equation (with respect to $\mu$) and can be solved numerically for $\mu$. Introducing this value of $\mu$ into the $p_k$ equations $(k = 1, 2, \ldots, n)$, we find the desired probabilities $p_k,\ k = 1, 2, \ldots, n$.

As a simple illustration of the *maximum entropy principle*, let us consider an unbiased ("honest") die. Here, $x_k = k$ $(k = 1, 2, 3, \ldots, 6)$, and so $\xi = (1 + 2 + 3 + 4 + 5 + 6) / 6 = 21 / 6 = 3.5$. The last equation here becomes

$$-2.5e^{2.5\mu} - 1.5e^{1.5\mu} - 0.5e^{0.5\mu} + 0.5e^{-0.5\mu} + 1.5e^{-1.5\mu} + 2.5e^{-2.5\mu} = 0$$

The solution of this equation is $\mu = 0$, and so we find

$$p_k = 1/6 \text{ for } k = 1, 2, \ldots, n$$

*Remark* In this example, the mean value $\xi = E(X)$ was known and used as a constraint on the corresponding probabilities. If $E(X)$ were not known, then no constraint would be on the probabilities. In this case, the maximum entropy principle would give equal probabilities (the uniform or equiprobable distribution $p_x(x)$), which are the only probabilities for which the entropy takes its absolute maximum value (see the properties of the entropy in Sect. 4.5.2.2). If the probabilities are subject to constraints, the MEP principle gives a maximum entropy of limited value, which is usually smaller than the entropy of the uniform distribution.

## 4.6 Concluding Remarks

This chapter is the first of three chapters that deal with the "*information pillar*" of human life and society. The topics that have been covered include the general definition of the information concept, the historical landmarks of its manifestations, namely, communication (speech, writing, printing, telegraph, telephone, computing, and computers), information theory, computer networks, multimedia, telematics, and informatics, and a technical review of communication and information theory.

Information storage, flow, and processing are inherent processes in nature and living organisms. Organisms do not live separately from other organisms but interact with each other within the ecosystem in which they exist. Many of the existing technological communication and modulation/demodulation models were inspired by life on Earth, or they can be used to model and better understand biological communication models. The exploration of the physical and biological world provides unexpected surprises and discoveries that increase steadily our information entropy.

In a similar way, modern information transmission and communication techniques are affecting and will continue to increasingly affect the social and economic/business activity of people over coming decades. An example of this is electronic commerce (e-commerce), which reduces substantially the sales and operational costs in comparison with the traditional stores. The Internet expands and moves e-commerce into an open and global market with big changes in the market structure, size, and players.

More generally, inter-networked information technology enables people to act in their own self-interests and affect the experiences of other people. Overall, it is anticipated that the prospect of *return-on-investment* (**ROI**) research in communications, information theory, and information technology is very promising, with directly measurable results in economic strength and social benefits.

## References


1. R. Capurro, Information: *A Contribution to the Foundation of the Concept of Information Based on Its Etymology and in the History of Ideas* (Saur, Munich, 1978). http://www.capurro.de/info.html
2. R. Capurro, B. Hjorland, The concept of information. Ann. Rev. Inf. Sci. Technol. **37**, 343–411 (2003). (B. Cronin, ed., Chap. 8)
3. R.J. Bogdan, *Ground for Cognition: How Goal-Guided Behavior Shapes the Mind* (Lawrence Earlbaum, Hillsdale, N.J., 1994)
4. G. Ropohl, The concept of information in the framework of the culturalist struggle. Ethik und Sozialwissenschaften **1**, 1–12 (2001)
5. W. Hofkirchner (ed.), The quest for a unified theory of information, in *Proceedings, 2nd International Conference on the Foundations of Information Science* (Gordon and Breach, Amsterdam, 1999)
6. F. Machlup, U. Mansfield (eds.), *The Study of Information: Interdisciplinary Messages* (Wiley, New York, 1983)
7. R. Capurro, On the genealogy of information, in *Information: New Questions to a Multidisciplinary Concept*, ed. by K. Kornwachs, K. Jacoby (Akademie Verlag, Berlin, 1996), http://www.capurro.de/cottinf.hbm
8. R.V.L. Hartley, Transmission of Information. Bell Syst. Tech. J. **7**, 535–563 (1928)
9. C. Shannon, A mathematical theory of communication. Bell System Tech. J. **27**, 379–423, 623–656 (1948)
10. C. Shannon, W. Weaver, *The Mathematical Theory of Communication* (The University of Illinois Press, Urbana, IL, 1949) (also 1972)
11. H. Titze, *Is Information a Principle? (Ist Information ein Prinzip?)* (Hain, Meisenheim, 1971)
12. T. Stonier, Towards a new theory of information. J. Inf. Sci. **17**, 257–263 (1991)
13. T. Stonier, Information as a basic property of the universe. Biosystems **38**, 135–140 (1997)
14. P.T. de Chardin, *The Future of Man* (Harper and Row, New York, 1964) (originally published in 1959)
15. S. Augarten, *BIT by BIT: An Illustrated History of Computers* (Ticknor and Fields, New York, 1984)
16. R. Moreau, *The Computer Comes of Age: The People, the Hardware, and the Software* (Trans. by J. Howlett) (MIT Press, 1984)
17. J.G. Butler, Information Technology History-Outline, http://www.tcf.ua.edu/AZ/ITHistory.Outline.htm
18. The Antikythera Mechanism Research Project, www.antikythera-mechanism.gr/project/general/the-project.html
19. Information Theory, http://www.eoht.info/page/information+theory
20. History of Information Theory, http://en.wikipedia.org/wiki/History_of_information_theory
21. L. Szilard, On the decrease in entropy in a thermodynamic system by the intervention of intelligent beings. Z. Angew. Phys. **53**, 840–856 (1929)
22. L. Brillouin, *Science and Information Theory* (Academic Press, New York, 1956)
23. R.M. Fano, *Transmission of Information* (MIT Press, Cambridge, MA, 1949)

24. E.T. Jaynes, Information theory and statistical mechanics: (I). Phys. Rev. **106**, 620 (1957); (II), ibid., **108**, 171 (1957)
25. M. Tribus, *Thermostatics and Thermodynamics: An Introduction to Energy, Information, and States of Matter* (Van Nostrand Reinhold, New York, 1961)
26. E.T. Jaynes, On the rationale of maximum entropy methods. Proc. IEEE **70**(9), 939–952 (1982)
27. G.N. Saridis, Entropy Formulation for Optimal and Adaptive Control. IEEE Trans. Autom. Control **33**(8), 713–721 (1988)
28. G.N. Saridis, *Entropy in Control Engineering* (World Scientific, London, 2001)
29. G.J. Klir, *Uncertainty and Information: Foundations of Generalized Information Theory* (Wiley-Interscience, Hoboken, NJ, 2006)
30. A. Ben-Naim, *Farewell to Entropy: Statistical Thermodynamics Based on Information* (World Scientific, London, 2007)
31. Wikipedia: Timeline of Information Theory, http://en.wikipedia.org/wiki/Timeline_of_information_theory
32. History of Computer Networks (Hammed Haddadi). http://www.ee.ucl.ac.uk/∼hamed/transfer/thesis/node10.html
33. Computer Networking History-Networking Computer Tips, http://www.onlinecomputertips.com
34. Internet World Stats, http://www.internetworldstats.com/stats.htm
35. R. Albarino, Goldstein's Light Works at Southampton, Variety **213**(12) (1966) (Aug 10)
36. D. Marshall, History of multimedia systems, http://www.cs.cf.ac.uk/Dave/Multimedia/node8.html
37. History of Multimedia (Calgary University) http://www.acs.ucalgary.ca/∼edtech/688/hist.html
38. S. Nora, A. Minc, *The Computerization of Society. A Report to the President of France* (MIT Press, Cambridge, MA, 1980) (translation of *L' Informatisation de la Société.* Rapport à M. le Président de la Republique, La Documentation Francaise, Paris, 1978)
39. The History of Telematics. Telematique, http://www.telematique.eu/telematics/
40. Telematics Valley, http://telematicsvalley.org/
41. C.D. Mortensen, *Communication: The Study of Human Communication* (McGraw-Hill, New York, 1972)
42. J.R. Barry, E.A. Lee, D.G. Messerschmidt, *Digital Communication* (Kluwer, Boston, 2004)
43. D.K. Berlo, *The Process of Communication* (Holt Rinehart and Winston, New York, 1960)
44. W. Schramm, How Communication Works, in *The Process and Effects of Communication*, ed. by W. Schramm, (University of Illinois Press, Urbana, Illinois, 1954)
45. Communication models, http://www.shkaminski.com
46. B.G. Korenev, *Bessel Functions and their Applications* (CRC Press, Boca Raton, FL, 2002)
47. E. Weisstein, Bessel Function of the First Kind, Wolfram Research, http://mathworld.wolfram.com/BesselFunctionoftheFirstKind.html
48. D.E. Foster, S.W. Sceley, Automatic tuning, simplified circuits, and design practice. Proc. IRE **25**(3), 289–313 (1937) (part 1)
49. FM Detectors (Discriminators), http://www.ycars.org/EFRA/Module%20B/fmdet.htm
50. Amplitude, Frequency and Phase Modulation, http://www.aubraux.com/dsp/dsp-modulation-all.php
51. A. Schwope, Modulation, http://www.andreas-schwope.de/ASIC_s/Schnittstellen/Data_Lines/body_modulation.html
52. J. Proakis, M. Salehi, *Digital Communications* (Mc Graw-Hill, New York, 2007)
53. H.L. Resnikoff, *The Illusion of Reality* (Springer, Berlin, 1989)
54. W.W. Reterson, E.J. Weldon Jr., *Error Correcting Codes*, 2nd edn. (The MIT Press, Cambridge, MA, 1972)
55. V. Press, *Introduction to the Theory of Error-Correcting Codes*, 3rd edn. (Wiley, New York, 1998)

56. W. Huffman, V. Pless, *Fundamentals of Error-Correcting Codes* (Cambridge University Press, Cambridge, 2003)
57. J.H. Van Lint, *Introduction to Coding Theory* (Springer, Berlin, 1981)
58. N.J.A. Sloane, *A Short Course on Error Correcting Codes*, vol. 188, CISM Courses and Lectures (Springer, Berlin, 1975)
59. A.M. Michelson, A.H. Levesque, *Error Control Techniques for Digital Communication* (Willey, New York, 1985)
60. C. Schlegel, L. Perez, *Trellis Coding* (IEEE Press, New York, 1997)
61. Error Correcting Codes, http://www.eie.polyu.edu.hk/∼enmzwang/adc/1-notes/node3.html
62. C. Langton (ed.), Convolution Codes, http://www.complextoreal.com/convo.htm
63. Y.S. Han, *Introduction to Binary Convolution Codes*, Graduate Institute of Communication Engineering, National Taipei University, Taiwan. http://web.ntpu.ed.tw/∼shan/intro_in_cod.pdf
64. C. Fleming, A Tutorial on Convolutional Coding with Viterbi Decoding. http://home.netcom.com/∼chip.f/viterbi/tutorial.html
65. F.J. Mc Williams, N.J.A. Sloane, *The Theory of Error-Correcting Codes* (North-Holland, Amsterdam, 1977)
66. R. Yates, A Coding Theory Tutorial, http://www.digitalsignallabs.com/tutorial.pdf
67. H. Nyquist, Certain topics in telegraph transmission theory. Trans. AIEE **47**, 617 (1928)
68. C. Shannon, Communication in the presence of noise. Proc. IRE **37**(1), 10–21 (1949). http://www.stanford.edu/class/ee104/shannonpaper.pdf
69. R. Wang, Shannon's source coding theorem (Dec 10, 2009). http://fourier.eng.hmc.edu/e161/lectures/compression/node7.html
70. B. Aazhang, Shannon's Noisy Channel Coding Theorem (May 19, 2004). http://cnx.org/content/m10180/latest/
71. Nyquist-Shannon Sampling Theorem Signal Frequency Mhz, http://www.economicexpert.com/a/Nyquist:Shannon:sampling:theorem.htm

# Chapter 5
# Information II: Science, Technology, and Systems

*Information technology and business are becoming inextricably interwoven. I don't think anybody can talk meaningfully about one without talking about the other.*

Bill Gates

*Information and communications technology unlocks the value of time, allowing and enabling multitasking, multichannels, multi-this, and multi-that.*

Li Ka Shing

**Abstract** This chapter is devoted to three modern subfields of information, namely information science, information technology, and information systems. These fields have an enormous impact on modern society and its development. Information science is generally concerned with the processes of storing and transferring information via the merging of concepts and methodologies of computer science, linguistics, and library science. Information technology (IT) or "infotech" covers all methodologies and technologies which are used for the production, storage, processing, transmission, and dissemination of information. Information systems use information science and information technology concepts and tools in the everyday operation of enterprises and organizations that needs the cooperation (symbiosis) of technology with human-controlled processes and actions. This chapter starts with a discussion of the fundamental general issues of information science including several classification schemes (knowledge maps), and continues with a guided tour to computer science, computer engineering, internet/www, and web-based multimedia. Finally, this chapter provides a general discussion of information systems which include their fundamental concepts, general structure, types, and development.

**Keywords** Information · Information science · Information technology
Information systems · Computer science · Scientific computing
Data bases · Thesauri · Programming languages · Artificial intelligence
Computer engineering · Operating systems · Parallel computing
Information processing · Telecommunications · Documentation
Computer network · Web-based multimedia · System development cycle



S. G. Tzafestas, *Energy, Information, Feedback, Adaptation, and Self-organization*,
Intelligent Systems, Control and Automation: Science and Engineering 90,
https://doi.org/10.1007/978-3-319-66999-1_5

## 5.1 Introduction

This chapter is a continuation of Chap. 4 which was devoted to the conceptual and technical issues of communication systems and transmission of information. Specifically, the present chapter provides a conceptual and technical tour of three important cornerstones of the "information human-life-and-society pillar," i.e., *information science*, *information technology*, and *information systems*. Information technology (**IT**) is the dominating branch, which is gradually becoming an "*umbrella*" that covers all the theoretical and practical issues of information processing (computing), including hardware architectures, software techniques, intelligent systems, knowledge engineering, embedded systems, multimedia, telecommunication networks, computer networks, Internet/web-based processes, signal/image processing, etc.

The structure of the chapter is as follows. Section 5.2 discusses the general issues of *information science* accompanied by some classification schemes (knowledge maps) of the relevant building blocks. Section 5.3, which is the main body of the chapter, provides a guided tour of the following ingredients of **IT**: *computer science* (theoretical CS, scientific computing, data bases, programming languages, artificial intelligence); *computer engineering* (hardware architectures, operating systems, parallel computing, software engineering, embedded systems); and *telecommunications* (telematics, cellular systems, computer networks, Internet/www, web-based multimedia,), and, finally, Sect. 5.4 deals with *information systems* (general concepts, general structure, types, development of IS).

All the above fields are extremely vast, and, therefore, the size of the present book does not allow a detailed technical presentation of the techniques and technologies involved. As a result, the chapter provides, in most cases, the definitions of the various concepts with brief descriptions of why they are needed and their role. But the material included is indeed sufficient for the conceptual and global purpose of the book. Comprehensive treatments are given in the textbooks and research books are referenced.

## 5.2 Information Science

*Information science (IS)* is the scientific field that is generally concerned with the processes of storing and transferring information through the merging of concepts and methodologies of computer science and engineering, linguistics, and library science [1, 2]. The specific purposes of information science are the collection, organization, storage, retrieval, interpretation, and use of information. The roots of IS come from *information theory* (Shannon), *cybernetics* (Wiener), and *electronic computers*. The transfer of information over time requires the availability of a *document* (which involves some storage medium). The use of documents is what is called *documentation*.

Actually, the field of IS is continuously changing, a fact that drives the IS scientists to regularly review and/or redefine basic building blocks. Information/

knowledge mapping constitutes an essential element for the construction, learning, and dissemination of knowledge. In the literature, one can find a huge amount of information/knowledge maps, frequently incomplete, inconsistent, and unsystematic. Such maps can be found in existing library classification schemes, classification of bibliographic resources, information services databases, and *thesauri* [3–7].

A comprehensive study of the systematic construction of knowledge maps of IS was recently (2007) conducted under the name *Critical Delphi Study*, which developed a qualitative research methodology for facilitating moderated, critical discussions among a group of 57 experts (called the panel) from 16 countries [8]. The indirect discussions were anonymous and were conducted in three successive rounds of structured questionnaires. Statistical details on the responses to these questionnaires and the revision of the responses are provided in [8] and the other related articles concerning the Critical Delphi study (e.g., [9]).

The *Delphi Method*, in general, was developed to assist a group of experts to arrive at an overall position regarding an issue under investigation. Full information on the Delphi method is available in [10], where the scope and objectives are described, along with course notes and a list of Delphi studies and applications. These applications include the topic of forecasting in transportation, the field of agricultural education, and the health-field sector. Very briefly, the Delphi method is based on a series of repeated questions (via appropriate questionnaires) posed to a group of experts whose opinions and judgments are important. After the initial interrogation of each expert, each subsequent interrogation includes information about the replies from the previous round, typically presented anonymously. So each individual expert has the opportunity to reconsider and possibly modify his/her previous reply in light of the replies of the other members of the group. The group's position is formed after two or three rounds by averaging. The Delphi method is typically conducted asynchronously through paper and mail, but it can also be executed via teleconferencing.

In the above mentioned *Critical Delphi Study* concerning the subject of *information science*, 28 panel members offered their reflections and classification schemes (maps) of IS, which are presented in [8]. Here, we present the first three of them provided by *Aldo Barreto*, *Shifra Baruchson-Arbib*, and *Clare Beghtol*.

**Aldo Barreto IS Map**

- **Information Production and Organization**

(i.e.,: Information nature, qualities, and value; Production of stocks of information; Information management and control; Technologies and practices of information).

- **Information Distribution**

(i.e.,: Users and information communities; Communication of information; Information sources; Channels of information and its flow).

- **Information Consumption and Use**

(Information availability and access; Information uses and applications; Cognition aspects of information; Assimilation of information; the production of knowledge).

- **History, Philosophy, Legal, Ethics, and Ancillary Aspects of Information**

(Legal structure of information, i.e.,: Copyright; Ethics of information; Policy and politics; Globalization aspects; History, philosophy, and environment).

*Aldo Barreto* says: "In my view, IS is a set of flows, processes, and actions that starts in the generator's (author's) mind and ends in a space where users (receptors) appropriate that information as knowledge. As it is a dynamic model, I cannot see a static table where headers do not match the whole idea" [8].

**Shifra Baruchson-Arbib IS Map**

- **Foundations of IS**

(History of IS; History of librarianship; Archival science; History of knowledge formats: Manuscripts, print and digital; IS epistemology).

- **Methodology**

(i.e.,: Quantitative and qualitative research; Bibliometrics and informatics; Bibliology; Domain Analysis; Webometrics).

- **Information/Learning Society**

(Social and cultural aspects of the information society; Sociology of knowledge; Social communication; Scientific communication; E-learning; Information literacy; IS Education; Lifelong learning).

- **Information Technology**

(Communication and Computer networks; Document delivery systems; Structure of computerized systems; Programming languages; Multimedia; Information retrieval systems; Systems analysis; Artificial intelligence; Human–computer interaction; Information architecture; Digital security systems; Websites construction; Network technologies; Knowledge representation; Search tools).

- **Data Organization and Retrieval**

(Classification schemes; Metadata; Indexing; Abstracting; Knowledge organization; Taxonomies; Thesauri; Ontology; Vocabulary control; Online-searching techniques; Reference work; The semantic web).

- **Information Industry Economy and Management**

(Competitive intelligence; Databases; Digital Libraries; Electronic publishing; Information industry market; Information management; Information Manipulation; Knowledge management; Information centers and libraries management; Collection management; Electronics comers).

- **Information Ethic and Law**

(i.e.,: Copyright; Digital security; Digital divide; Censorship; Internet crime; Free access to information; Information policies).

- **User studies**

(i.e.,: Human information behavior; Information seeking behavior; Information needs; Reference interview; User-information scientist-interaction).

- **Diffusion Studies**

(i.e.,: Information dissemination; Communication theory; Message theory; Information centers and Libraries).

- **Social Information Science**

(Information needs of various cultures; Information education, power, and ethics; Social information banks; Social information sections in school and public libraries; Self-help sources (printed and electronic); Social information scientist; Community information; Information diffusion in multicultural societies; Health information centers).

**Clare Beghtol IS Map**

- **People**

(*By group*: Community; Culture; Domain; User group; *By individual*: Researcher; User).

- **Object of Study**

(*By element*: Data; Information; Knowledge; Message; *By conceptual foundation*: Epistemology; History; Philosophy; Practice(s); Theory; *By purpose*: Communication; Creation; Discipline area; Dissemination; Evaluation; Management; Organization; Representation; Retrieval; Search; Storage; *By methodology*: Qualitative; Quantitative).

- **Systems**

(*By cultural factor*: Economic aspects; Education; Ethical aspects; Legal aspects; Professions; Societal aspects; *By technology*: Electronic; Manual; Mechanical).

- **Space**

(By Universal Decimal Classification).

- **Time**

(By Universal Decimal Classification).

*Clare Beghtol states*: "It is interesting that no one has produced a faceted (analytico-synthetic) system, so I've provided the basis for one… The fundamental facets are People, Object of Study, Systems, Space, and Time. These are subdivided into sub-facets and foci at a general level. It would need further conceptual development for sub-facets and foci, and a synthetic notation that would allow both inter-and intra-facet synthesis. The Universal Decimal Classification has been chosen for subdividing Space and Time because it is more highly developed in those areas than other general systems [8]."

It is clear from these three maps of IS, that the field of IS is very wide and includes a large repertory of subfields and topics. Detailed discussions of these topics, which are out of the scope of the present book, can be found in [2–9] and the references cited therein.

One of the basic topics of information science is "*information taxonomy,*" which is the first step in the design of efficient software architectures from the ground up (http://articles.techrepublic.com/5100-10878_11-5055268.html). Thomas H. Davenport and Laurence Prusac define *information architecture* as follows: "Information architecture, in the broadest sense, is simply a set of aids that match information needs with information resources" and involves the *standards and guidelines element*, plus the following five elements:

**Information usage patterns** These patterns determine how information is used and the information flows within a system (enterprise, organization, etc.).

**Information access points** These access points are the various channels through which the information needed is located and extracted.

**Taxonomy** This is the core element of information architecture, which interconnects all the other elements and guides the information flow and management so as to meet the relevant standards and guidelines.

**Administration structure** This structure helps the internal users to create and maintain information through a proper visual way in accordance with the content-management flow.

**Information flow** It describes the paths of information flows within the system.

An illustration of how the above six elements are organized and interconnected in the information architecture is shown in Fig. 5.1.

The *taxonomy structure*, a term typically used in life sciences for categorizing animals and plants, is actually a conceptual framework and not a product, and, when implemented on the World Wide Web, it provides an effective structuring of content to ensure easy and accurate access.

Taxonomy can be implemented in two ways, namely: (i) as a *classification system*, *thesauri* and *controlled vocabulary*, and (ii) as a *directory listing* or *website map*.

The taxonomy puzzle can be split into two parts:

- *Taxonomy structure* This is similar to standard classification systems (e.g., those used in life sciences or libraries).

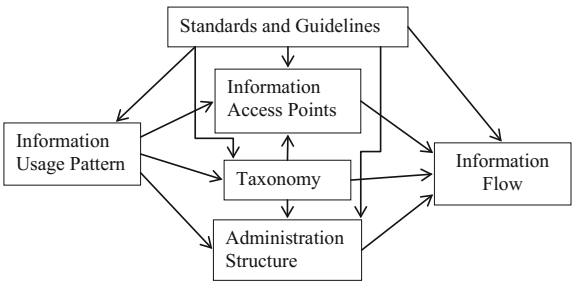



**Fig. 5.1** The information architecture

- *Taxonomy view* This is the visual aspect of taxonomy structure, presenting web content grouped according to topics and forming an easy to use sitemap. See, e.g., the Yahoo directory (www.yahoo.com).

## 5.3 Information Technology

*Information technology* (**IT**) is a modern term that embraces a wide variety of interrelated information processing and computing fields which were developed almost concurrently and have arrived today at very high levels of maturity and development. These fields are:

- Computer science.
- Computer engineering.
- Telecommunications.

Some authors include in IT the field of signal and image processing, which is closely related to IT, but typically it is studied as a separate field. A compact definition of IT that covers all the above subfields is the one given by the *Information Technology Association of America* (**ITAA**). This definition states: "IT is the study, design, development, implementation, support or management of computer-based information systems, particularly software applications and computer hardware [11]."

In the following, we will present a brief outline of these subfields sufficient for the purposes of this book.

### *5.3.1 Computer Science*

#### 5.3.1.1 General Issues

*Computer science (CS) seeks* to provide a scientific basis for the study of information processing, the solution of problems by algorithms, and the design and programming of computers [12].

Martin Davis and Elaine Weyke state that: "Theoretical computer science is the mathematical study of models of computation" [13].

Computer science was first recognized as a distinct scientific field in the 1950s and early 1960s, and, in the present days, the use of computers has shifted from experts or professionals to the much wider class of users. The principal achievements of computer science include (but are not restricted to) the following [14–16]:

- The theoretical development of computation and computability, the major result being the proof that there exist problems that are computationally intractable.
- Development of the algorithmic approach and the concept of a programming language.

- The scientific computing concept (computer solution of scientific problems) through the use of numerical mathematical techniques.
- The development of intelligent systems via the use of artificial intelligence and knowledge-based techniques.

The computer science field is now a well-established field of study and is included in the curricula of most universities worldwide [17]. More specifically, computer science has the following main subfields:

- Theoretical computer science
- Scientific computing
- Computer programming and languages
- Artificial intelligence and knowledge-based systems

#### 5.3.1.2 Theoretical Computer Science

*Theoretical computer science (TCS)* includes both the traditional computation theory and the more recent complex and abstract topics of computing. The two basic questions that are studied by TCS are:

- What is computable?
- What resources are needed to carry out these computations?

The first question is the particular subject of the so-called *computability theory*, and the second question is addressed by the *theory of computational complexity*. Computability theory investigates the solvability of theoretical and practical problems using several computational models, whereas the computational complexity theory examines the computational (time and memory) costs of the techniques developed for solving these problems. The well-known **P = NP** (*P versus NP*) problem is still an unsolved (open) problem of computation theory [18, 19]. Of course, since Steve Cook presented his seminal *NP-completeness paper*: "The Complexity of Theorem-proving Procedures" in Shaker Height, OH (May, 1971) [20], the computational power of digital computers has vastly increased, and the cost of computing has dramatically decreased. As we solve larger and more complex problems with better algorithms and greater computational power, the problems that we cannot tackle start to stand out. Here the theory of *NP-completeness* (**NP-C** or **NPC**) helps in understanding the existing limitations, and the P versus NP problem emerges both as an important theoretical question in computational complexity theory and as a fundamental principle that permeates all sciences. It should be noted that the standard model in computability is the *Turing machine* (introduced in 1936), which, although it was introduced before computers were built physically, it was continuously recognized as the suitable computer model for the purpose of defining the concept of a "computable" function.

Informally, the class **P** (*Polynomial Time*) is the class of decision problems solvable by some algorithm in a finite number of steps bounded by some fixed

polynomial in the length of input. The class **P** is robust and has equivalent definitions over a large repertory of computer models. Formally, in terms of Turing machines, the elements of the class **P** are languages.

Many problems in practice do not seem to have such an efficient algorithm. However, all these problems have a common feature: Given a potential solution, we can validate that solution efficiently. The collection of problems that have an efficiently verifiable solution is called **NP** (*Non-deterministic Polynomial-Time*). Therefore, **P = NP** means that, for every problem that has an efficiently verifiable solution, we can find that solution efficiently as well. The very hardest NP problems (e.g., partition into triangles, Hamiltonian cycle, and three-coloring) are called "**NP-complete**" problems, i.e., given an efficient algorithm for one of them, one can find an efficient algorithm for all of them, and in fact any problem in NP. The majority of computer scientists arrived at the conclusion that $\mathbf{P} \neq \mathbf{NP}$, and their efforts to prove it has been recognized very quickly as the single most important question in theoretical computer science, plus an important computational issue in almost all scientific disciplines. Examples of problems are:

- Finding a DNA sequence that best fits a collection of fragments.
- Finding the Nash equilibrium with given properties in a number of environments.
- Finding optimal protein-threading processes.

The first problem that was shown to be NP-complete was the *Boolean satisfiability problem,* and the relevant theorem is known as the *Cook-Levin Theorem*. Many computer scientists look at the negative, that if P = NP, then public-key cryptography becomes impossible. Three comprehensive books on the subject are [21–23]. These books provide full treatments of the theory of complexity and algorithms including such mathematical topics as logic, number theory, probability theory, and combinatorial theory.

#### 5.3.1.3 Scientific Computing

*Scientific computing (SC)* is a very broad subfield of computer science devoted to the study and construction of mathematical models and numerical techniques for solving scientific problems using as a tool the computer. The standard and classical topics covered by scientific computing are presented and studied in textbooks and research books of the field, good examples of which are the books of X. Yang, W. H. Press, M.T. Heath, and R.E. Crandall [24–27]. A dominant place in the field is held by computational physics, computational chemistry, and computational biology. The principal topics of scientific computing are [26]:

- Approximation theory.
- Computing arithmetic.
- Solution of systems of linear algebraic equations.
- Linear least squares and model fitting.

- Eigenvalue theory.
- Nonlinear algebraic equations.
- Numerical optimization algorithms.
- Interpolation and extrapolation.
- Numerical integration and differentiation.
- Initial-value and boundary-value problems of ordinary differential equations.
- Numerical solution of partial differential equations.
- Fast Fourier transform.
- Random numbers and stochastic simulation.

In [27], Crandall includes the following advanced research topics of scientific computation:

| | |
|---|---|
| • Fast Fourier transform | • Chaos |
| • Wavelets | • Fractals |
| • Prime numbers | • Fast arithmetic |
| • N-body problems | • Sound synthesis |
| • Zeta functions | • Matrix algebra |
| • Factoring | • Convolution |
| • Image processing | • Genetic algorithms |
| • Monte Carlo simulation | • Compression |
| • Nonlinear systems | • Polynomial methods |

A useful practical set of educational modules, written in Java, that illustrate the basic concepts and algorithms of scientific computing is provided in [28]. These modules support the material given in [26] but can also be used in conjunction with any textbook on scientific computing. Each module is accessible through a web browser, the problem data can be selected interactively, and the results are provided both numerically and graphically. Scientific computing is supported by many programming languages and tools including C, FORTRAN, MATLAB, MATHEMATICA, PYTHON, GNU Octave, and algebra libraries BLAS and LAPACK. Web links for them can be found in [29] and other sites [30–33].

A modern branch of scientific computing is the so-called "*high-performance scientific computing*" or "*parallel scientific computing*" which aims at improving the quality of computing (accuracy, speed) through the use of parallel computers and high-performance computer systems (supercomputers). A useful educational book on this field is [34].

#### 5.3.1.4 Data Structures and Databases

*Data structure* is called a way of representing data for efficient storage and organization in a computer. Data structures are based on the capability of computers to bring and store data at any place in their memory, specified by an address. Actually, over the years, various forms of data structures have been developed that are appropriate for particular applications. The implementation of a data structure is performed by writing a program that creates and manipulates instant ions of this

structure. The efficiency of a data structure is closely related to the procedures and operations of the computer program with which it is associated and implemented. For this reason, many formal software-design methods and programming languages focus on data structures rather than the algorithms of the design [35].

In their book, M.T. Goodrich and R. Tamassia present the implementation of data structures and algorithms using Java [36]. After an introduction to Java programming and object-oriented design, they deal with the following data structures and topics:

- Analysis tools.
- Arrays and linked lists.
- Stacks and queues.
- Lists and iterators.
- Priority queues.
- Maps and dictionaries.
- Search trees.
- Sorting, sets, and selection.
- Text processing.
- Graphs.

A *database (DB)* is defined as a structured collection of records or data which is stored in a computer, such that one can consult it via a program to get answers to queries. Databases are classified into several types, according to the scope of the data or the goal of storage. Many databases provide access rights that permit the user to perform various operations on the data such as update, edit, delete, and so on. To this end, databases involve software-based containers that perform the collection and storage of information in such a way that the user can retrieve, add, update, or delete information in an automated way.

The types of databases classified according to their purpose are the following:

- *Analytical databases* (Static read-only databases that allow the user to analyze enterprise data about personnel, sales, marketing, etc., depending upon a selected management policy).
- *Data warehouse* (A large collection of data extracted from several other databases which can be screened and edited or standardized and become available to managers and other staff members of an organization).
- *Operational or subject-area databases* (Databases involving information required for the operations of an organization, such as information about a particular member, subject or department).
- *Distributed databases* (Databases of an organization distributed at various local branches or offices. They can be either common to all sites or they can be specific for a local site only, or of combined common/specific type).
- *External databases* (Databases that refer to online access to external, privately owned data, e.g., wealth data about a person, directory data, library data, etc.).

- *End-user databases* (Databases developed by end-users at their own computers, such as spread sheets, word documents, or downloaded files).
- *Hypermedia databases* (A set of interconnected multimedia web pages. The relevant information is stored online and the data—text, graphics, still pictures, video, audio, etc.—can be accessed by many users simultaneously).

The database types classified by the scope of data are the following:

- *General-interest databases* (Databases that provide information about large numbers of fields and topics and are mainly used for research purposes).
- *Discipline-specific databases* (Here the information is restricted to particular fields, and they are useful to the professionals in these fields).
- *Subject-specific databases* (These databases are focused on a particular subject only. They are mainly used by academic people, and involve papers from journals and conferences).

The principal working models of databases are the following:

- Hierarchical.
- Network.
- Relational.
- Object relational.
- Post-relational.

The relational database is now the standard of business computing and employs the *table* data structure which allows easy searching. The object-database model is frequently used in spatial, engineering, and telecommunication databases. This model is the result of merging object-oriented programming with database technology. Post-relational databases employ more general data models than the relational model, e.g., extending relational database systems with non-relational properties or allowing the representation of a directed graph with trees on the nodes.

The control of the creation, maintenance, and use of a database of an organization is performed by the so-called *database management system* (**DBMS**). A DBMS is actually a software program that organizes the storage, modification, and extraction of information from a database. Today, we have available for use many different types of **DBMSs,** ranging from small systems that run on personal computers to very large systems operating on mainframes.

Some examples of database applications are:

- Library-database systems.
- Flight-reservation systems.
- Large inventory systems.
- Medical-database systems.

Technically a **DBMS** depends on the way it organizes the information internally. For example, we have relational DBMS, network DBMS, hierarchical DBMS, etc.

The requests for information from a database are made by *queries* that have a stylized form.

*A relational DBMS* **(RDBMS)** has the following components:

- Interface driver(s).
- SQL (Structured Query Language) engine.
- Transaction engine.
- Relational engine.
- Storage engine.

Likewise, an object *DBMS* (**ODBMS**) has the following components:

- Language driver(s)
- Query engine
- Transaction engine
- Storage engine

Three very popular databases available in the open market are: **Oracle, MySQL,** and **DB2** [37–40]. The data in an Oracle database (licensed by Oracle Corp.) are handled by an Oracle DBMS, which is a RDBMS.

*Oracle Corporation* itself blurs the distinction between data managed by an Oracle RDBMS, an Oracle database, and the Oracle RDBMS software itself. MySQL is an extremely popular open-source database, selected by a wide range of database developers. It is noted that organizations and enterprises are primarily looking to open databases for supporting new applications (e.g., Web 2.0 applications, small portal applications, etc.). DB2 has a long history and is recognized to be the first DB product using SQL. It was introduced by IBM in 1982 when IBM released SQL/DS and DB2 on its mainframe platform. DB2 has its origin at the early 1970s when the IBM scientist E.F. Codd developed the theory of relational databases and published the model for data manipulation. Originally, DB2 was exclusively available on IBM mainframes, but later it was brought to other platforms (UNIX, Windows and Linux servers, and PDAs). Another family of DBMS products from IBM is called *Informix,* which was first released in 1986 and, during the 1990s, was the second most popular DB system, after Oracle. Other databases currently available are: *Microsoft Access, Postgre SQL, System 2000, Ingress,* and *Sybase*. Extensive information on data structures, databases, and DBMS can be found in [41, 42].

#### 5.3.1.5 Computer Programming and Languages

**Computer Programming**

*Computer programming* is writing instructions used by a computer for problem-solving purposes. The principles of computer programming are generic and can be used independently of the programming language employed. Computer programming can be done in one of a large variety of languages, which range from low-level machine code (known as microcode) to higher level languages.

According to their size, computer programs are classified into the following three categories [43]:

- **Trivial** (Programs that an experienced programmer can write in one or two days).
- **Small** (Programs that can be coded by a skilled programmer in less than one year of full-time work).
- **Large** (Programs needing more than 2 to 5 man-years, which usually are coded by large programming teams).

Of course, these categories are not sharply delineated, depending on the skills and programming abilities of those who write the programs. Since the majority of languages use the same fundamental principles, a good programmer should first learn these principles, before he/she learns the specific details of a certain language. Actually, a good programmer should be competent in more than one language, in order to be able to apply different methods (algorithms) for the solution of computer problems. An algorithm consists of a set of suitable instructions necessary to complete a task. Thus, knowing more than one language enables one to implement a given algorithm in many distinct ways. Typically, writing databases requires large programming groups. A known exception to this is *Larry Ellison,* who wrote the original version of Oracle on his own in 6 months. Programs are successful if they have long life times, which means that are used either as they stand or suitably modified by the programmers employing them.

**Computer Programming Languages**

All languages translate the computer programs into machine code (string of zeros and ones). However, they can vary in several respects [44].

A general classification of computer languages was given by Brian Hayes [45] so:

- Imperative languages.
- Object-oriented languages.
- Functional languages.
- Declarative languages (including logical programming).

Imperative and object-oriented languages are mainly used in applications, while functional and declarative languages are popular in academic and research environments, where the focus is on solving problems well before these problems are considered in mainstream practice.

The bottom-up classification of languages is the following [43]:

- Direct programming.
- Assembly languages.
- High-level languages.

**Direct programming** was the original programming mode in a "language" that the computer understands directly (hard wiring or plug-board programs). In early computers, the programming was performed via a set of switches (front-panel

switches accompanied by associated indicator lights). An alternative method developed for use in early industrial-age factories was the *punch-card* method, which allowed easy switching to new designs. The front-panel switch and punch-card methods used *numeric codes* which encapsulated various machine instructions. *Machine code* involves the numbers used internally, whereas the numbers used on external media (punch cards or disk files) are called *object code*.

**Assembly languages** are based on writing a list of machine instructions using human-recognizable symbols. A suitable program, called an *assembler*, converts these symbolic instructions into machine or object code. Assembly languages are easier to understand and use than raw machine code, but they are still very near to machine language. The human-readable version of assembly code is called the *source code* because it is the source that the assembler converts into object code.

**High-level languages** are easier to understand than assembly languages and can be run on many different computers. The source code written in a high-level language is translated into object code via a *compiler*, an *interpreter* or *both* (compiled and interpreted versions). A *compiler* translates a program (or part of it) into object code, through a number of steps (sometimes first converting the program into assembly language and then to object code via an assembler or first converting the program into a platform-independent intermediate language). Compiled codes run faster than interpreted codes. Sometimes, an optimizer compiler is used to identify ways in which the compiled code can run even faster. An *interpreter* converts each high-level instruction into a series of machine instructions that are immediately executed.

Large programs, written by multimember teams of programmers, can be broken down into several parts, each once of which is separately compiled. The individual programmers write their own part of the program, which are then merged into a single object program by a linker. A program can be loaded into the main memory by a special program called a *loader*. A program that helps the programmer to create or edit the source files for programming is called an *editor*. An editor must have the essential tools needed in order to support syntax highlighting of the language concerned. Clearly, the editor is one of the most useful tools in the hands of the programmer.

**Widely Used Languages** Some widely used languages which played a significant role in the development of structured and object-oriented programming are: Algol, Basic, Simula, Smalltalk, Pascal, C++, Objective C, Lisp, Prolog, Concurrent Prolog, Java, JavaScript, HTML, HTTP, and PHP. A listing of books concerning five modern languages is provided in [43].

#### 5.3.1.6 Artificial Intelligence and Knowledge-Based Systems

The *artificial intelligence* (**AI**) field is concerned with intelligence machines, or rather with embedding intelligence in computers. According to *McCarthy*, who introduced the term in 1956, "*artificial intelligence is the science and engineering of making intelligent machines*" [46]. Today, AI has become an important element

of the computer industry, helping to solve the extremely difficult problems of society. *Knowledge engineering* (**KE**) and *expert systems* (**ES**) belong to AI but have been also developed separately especially on the applications side. The term "knowledge engineering" was coined in 1983 by *Edward Feigenbaum* and *Pamela McCorduck* [47], as "the engineering field that integrates knowledge computer systems in order to solve complex problems normally requiring a high level of human expertise." Closely related to KE are the expert systems (ES) each of which deals with a specific problem domain requiring high-level and professional expertise. Actually, expert systems are computer programs that simulate the reasoning and performance of an expert. Alternatively, one can say that an ES is a computer application that solves complex problems that would otherwise require extensive human expertise. To this end, it simulates the human-reasoning problem by using specific rules or objects representing human expertise.

Some of the problems that fall within the framework of AI are:

- Game playing.
- Theorem proving.
- General problem-solving.
- Natural-language understanding.
- Pattern recognition.
- Perception and cognition.
- Symbolic mathematics.
- Medical diagnostics.
- Fault diagnosis/restoration of technological systems.

The basic steps in the development of AI in the decades 1950–1980 are the following:

| Decade | Area | Researchers | Developed system |
|---|---|---|---|
| 1950 | Neural networks | Rosenblatt (Wiener, McCulloh) | Perceptron |
| 1960 | Heuristic search | Newell and Simon (Turing, Shannon) | GPS (General problem-solving) |
| 1970 | Knowledge representation | Shortliffe (Minsky, McCarthy) | MYCIN |
| 1980 | Machine learning | Lenat (Samuel, Holland) | EURISCO |

The 1980s was the age of expert systems and neural networks, and the 1990s the age of intelligent robots (e.g., Sojourner to Mars, the volcano-exploring robot Dante, etc.). *Newell* and *Simon* coined the physical-system hypothesis which says that "a physical system has the necessary and sufficient means of general intelligent action" and implies that the essence of intelligence is symbol manipulation. Contrary to this hypothesis, *Godel* has derived the so-called "*incompleteness theorem*", which states that: "A physical symbol system cannot prove all true statements" and limits what machines can do. *Hans Moravec* and others have argued

that the human brain can be copied (simulated) directly into hardware and software (a statement that supports the materialistic position that the mind is the result of a physical process in the brain). Today there exist powerful AI-based *natural-language processing systems* (e.g., advanced grammar and spelling checkers, speech systems, machine translators, chatbots, etc.). Artificial intelligence has entered into almost all industrial and real-life applications. New intelligent robots (mobile robots, humanoids, telerobots, surgical robots, etc.) are used for social and medical services, as well as for entertainment, also possessing human-like-emotion features (e.g., the MIT KISMET and the Hertfordshire University KASPAR social robots).

Due to the broad spectrum of human application areas and problems that are dealt with by AI, the approaches of solution required are numerous and quite different from each other. However, there are some basic methods that play major roles in all cases. These are the following [48–50]:

- Knowledge acquisition and maintenance.
- Knowledge representation.
- Solution search.
- Reasoning.
- Machine learning.

**Knowledge acquisition** is the first step in any AI-based knowledge-based solution process, and a format has to be used for the general representation of facts and the relationships between the facts that will cover the entire (or at least the greater part) of the domain knowledge.

**Knowledge representation** should capture the essential aspects of the knowledge domain of concern in a form suitable for later processing. In no way should the knowledge-representation methods be domain or content restricted. The knowledge hierarchy ranges from simple data up to meta-knowledge as shown in Fig. 5.2.

The most important and popular methods of knowledge representation are the following:

- Predicate logic (calculus)
- Propositional calculus
- Production rules
- Semantic networks
- Frames and scripts
- Objects
- Model-based representation
- Ontologies

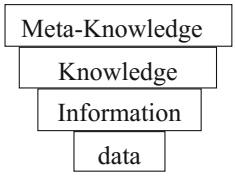



**Fig. 5.2** Hierarchy of knowledge levels

The last method that is based on the "*ontology*" concept is relatively newer than the other methods, and we will discuss it briefly here. The term "ontology" was borrowed from philosophy, where ontology is a branch of metaphysics that deals with the study of *being* or *existence* and their categories. The term "ontology" has its origin in Greece (όντα = onta) = beings and λόγος (logos = study).

*Aristotle* described ontology as "the science of being: qua/being" ("qua" = "with respect to the aspect of"). *Plato* considered that ontology is related to "ideas" and "forms." The three concepts that play a dominant role in metaphysics are: "*substance,*" "*form,*" and "*matter.*"

In knowledge engineering, the term "ontology" is used as a "representation" of knowledge in knowledge bases [51, 52]. Actually, ontology offers a shared vocabulary that can be used to model and represent the types of objects or concepts of a domain, i.e., it offers a formal explicit specification of a shared conceptualization [53]. In practice, most ontologies represent individuals, classes (concepts), attributes, and relations. A method that defines and uses ontologies, which are specific for problem-solving processes, is discussed in [54]. These ontologies are called "*method ontologies*" and were used, e.g., in Protégé-II.

The principal ways *for solution search* in the state space of AI problems are:

- Depth-first search.
- Breadth-first search.
- Best-first search.

All of them belong to the so-called *generate-and-test* approach in which the solutions are generated and subsequently tested in order to check their match with the situation at hand.

*Reasoning* with the stored knowledge is the process of drawing conclusions from the facts in the knowledge base, or, actually, inferring conclusion from premises. The three classical ways of knowledge representation directly understandable by the computer are:

- Logic expressions.
- Production rules.
- Slot-and-filler structures.

A form of reasoning, known as *automated reasoning*, is very successfully employed in expert systems. This is based on logic programming and implemented in the PROLOG language.

**Machine learning** is an AI process difficult to define uniquely, since it ranges from the addition of any single fact or a new piece of new knowledge to a complex control strategy, or a suitable rearrangement of system structure, and so on. A useful class of machine learning is *automated learning,* which is the process (capability) of an intelligent system to enhance its performance through learning, i.e., by using its previous experience. Five basic automated learning aspects are:

- Concept learning.
- Inductive learning (learning by examples).
- Learning by discovery.
- Connectionist learning.
- Learning by analogy.

A complete expert system involves the following four basic components (as shown in Fig. 5.3) [55, 56]:

- The knowledge base (knowledge network)
- The inference engine
- The knowledge-acquisition facility
- The justification and explanation facility
- The system-user interface

These components perform the AI operations outlined above. A knowledge-based system must not necessarily have all the above components. The initial expert systems were built using the higher level languages available at that time. The two higher level languages most commonly used in the past for AI programming are LISP and PROLOG. At the next level, above the higher level

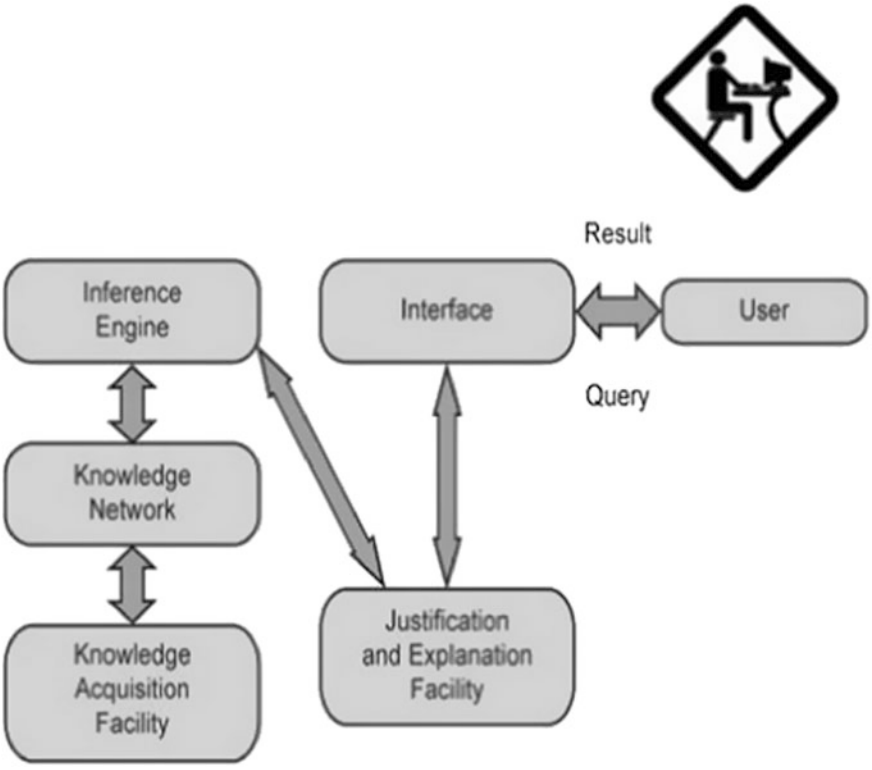



**Fig. 5.3** Typical structure of a complete expert system

programming, are *programming environments* designed to help the programmer/designer to accomplish his/her task. Other, more convenient programs for developing expert systems are the so-called "*expert-system building tools,*" or "*expert-systems shells,*" or just "*tools.*" The available tools are sortable into the following five types [57]:

- Inductive tools.
- Simple rule-based tools.
- Structured rule-based tools.
- Hybrid tools.
- Domain-specific tools.
- Object-oriented tools.

Among the tools available in the market for commercial or academic/research applications, we mention the following:

- EMYCIN (Empty MYCIN, i.e., MYCIN without the knowledge base).
- KS 300 (a slight modification of EMYCIN).
- KEE (object-oriented).
- Knowledge Craft (object-oriented).
- VP EXPERT (a good small expert system).
- M.1 and Nexpert (medium-size E.S.).
- S.1, OPS83, OPS5 (large-size E.S.).

### 5.3.2 Computer Engineering

*Computer engineering*, which considerably overlaps with computer science, is the area of information technology that merges and uses electronic engineering and computer science. It involves the following primary subfields:

- Logic design and computer hardware.
- Computer architectures.
- Parallel computing.
- Software engineering.
- Operating systems.
- Embedded systems.

As seen by the above list of subfields, computer engineering deals with practical aspects of computing from the logic, hardware and software design of microprocessors, personal computers, and mainframe computers to high-performance computers (supercomputers). In many university curricula, computer science and computer engineering are organized jointly under the name *computer science and engineering*.

#### 5.3.2.1 Logic Design and Computer Hardware

The basic logic analysis and design topics used for the design of the computer components are the following:

- Binary and hexadecimal number systems.
- Boolean-algebra and Boolean-equations' minimization.
- Combinatorial-logic design.
- Sequential-logic design.
- State sequencers and controllers.

The computer hardware design is performed at three levels, namely:

- Device level.
- Logic level.
- System level.

Classical books on digital computer design involve the following: "Logic and Computer Design Fundamentals" (*Mano/Kime*) [58], "Digital Design" (*Mano/Ciletti*) [59], and "Fundamentals of Digital Logic Microcomputer Design (*Rafiquzzaman*) [60].

#### 5.3.2.2 Computer Architectures

The theoretical foundations of computer architecture were established by *J. von Neumann* in 1945. A large class of electronic computers is based on the so-called *von Neumann architecture* which uses the concepts of stored-program, central processing unit (**CPU**) and "*single separate memory*" that holds both the program and the data.

Digital computers can be classified according to their architecture, the electronic technology, the type of processing, and the programming languages they use, into the following five generations [61, 62]:

- **First generation: Vacuum Tubes (1940–1956)**

First-generation computers used vacuum tubes for their logic circuits and magnetic drums for memory. They occupied entire rooms, and they were very expensive to operate, using a lot of electrical energy and generating a lot of heat, often leading to malfunctions (e.g., IBM 761)

- **Second generation: Transistors (1956–1963)**

The basic circuit element in these systems was the transistor (invented in 1947) which generates much less heat. Second-generation systems were still using punch cards for input and printout for output. They evolved from machine language to assembly (symbolic) languages and higher level languages (ALGOL, FORTRAN, COBOL, etc.).

- **Third generation: Integrated Circuits (1964–1971)**

Here, the basic element was the *small-scale integrated* (**SSI**) and the *medium-scale integrated* (**MSI**) circuits. Computers became, for the first time, accessible to large numbers of users because they were smaller and cheaper. Among their principal features were the multiprogramming operating systems and the time-sharing concept. Examples of third-generation systems are IBM 360/91, Illiac V, and CDC 7600.

- **Fourth generation: Microprocessors (1971-Present)**

This generation is based on *large-scale integrated* (**LSI**) circuits (thousands of integrated circuits built onto a single silicon chip), which are the building elements of microprocessors. The first computer for the home user (**PC**) was created by IBM in 1981, and Apple in 1984 introduced the Macintosh. In this generation, *"supercomputers"* were also designed and released.

- **Fifth generation: Artificial Intelligence (Present and future)**

These computers use *very large-scale integration* (**VLSI**) circuits, pipelining, artificial intelligence, and knowledge-based techniques. Actually, this generation of computers is still in development, and parallel processing, quantum computation, and nanotechnology are expected to contribute a great deal to new capabilities and prospects in the years to come. Natural language and self-organization features are among the main goals of fifth-generation computers.

Very broadly, computer architecture is considered the organizational design and the operational structure of a computer and involves the following three elements [63]:

- **Instruction set architecture (ISA)** The computing-system structure seen by a machine language or assembly language programmer (memory instruction repertory, processor registers, and data/address formats).
- **Micro-architecture** The computer organization that specifies how various parts are interconnected and how they cooperate in order to realize the ISA.
- **System design** This design deals with all the other constituents of the computer, e.g., buses, switches, memory controllers, multiprocessing elements, CPU-off-load elements, etc.

The *ISA architecture* is implemented at three levels:

- Logic-design level
- Circuit-design level
- Physical-implementation level.

The *von Neumann architecture* involves the following five components (units) (Fig. 5.4):

- Arithmetic and logic unit (**ALU**).
- Control unit/Central processing unit (**CPU**).
- Main Memory (**RAM**, **ROM**).

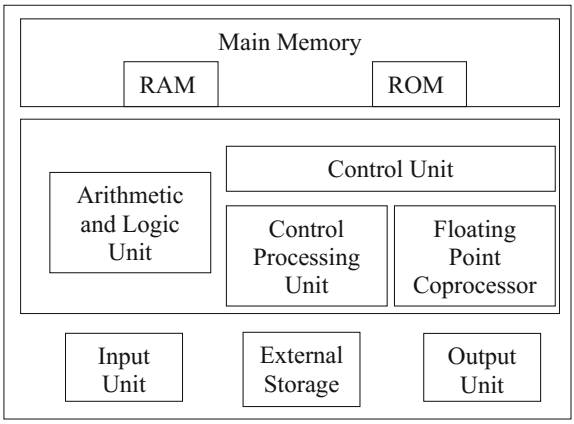



**Fig. 5.4** Units of the von Neumann computer architecture

- Input unit.
- Output unit.

The "*von Neumann architecture*" is also known as "*stored-program computer.*" Due to the separation of CPU and memory in this architecture, there is a limited data transfer (throughput) between the CPU and memory compared to the size of memory. This drawback is called the "*von Neumann bottleneck.*" Present-day computers have a throughput much smaller than the rate at which the CPU works. The CPU communicates with the other units via the so-called "*system bus.*" Some computers have more than one CPU. These are called *multiprocessing* computers.

The principal types of digital processors are [43, 63]:

- **CISC** (Complex-Instruction-Set Computer).
- **RISC** (Reduced-Instruction-Set Computer).
- **DSP** (Digital Signal Processor).
- **Hybrid** (Combination of features of CISC, RISC, and DSP).

*The arithmetic and logic unit* (**ALU**) performs integer arithmetic and logic operations, and also shift and rotate operations. Floating-point arithmetic is usually performed by dedicated *floating units* (**FPU**) which can be implemented as a co-processor (see Fig. 5.4).

The *control* of the computer is performed by control units which fetch and decode machine instructions and may control several external devices.

The computer has actually two data buses:

- An *internal bus* inside the processor (for the movement of data, instructions, addresses, etc., between registers and other internal components).
- An *external bus* outside the processor, but inside the computer (for the movement of data, addresses, and other information between major components inside the computer). Common types of buses include the system bus, data bus, address bus, cache bus, memory bus, and I/O bus.

The *main memory* (also called *internal memory* or simply *memory*) involves **RAM** (Random Access Memory) that enables the computer or processor to access any location in memory, while sequential access memories can be accessed in order. It also involves a **ROM** (Read Only Memory) which is also a random access memory, but only for reads. In its interior, the processor contains special types of memory such as *registers* and *flags*.

*External (or auxiliary) storage* is any storage other than main memory (e.g., floppy disks, zip disks, optical devices, etc.).

*Input devices* bring data into the computer and include punch-card readers, paper-tape readers, keyboards, mice, touch pads, trackballs, etc.

*Output devices* transfer information out of a computer. Pure output units include devices such as card punches, paper-tape punches, LED displays for light-emitting diodes, monitors, printers, and pen plotters.

*Input/Output (**I/O**) devices*, regardless whether they read-only or write-only or read and write, are classified into:

- *Character devices* that treat the data in streams of characters, bytes, or bits.
- *Serial devices* that treat streams of data as a series of bits.
- *Parallel devices* that treat streams of data in small groups of bits concurrently.
- *Block devices* that move large blocks of data all at once.

The behavior of a computer is application-dependent and can be categorized as:

- *CPU bound* (e.g., in scientific computing).
- *I/O bound* (e.g., web serving applications).
- *Memory bound* (e.g., video editing).

Other computer architectures, beyond the von Neumann architecture, involve the following [53, 63–69]:

- Harvard architecture.
- Stack architecture.
- Pipeline architecture.
- Non-uniform memory-access architecture.
- Vector or array-processor architecture.
- Quantum-computer architecture.

The most recent development in the computer field is the *quantum computer* which works by making direct use of quantum-mechanical phenomena, namely, *superposition* and *entanglement*. In contrast to traditional transistor-based computers, quantum computers use quantum properties to represent data and process these data. A quantum computer has the potential to improve enormously the processing speed of existing computers. However, until now, they only exist in the laboratory as small-scale prototypes with a capacity of a few bits. Due to the laws of quantum mechanics, if a quantum computer is expanded by just one single computer bit, then its computing power is immediately doubled, whereas, in a classical computer, the computing power grows only linearly with its components in the

ideal case. For example, 5% more transistors result in a 5% performance improvement. This means that the computing power of a quantum computer grows exponentially with its size, i.e., a quantum computer with n **qubits** (**qu**antum **bits**) can be in an arbitrary superposition of up to $2^n$ different states simultaneously (not in just one of these $2^n$ states at any one time, as in a classical computer). The implementation of *qubits* for a quantum computer can be made by using particles with two *spin states*: "*down*" (symbolically ↓ or | 0>) and "*up*" (symbolically ↑ or | 1>).

#### 5.3.2.3 Parallel Computing

Very broadly, *parallel computing* is the simultaneous use of multiple computing resources for solving a computational problem employing multiple CPUs. To this end, the problem is split into discrete parts so that each processing element can execute its part of the problem concurrently with the others. Each problem part is broken down into a series of instructions. The instructions of each part are executed concurrently on different CPUs [70–72]. In contrast, traditional serial computing is performed on a single computer that has a single CPU. The problem is split into a discrete series of instructions and the instructions are executed one after the other (serially). Only one instruction can be executed at any instant of time. Pictorially, the serial and parallel computation can be illustrated as shown in Fig. 5.5.

In parallel computing, the computing resources may be:

- A single computer with multiple CPU.
- An arbitrary number of computers forming a network.
- A combination of the above.

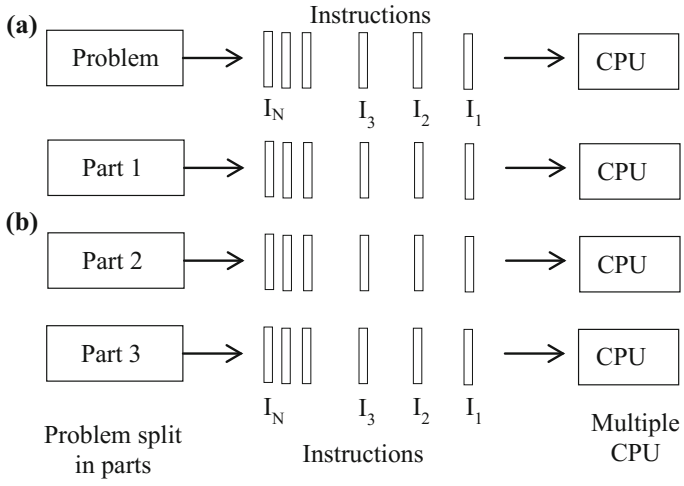



**Fig. 5.5** Serial (**a**) versus parallel (**b**) computation

The characteristics of parallel computing are the following:

- The problem is broken into discrete pieces of work that can be solved simultaneously.
- Multiple program instructions can be simultaneously executed at any moment in time.
- The problem can be solved in less time with multiple resources (than with a single computing resource).

In the ideal case, the speed-up due to parallelization would be linear, e.g., if we use three times more processors, we would get three times less runtime. In practice, however, the speed-up is nearly linear, because a problem may not be 100% parallelizable. The relation of speed-up and fraction P that is parallelizable is given by **Amdahl's law** coined in the 1960s by *Gene Amdahl* [73, 74]. This law is:

$$\text{Overall} \quad \text{speed-up} = \frac{1}{(1-P)+P/S}$$

where $P$ is the fraction which is parallelized and $S$ is the speed-up of the parallelized fraction.

For example, if we make 90% of a program run 10 times faster, we have $P = 0.9$ and $S = 10$. Therefore, the overall speed-up is equal to:

$$\text{Overall} \quad \text{speed-up} = \frac{1}{(1-0.9)+(0.9/10)} = \frac{1}{0.1+0.09} = 5.26.$$

Similarly, if we make 80% of a program run 20% faster, then $P = 0.8$ and $S = 1.2$, and

$$\text{Overall} \quad \text{speed-up} = \frac{1}{(1-0.8)+0.8/1.2} = \frac{1}{0.2+0.66} = 1.153.$$

Suppose that a special processor performs via floating-point operations 50% of our computations with a speed-up of 15. Then the resulting overall speed is:

$$\text{Overall} \quad \text{speed-up} = \frac{1}{(1-0.5)+0.5/15} = \frac{1}{0.5+0.033} = 1.876.$$

An augmentation of Amdahl's law with a corollary for multicore hardware, which is useful for future generations of chips with multiple processor cores, is described in [74].

Parallel computers are classified in many different ways. The most popular way is the *Flynn classification* which is based on the characteristics *Instruction* and

*Data*, each of which has only one of two possible states: *Single* or *Multiple*. Thus, Flynn's classification is represented by the following matrix:

| | |
|---|---|
| SISD | SIMD |
| MISD | MIMD |

where

SISD Single Instruction/Single Data
SIMD Single Instruction/Multiple Data
MISD Multiple Instruction/Single Data
MIMD Multiple Instruction/Multiple Data

Regarding parallelism, we have the following types:

*Bit-level parallelism:* Increasing the word size reduces the number of instructions a processor must execute to carry out an operation on variables with a size greater than the length of the word.

*Instruction-level parallelism*: The re-ordering and combination into groups of the computer-program instructions, which are then executed in parallel without changing the result of the program is called instruction-level parallelism.

*Data parallelism*: This is the parallelism that is inherent in program loops. The data are distributed across various computing nodes and are processed in parallel.

*Task parallelism*: This type of parallelism is a feature of parallel programs according to which "completely different calculations can be carried out on either the same or different sets of data." In data parallelism, only the same calculation can be performed on the same or different sets of data.

Regarding the parallel-computer memory architectures, we have the following types:

- *Shared-memory parallel computers* (Uniform Memory Access: UMA, Non-Uniform Memory Access: NUMA).
- *Distributed-memory parallel computers* (they require a communication network to connect the processing elements).
- *Hybrid shared-distributed-memory parallel computers* (The shared memory component is typically a cache coherent symmetric multiprocessing (SMP) machine. The distributed element component is the networking of multiple SMPs).

Parallel computing has been inspired and used over the years in complex and large real-time applications, including:

- *Physics* (nuclear, particle, photonics, high-energy).
- *Bioscience* (genetics, biotechnology)
- *Astronomy*
- *Geology and seismology*
- *Environment* (ecology, meteorology)
- *Chemistry and molecular science*

- *Electrical and mechanical engineering*
- *Mathematics and computers science*
- *Information systems*.

According to [70] "the most exciting development in parallel computer architecture is the convergence of traditionally disparate approaches on a common machine structure." In [70], a number of critical hardware and software issues for all parallel-architecture-communication latency, communication bandwidth, and coordination of cooperative work across modern designs are examined.

#### 5.3.2.4 Software Engineering

*Software engineering* is a term that was first introduced in the 1968 *Software Engineering Conference* of NATO and has created a scientific debate because this subject is more an art than an engineering field [75]. Despite this fact, software engineering has been developed considerably and been established as a separate subfield of computer science and engineering that employs systematically scientific, mathematical, management, and engineering methodologies to successfully build large computer programs (software). The general goal of software engineering is to develop new software (or modify existing software) so as to be more affordable economically, faster to build, maintainable, and, in sum, to achieve higher quality and better performance characteristics.

Software systems are created and built following an organizational and R&D structure called a *software project* [76, 77]. Building software systems is a difficult and complex process that requires high expertise, special skills, high-level capabilities, and interdisciplinary cooperation. Very broadly, software engineering involves the following:

- Strategies and goals.
- Software-requirements analysis.
- Software design.
- Advanced computer programming.
- Testing algorithms and procedures.
- Validation and verification.
- Software quality control.
- Software maintenance.
- Analysis and design tools.
- Project management (planning execution)
- General computer-engineering tools and methods.

The field of software engineering has now arrived at a very mature state and has a very broad scope, involving both system-analysis and system-design issues. A good picture of software engineering can be obtained by the following selected list of topics which are examined in the introductory tutorial provided in www.freetutes.com/systemanalysis:

- **Information System**. A general overview of information systems and their relevance to the operation of any organization.
- **Software system development life cycle models**. Different available models and approaches towards software development (Waterfall Model, Prototype Model, Dynamic Model, Object-Oriented Models).
- **Preliminary analysis**. It concerns the feasibility study for the system under development.
- **Fact-finding and decision-making**. It deals with the fact-finding techniques, decision-making, and documentation techniques.
- **Functional modeling**. System-design issues and concepts (input/output, databases, data flow diagrams), and modular programming approach to software development. (i.e.,: structure of charts, cohesion, and coupling concepts).
- **Data modeling**. It is concerned with the data modeling phase of the system development (entity relationship model, relational model, and object oriental models).
- **Testing and quality assurance**. It includes the various testing, validation, and quality-assurance techniques during the entire cycle of system design and implementation.

Of course, in practice, the managerial and personnel aspects should be properly taken care of over the entire software-system design and development process.

*Software engineering* is included in academic curricula which are evaluated according to the established accreditation practice. Most software engineers are working as company employees or contractors. Usually, professional certification is required by the employing companies.

#### 5.3.2.5 Operating Systems

An *operating system* (**OS**) is a software program which connects the computer hardware with the computer software. The purpose of an OS is actually to organize and control the hardware and software such that they cooperate efficiently in a flexible but predictable way. The operating system is the first program that must be loaded onto a computer. Without an OS, a computer is a useless machine.

The evolution of operating systems has closely followed the evolution of computers. As a result, there is available a very wide repertory of OS types. Many of the marketed or freely available operating systems have characteristics which belong to more than one type. The main types of operating system are the following [78]:

- **Graphical user-interface (GUI) operating systems**. A GUI OS involves icons and graphics and is usually navigated via a computer mouse. Examples of GUI operating systems are: Windows 98, System 7x, and Windows CE.
- **Multi-user operating systems (MU/OS)**. These OSs allow the use of a computer by multiple users at the same and/or different times. Examples of MU/OS are: Unix, Linux, and Windows 2000.

- **Multitasking operating systems (MT/OS)**. Any operating system that can allow multiple software processes (tasks) to run at the same time (concurrently) is called a multitasking operating system. Examples of MT/OS are: Unix, Linux, and Windows 2000.
- **Multiprocessing operating systems (MP/OS)**. MP/OS is an operating system that can support and employ more than one computer processor. Examples of MP/OS are: Unix and Windows 2000.
- **Multithreading operating systems (MTH/OS)**. These are operating systems that enable various parts of a software program to run simultaneously. Examples of this OS type are: Unix, Linux, and Windows 2000.

A list of several distinct operating systems currently available, with their dates of release, is provided in [78, 79].

#### 5.3.2.6 Embedded Systems

An *embedded system* is a computer system with an embedded or built-in processor, designed for dedicated tasks for some specific kind of real-time applications. Embedded systems include not only one or more processing cores, but also some or all of the mechanical devices of the application concerned. Thus, very broadly speaking, embedded systems are computers inside a product (machine, turbine, lift, car, train, air conditioner, etc.). Most of the applications for which appropriate embedded systems must be designed are *time-critical*, and, therefore, the embedded system should be a *real-time* system. An embedded system is specified by the application rather than the hardware itself, in contrast to a general-purpose computer (e.g., PC) which is designed so that it is adaptable to a large repertory of applications and end-user requirements [80–83].

The two basic components of a typical embedded system are a *computer* and an *interface* to the physical world device (a keyboard, a car, a chemical plant, and so on).

The *computer* may be *small*, *medium,* or *large* according to the application concerned. For a TV remote control, a small four-bit microcontroller is sufficient. Data acquisition systems in laboratories or factories can be handled by medium-sized (eight or 16-bit) microcontrollers or PCs. Plant monitoring and control systems need large high-end computers.

The *interfaces* in embedded systems belong to the following types:

- Common serial and parallel interfaces.
- Industrial interfaces.
- Networking interfaces.
- Analog-to-digital converters (ADC) and Digital-to-analog converters (DAC).

The most popular type of interface is the first type involving standard serial buses (RS-232, RS-423, RS-422, and RS-485), while in industrial systems use is

made of interfaces like the IEEE 488, Computer Automatic Measurement and Control: CAMAC, Controller Area Network: CAN, and IEEE 1394.

Real-time embedded systems can be categorized into two kinds:

- *Hard real-time embedded systems*, which have to meet deadlines always under all conditions (e.g., in temperature control of a critical process).
- *Soft real-time embedded systems*, where occasional failures to meet a deadline are acceptable.

Embedded systems can be used not only as primary or secondary controllers, but also as protocol converters and data processors in heavy-duty applications.

Early examples of embedded systems appeared in the 1950s and include a closed-loop control system developed at a Texaco refinery in Texas and an analogous system at Monsanto Chemical Company, namely, an ammonia system in Louisiana. Two other well-known embedded systems are the *Apollo Guidance Computer* developed at MIT (instrumentation Laboratory) and the *Autonetics D-17* guidance computer for the Minuteman missile (1961). Since then, embedded systems feature continuously increasing computing power at radically reduced costs.

Advances in communications have also considerably contributed to the growth of embedded systems. Data communication devices and cellular phones are two fundamental classes. Today, web-based embedded systems are very-low-cost systems that can be connected in distributed environments. The modern web-based embedded module functions can be connected to the Internet in two alternative ways:

- Ethernet connection.
- Serial connection.

A set of 43 papers and presentations on embedded systems is available On the Carnegie Mellon University site (P. Koopman) [81].

Today, embedded systems are used in a wide variety of real-life applications. Some primary examples are:

- *Telecommunications* (e.g., pagers, telephones, mobile phones, dedicated routers, network bridges, etc.).
- *Process control* (This is the predominate type of systems in which embedded systems have found applications from the very beginning of their history).
- *Manufacturing* (Here, specialized equipment is included, such as fault-diagnosis equipment, production-line equipment, robots, etc.).
- *Transportations* (e.g., automobiles, trains, avionics equipment, and, in general, motor control).
- *Consumer electronics* (e.g., household equipment, personal electronics, audio-visual devices, etc.).
- *Office automation* (e.g., computers, printers, scanners, photocopying machines, fax devices, surveillance equipment, etc.).
- *Internet* (e.g., PCs and accompanying accessory equipment).

### 5.3.3 Telecommunications

The history of telecommunications and a tour of communication and information theory was provided in Chap. 4. Here we will present a short conceptual account of the following additional topics:

- Telematics.
- Cellular communications.
- Computer networks.
- Internet and the World Wide Web.
- Web-based multimedia.

#### 5.3.3.1 Telematics

As mentioned in Sect. 4.3.6, "*telematics*" is concerned with the long-distance transmission and computerized processing of information, and so it is actually a field belonging to the general field of information technology. Practically, telematics is related to the modern industry that combines computers and telecommunication systems, including dial-up service to the Internet and all kinds of networks used for transmitting data. Moreover, specific applications of telematics include automobiles with **GPS** (*Global Position System*) tracking that are integrated with computers and wireless communications for use in emergency working situations, etc. As the size of computers becomes ever smaller and their energy needs are continuously reduced, computing equipment becomes mobile, and the pocket computer becomes a basic element of our everyday life wherever we are. These advancements produce new areas of telematics applications, frequently called *ubiquitous computing*, *mobile computing*, or *pervasive computing* [84–86].

A short list of such applications is the following:

**Intelligent transportation systems** These include many embedded components, such as engine control, instrumentation, in-vehicle comfort, emergency warning systems, etc. The use of telematics enables instantaneous travel-direction recognition of a vehicle which can be sent, in real-time, to surrounding vehicles equipped with proper collision avoidance and warning systems (CAWS) in order to actually avoid collisions and crashes.

**Fleet control** This refers to the management and control of the vehicle fleet of a given company, which includes vehicle maintenance, vehicle tracking, vehicle diagnostics, driver management, fuel management, and safety management.

**Mobile television** This kind of TV works in a way similar to mobile phones and displays TV channels and programs via LCD devices.

**Monitoring water and air pollution** This is based on distributed metering and computing devices.

**Positioning systems** These systems are important for autonomous navigation, tracking of commercial vehicles, and total traffic monitoring. The position of a vehicle can be determined by several methods, such as dead reckoning, satellite positioning, signpost systems, geographic data bases, etc.

**Telemedicine and telehealth** Telematics in health-care needs, besides the software and hardware technology, requires acceptance by the community. This can be enhanced through proper education of both the public and the physicians [87].

#### 5.3.3.2 Cellular Communications

The two broad categories of wireless telecommunications are [88–90]:

- Fixed wireless communications
- Mobile communications

*Fixed wireless* is just an alternative to wired communications, where the user does not need mobility to obtain low-cost telecommunications from fixed locations.

*Mobile communications* must have the capability to provide service to mobile-phone users outside their home environment. After the appearance of *cellular networks*, the so-called *Personal Communications Service* (**PCS**) technology was developed to satisfy the needs of personalized communications "anytime and anywhere." PCS networks were developed employing radio-frequency designs similar to cellular ones. However, although "*cellular networks*" and *PCS* use the same fundamental technologies, they have considerable and notable differences.

*Cellular systems* employ large numbers of wireless transmitters (of low power) to create communication cells spanning certain geographic service areas. The cells are sized according to the users' density and demand within the region concerned. As the users, in their travels, pass from one cell to another, their communication is "*handed off*" between cells providing a seamless service. Each mobile uses a separate, temporary radio channel to talk to the *cell site* or *cell tower* that can talk to multiple mobiles simultaneously, using one channel per mobile. Two frequencies are used: one for the *forward link* and another for the *reverse link*. The mobiles must move near the base station to maintain communication. This is because the energy dissipates over distance. The improved capacity of a cellular network compared with a single transmitter, is due to the possibility of reusing the same radio frequency in a different area for a different transmission, whereas, in the case of a single transmitter, only a single transmission can be used per radio-frequency channel. Not all channels can be reused in every cell due to the interference problems caused by mobile units. The cellular-radio base station can communicate with mobile units only if they are within range.

The signals coming from several different transmitters are distinguished using *frequency division multiple access* (**FDMA**) and *code division multiple access* (**CDMA**). Cellular (and PCS) networks can use several mobile networking protocols that can work and enable high-level services while the user is moving away from the home area. Such protocols include the European **GSM** *(Global System for*

*Mobility)* protocol, the **GSM** *Mobile Application Part* (**MAP**) standard protocol (used worldwide), and the **ANSI-41** protocol standardized by the Telecommunication Industry Association (**TIA**) and the American Standards Institute (**ANSI**).

A *communication cell* is the basic geographic unit of a cellular system. Rural and urban areas are split into regions on the basis of suitable engineering rules. The name cell comes from the *honey comb (hexagon) shape* of the regions into which the area concerned is divided (Fig. 5.6). Of course, not all cells are *perfect hexagons* because of variations of natural terrain and possible man-made constraints.

*Cluster* is defined to be a group of cells (Fig. 5.7) [88]. Within a cluster, it is not possible to reuse any channels.

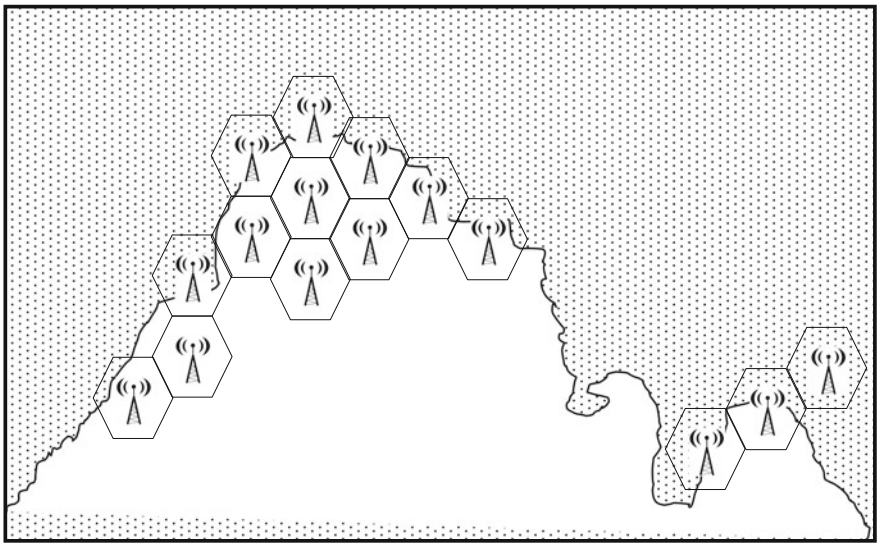


**Fig. 5.6** Typical structure of a cellular mobile communication system [88]

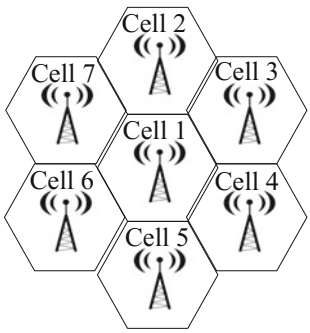



**Fig. 5.7** Cluster with seven cells [88]

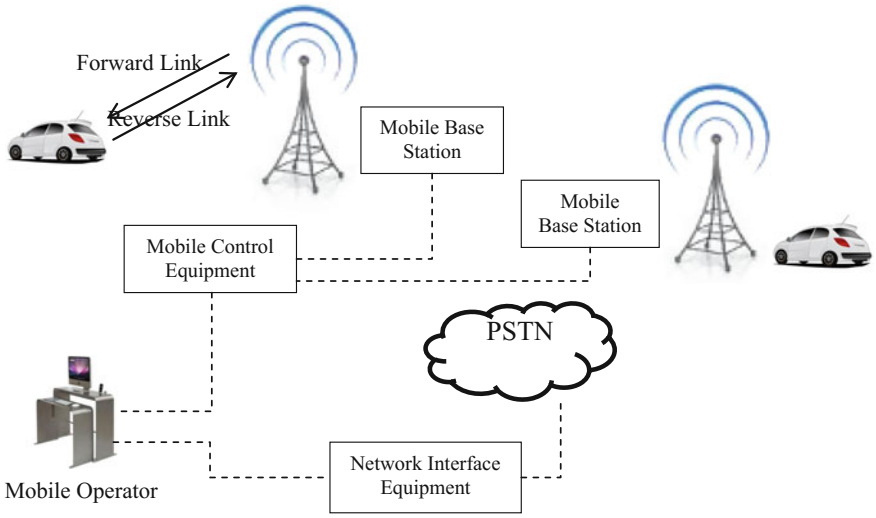



**Fig. 5.8** The four subsystems of a cellular mobile telephone service network

The concept of *frequency reuse* was coined because only a small number of frequencies are available for mobile systems.

The four principal subsystems of a cellular communication system are (Fig. 5.8):

- *PSTN*: Public switched telephone network.
- *MTSO*: Mobile telephone switching office.
- *Cell center* (or site) with the antenna.
- *MSU*: Mobile subscriber unit.

The PSTN consists of local networks, exchange area networks, and the long-range network for communication connections worldwide. The MTSO is the main office for mobile switching via the mobile switching (MSC) and the relay stations for connection with the PSTNs (network interface). The cell center is the physical place of the antenna and the accompanying radio equipment. Finally, the MSUs involve the mobile control equipment, a transceiver for transmission, and receipt of radio transmission to and from a cell tower.

#### 5.3.3.3 Computer Networks

A *computer network* (or simply *network*) is a group of two or more computers with their accessories connected by communication channels, and primarily intended to share resources among the network users. As mentioned in Sect. 4.3.6, the starting point of the computer-network field and the Internet is the *ARPANET* (Advanced Research Projects Agency Network) where the first link was realized between the University of California and Stanford (21 November 1969). Besides the resource

sharing ability, computer networks enable communications among people in various ways (e-mail, chat rooms, video telephones, video conferencing, etc.), and the sharing of software, files, and data.

Computer networks [91–93] are realized using two different technologies, namely:

- *Wired technologies* (optical fiber, coaxial cable, etc.).
- *Wireless technologies* (Earth microwaves, communication satellites, wireless LANs, cellular and PCS technology, wireless web).

A computer network consists of two kinds of components, namely: *nodes* and *communication lines*. The nodes treat the network protocols and provide switching capabilities. "A node is typically itself a computer called a *host*" which runs specific communication software. The communication lines may be of several forms, e.g., copper wire cables, optical fiber, radio channels, and telephone lines.

An abstract network with several nodes and hosts has the form shown in Fig. 5.9a.

Figure 5.9b depicts how different computer networks can be interconnected, and Fig. 5.9c shows an ad hoc computer-network diagram.

Every host is connected to one or more nodes of the network via separate communication lines. Each host has its own address allocated to it by the network. All communication among hosts is routed via the nodes, which have the task to find how to direct messages across the network from one point to another. A host can communicate with another host if the former knows the address of the latter.

Actually, there are several network types worldwide, of which LAN and WAN are the ones most widely used. These network types are:

**LAN-Local Area Network**

This is a network bounded in a small area, e.g., a home, an office, a hospital, an enterprise building or a small group of neighboring premises. A LAN uses the TCP/IP network protocol for the communication of the computers. LANs are usually controlled and administrated by a single person or organization and implemented as single IP subnets.

**WLAN or Wi-Fi–Wireless LAN**

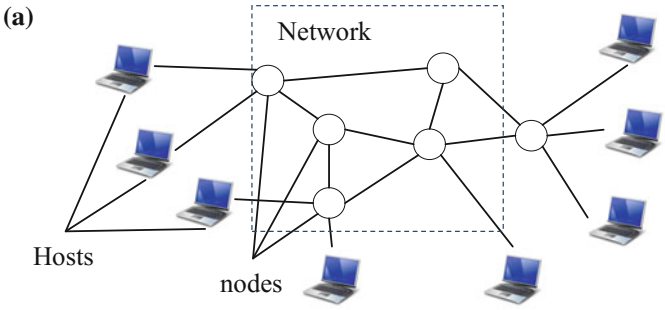
In WLAN, usually called Wi-Fi, no wires are used, but the communication medium consists of radio signals. Wi-Fi's need the installation of network cards in order to access any nearby wireless network, through wireless routers.

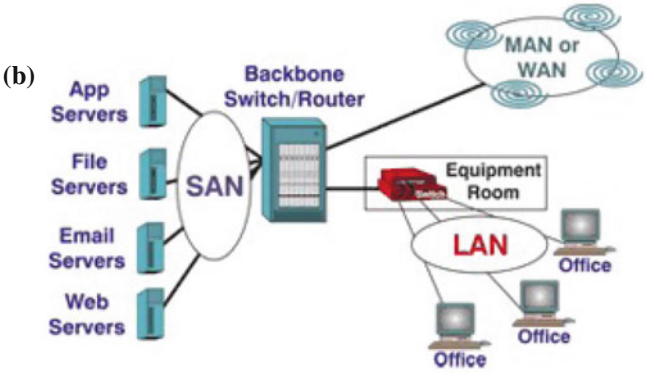
**MAN-Metropolitan Area Network**

This is a network for a medium-sized area, larger than that of a LAN, such as a city. This type of network is not very frequently used, but is important for governmental bodies and large-scale organizations.

**CAN-Campus Area Network**

This is a network of several LANs which is used in areas smaller than metropolitan ones, e.g., a large university campus or a cluster of local offices and buildings.


(a)
Network
Hosts
nodes


(b)
App Servers
File Servers
Email Servers
Web Servers
SAN
Backbone Switch/Router
MAN or WAN
Equipment Room
LAN
Office
Office
Office

(c)


Internet
Broadband modem
Ethernet or USB cable
Computer 2
Wi-Fi
Printer
Computer 1
Wi-Fi
Computer 4
PDA
iPhone
Computer 3

◂**Fig. 5.9** **a** A general network with several hosts and nodes, **b** Interconnection of various computer networks (LAN, SAN, MAN, and WAN), **c** Diagram of an ad hoc wireless computer network, [www.fiber.info/articles/fiber_optic_network_topologies_fiber_for_its_and_other_systems, www.conceptdraw.com/examples (The reader is informed that Web figures and references were collected at the time of the writing the book. Since then, some of them may no longer be valid due to change or removal by their creators, and so they may no longer be useful.)]

**WAN-Wide Area Network**

A WAN covers a large geographical area and consists of a collection of LANs distributed over this area. LANs and WANs are connected by routers. The WANs use protocols such as ATM, X.25 etc. Similar to LAN, a WAN can be wireless (WWAN), the most common of which is the IEEE 802.16.

**SAN-Storage Area Network/System Area Network**

This network connects data-storage equipment to servers through fiber channels. They are typically used in data-oriented companies. The same acronym SAN is also used for system-area networks (sometimes called cluster-area networks), which connect supercomputers via high-speed connections in cluster configuration.

**GAN-Global Area Network**

A GAN is a network that can support mobile communications over a large number of LANs, MANs, satellite coverage areas, and so on.

Two other types of small computer networks are: (i) **PAN**: *Personal Area Network* or **HAN**: *Home Area Network*, which is used for the connection of personal digital equipment at home, such as RCs, printers, FAX machines, and mobile devices, and (ii) **DAN**: *Desk Area Networks*. There is also the class of the so-called *virtual private network* (**VPN**) in which several links between nodes are implemented by open connections or virtual circuits within a larger network such as the Internet, instead of using wires. VPNs find application, among others, for safe communications via the Internet or for separating the traffic of different communities over a network with high-security performance.

The hardware components used for interconnecting the network nodes are standard network-interface cards, bridges, hubs, switches, and routers. A hub is a physical layer device, which connects multiple user stations, each with a dedicated cable.

The network topologies specify the way in which the network components are organized. The five widely used *LAN topologies* are the following: **Bus**, **Ring**, **Star**, **Tree**, and **Mesh** topologies. The simplest LAN connection is the Peer-to-Peer topology shown in Fig. 5.10a, which is the basic element of the bus topology.

**Bus topology**—This is implemented as shown in Fig. 5.10b, i.e., in a linear LAN structure where the transmissions from the network stations propagate the length of the medium and are received by all other stations.

**Ring topology** This is a LAN architecture in which the devices are connected in series by unidirectional transmission links to create a closed-loop (Fig. 5.11).

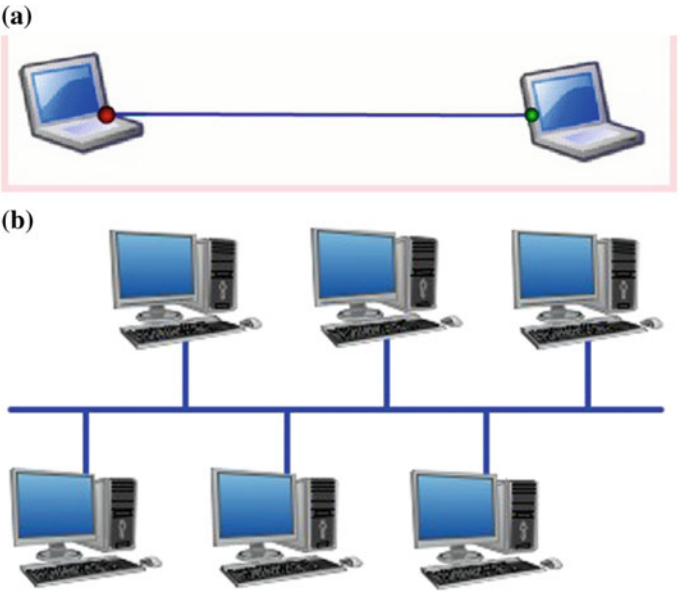



**Fig. 5.10** **a** LAN peer-to-peer topology, **b** LAN bus topology [(**a**): https://kelomptigaa.wordpress.com/tag/computer/; (**b**) www.conceptdraw.com/diagram/images-for-common-networking-tools]

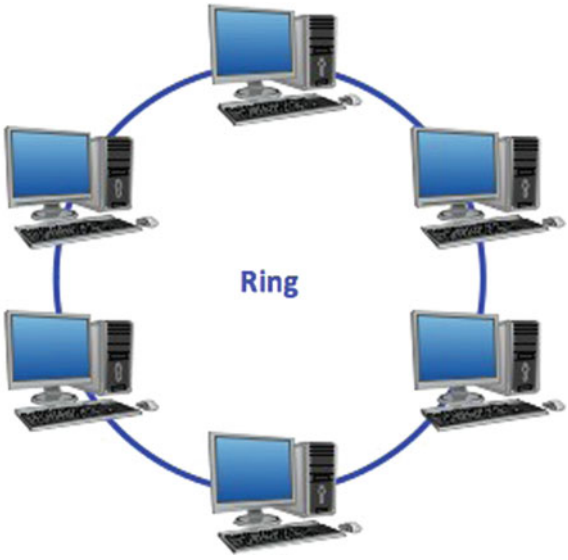



**Fig. 5.11** LAN ring topology (www.conceptdraw.com/diagram/images-for-common-networking-tools)

**Star topology** Here a single hub or switch acts as the central connection point for several other points via dedicated links (Fig. 5.12). Bus and ring topologies are sometimes implemented physically in a star topology, in which case we have a star-bus or star-ring topology.

**Tree topology**—This topology, which is also called *hierarchical topology*, is identical to the bus topology, but here branches with multiple nodes are possible as shown in Fig. 5.13.

**Fig. 5.12** LAN star topology (www.conceptdraw.com/diagram/images-for-common-networking-tools)

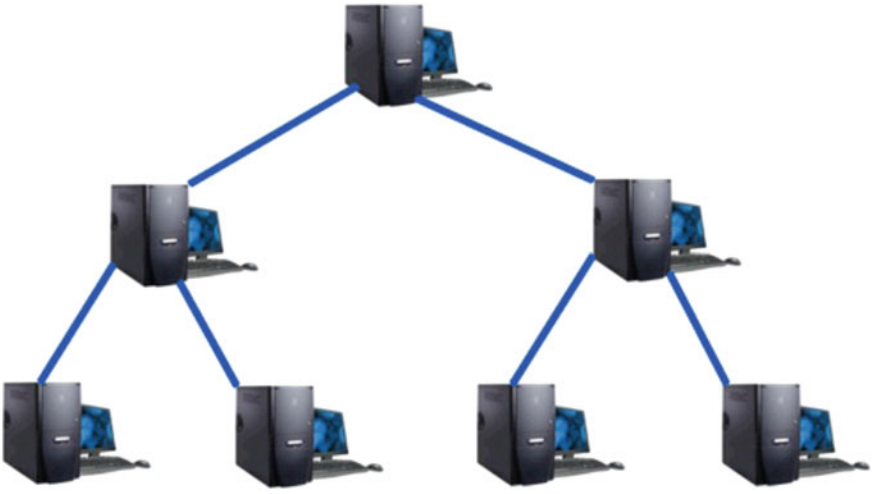


**Fig. 5.13** LAN tree or hierarchical topology (www.ulastboy31.com/computer-networking-lessons.html)

**Mesh topology**—A mesh-network topology is similar to an enterprise-wide mesh, incorporating interconnected nodes. Because all nodes are interconnected, data can go directly from the original node to its destination. If some connection has a problem, then routers can redirected the data easily and efficiently. Mesh LANs are the most reliable and fault-tolerant type of LAN because data can follow multiple routes between any two nodes. The disadvantage of full-mesh architecture (Fig. 5.14) is the cost because every node of a network must be connected to every other node, and so the cost increases as the size of the network increases. More practical and economical are the partial-mesh LANs, in which some links are omitted; so they are preferred over full-mesh networks.

LAN topologies are also used to WANs. In the bus topology (also *called "peer-to-peer" topology*), each site depends on every other site in the network to transmit and receive its traffic. Although in LAN peer-to-peer topology the computers share access to a cable, in WAN "peer-to-peer" topology, the various locations are usually connected to another one via dedicated circuits. Just for illustration, Fig. 5.15 shows the WAN peer-to-peer topology where each site (building) has

**Fig. 5.14** LAN full-mesh topology (www.contrib.andrew.cmu.edu/~ma1/computer_-networks.html)

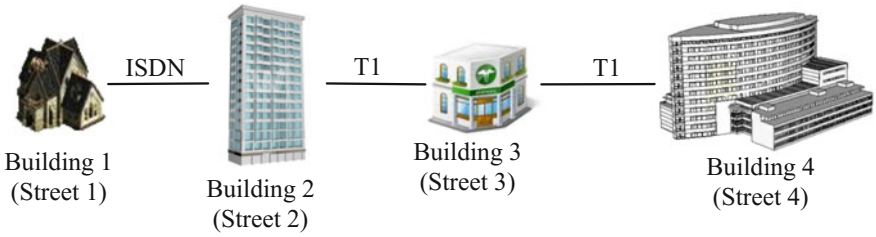



**Fig. 5.15** WAN peer-to-peer topology

continuously available communication channels between the access points leased from a telecommunications provider (e.g., an ISP).

Two other common building hardware devices of LANs are *repeaters* and *extenders*. A *repeater* is a device of the physical layer which is employed to interconnect the segments of an extended network. An *extender* is a remote-access multilayer switch which connects to a host router. Multiple LAN extenders connected to a host router via a WAN have the typical organization shown in Fig. 5.16.

#### 5.3.3.4 Internet and World Wide Web

The *Internet* is the largest of the so-called *inter-networks* (briefly *internets*) which are computer networks that connect several networks. The connection of several networks is achieved using devices which belong to the Layer 3 (Network Layer) of the OSI reference model. These inter-networks involve public, private, industrial, commercial, or governmental networks. The *Internet* is the largest of these inter-networks which is a publicly available internationally interconnected system of computers which employs the *TCP/IP* suite of packet switching protocols on top of any services or information that these computers offer. The *World Wide Web* (**www** or simply the **Web**) is one of the Internet applications (actually the most-widely used) organized with hypertext [94–97].

The web displays information in the form of *pages*, which are written in **HTML** (*HyperText Markup Language*). HTML specifies the way the information should be displayed no matter what is the type of computer or the browser used. Pages include hypertext links that enable the users to jump to other related information on the Internet. A *website* is called a collection of related Web pages having a common Web address (i.e., **URL**: *Uniform Resource Locator*, in Internet terminology). It is

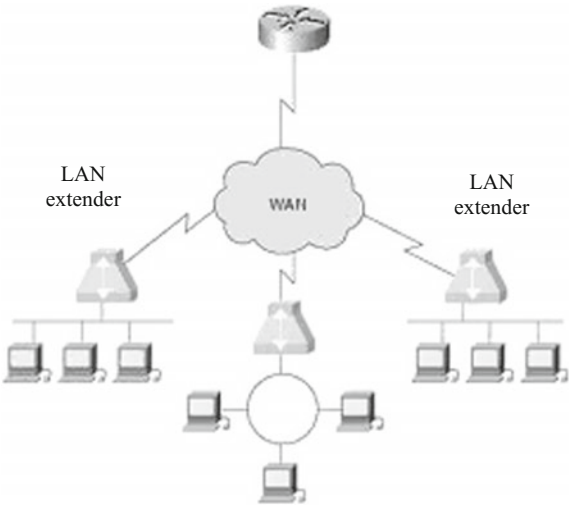



**Fig. 5.16** Organization of LAN extenders connected to a host router via a WAN

clear that the Web is just one way of accessing information over the Internet medium. It is an information-sharing facility that is built on top of the Internet. The Web and Web services employ the **HTTP** (*Hyper Text Transfer Protocol*), which is the standard application-level protocol used for transporting files on the Web. The Web employs *browsers* to access documents (i.e., web pages) which are linked to each other through hyperlinks. In sum, the Web is simply a part (a large part) of the Internet. Although many persons speak about the Internet and the Web interchangeably, the two concepts actually do not represent the same thing and are not synonymous. Therefore, they should not be confused. Popular Web browsers are: *Internet Explorer, Netscape Navigator, Mozilla Firefox,* and *Mosaic*.

A typical view of a browser with a short description of the commonly used features is shown in Fig. 5.17.

Other extensions of a computer network, typically a LAN, are the *intranets* and *extranets*. An intranet is a group of networks which are controlled by a single administrator via *Internet Protocol* (**IP**) and *IP-based tools* (e.g., Web browsers or file transfer applications). An extranet is a controlled private network that uses Internet technologies and the public telecommunication system to securely share business information of a company between partners, suppliers, distributors, etc. To handle different tasks over the Web, there are three principal services available: *Directories*, *Search Engines*, and *Meta Engines*. Directories are Web sites providing the major sites and companies. An example of a directory is *Yahoo* (www.yahoo.com). Search engines create an index based on the occurrence of key words in the full texts of all sites existing on the Web. Finally, metasearch engines enable users to query both directories and search engines. Useful search engines are *AltaVista* and *Infoseek* [98, 99], and a powerful metasearch engine is *Metacrawler* (www.metacrawler.com).

The functionality of *Netscape* and *Explorer* can be extended by downloading extra programs that can work within the browser [100, 101].

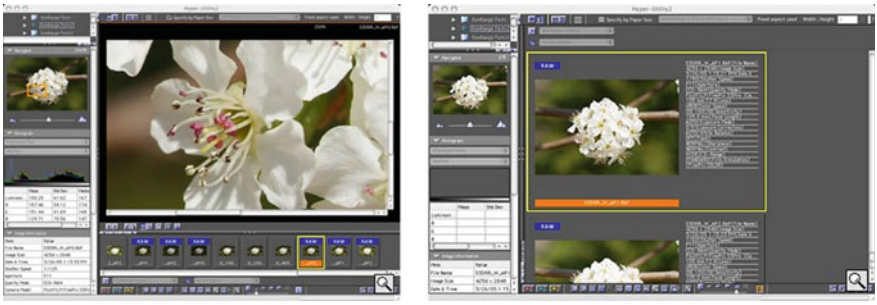


**Fig. 5.17** Typical features of a Web browser

Internet usage and population statistics around the world are recorded by *Internet World Stats* [102] and compiled by the *e-consultancy company* [103], which is a leading online publisher of best-practice Internet marketing reports, research, and guides.

The Web is a very useful distribution channel of knowledge or skill transfer [104]. This is called *Web-based training* (**WBT**) and reduces the principal barriers for distributed learning. For example, in the field of control and robotic engineering education, there are today important results and interactive Web-based platforms that provide e-course material and web-based laboratory exercises. These web-based platforms fall into the following categories [105]:

- E-course material and e-classroom environments
- Virtual laboratories
- Remote (physical) laboratories
- Combination of the above

Two principal protocols running on top of *IP networks* are the **TCP** (*Transmission Control Protocol*) and **UDP** (*User Datagram Protocol*). Whereas the IP protocol deals only with packets, TCP enables two hosts to establish a connection and exchange streams of data (i.e., sequences of bytes). At the program level, the TCP stream appears like a file. To receive data, a program has to wait until it receives the On Data Available event. UDP is used principally for broadcasting messages over a network. UDP preserves datagram boundaries between the sender and the receiver and is a connectionless protocol, i.e., a datagram can be sent any time without advance advertising, negotiation, or preparation. The sender sends a datagram and hopes that the receiver will be able to handle it. In sum, UDP provides no guarantee that the datagram will be delivered to the host, although the rate of failure is actually very low on the Internet, and almost zero on a LAN, unless the bandwidth has been exhausted.

#### 5.3.3.5 Web-Based Multimedia

Web-Based Information Systems are called **WIS** (*Web Information Systems*). A *Multimedia Web Information System* (**MWIS**) is a WIS that employs multimedia as its basic component [106, 107]. Currently, there exist several multimedia technologies and browser add-ons which enhance the Web with MWIS. Although correct delivery of documents and images must be secured (and for this reason are delivered by HTTP on top of TCP), this is not a requirement for time-sensitive information such as audio and video, which can tolerate frame losses. This has inspired the development of the *Real-time Transport Protocol* (**RTP**), together with the *Real-Time Streaming Protocol* (**RTSP**), in order to stream media of computer networks (and the Internet). MWIS applications typically run RTP on the top of UDP in order to employ its multiplexing and check-sum services. Of course, RTP can work with other transport protocols, as well. RTP is enhanced with a control protocol, making the **RTCP** protocol, which allows monitoring the data so that the

receiver is able to detect any packet loss and to compensate for any delay jitter. Using special *plug-ins* enables Web browsers to play audio and video streams, a feature that is not inherent in the browsers (see, e.g., www.netscape.com/plugins and www.activex.com). These special plug-ins, which make use of some RTP protocol, enable the Web browsers to establish their connections to servers and present multimedia within the flow of the hypertext document or in a separate window, by encoding the audio and video streams via low-bit rate compression techniques. Embodying such special plug-ins to browsers is very useful in 3D applications. Real-time communication of 3D data for stand-alone or networked applications can be realized via *Extensible 3D* (X3D), which is the successor to **VRML** (*Virtual Reality Modeling Language*) which is an XML-based 3D file format. X3D and VRML can be embedded in Web applications, thus leading to new enhanced human–computer interaction modes. The treatment of interactive and animated 2D data is now possible using **SVG** (*The Scalable Vector Graphics*) specification [1506, 107]. The major components of multimedia are the following [108]:

- *Capture equipment* (keyboards, mice, video camera, video recorder, audio microphone, 3D input devices, and haptic sensors).
- *Storage equipment* (hard disks, CD-ROMs, Zip/Jaz drives, DVDs, etc.).
- *Communication networks* (Ethernet, ATM, Intranets, the Internet, etc.).
- *Computer systems* (Workstations, multimedia desktop machines, and MPEG/Video/DSP hardware).
- *Display equipment* (high-resolution monitors, color printers, CD-quality speakers, HDTV, SVGA, etc.).

The multimedia applications include, but they are not limited to, the following:

- *Education* (MWIS-based training, courses, and e-books).
- *Engineering* (computer-based simulation, animation, virtual reality).
- *Industry* (presentations, reports, seminars, marketing, etc.).
- *Medicine* (virtual surgery, training, simulation of bacteria growth dynamics, robotic surgery, etc.).
- *Scientific research* (system modeling and simulation of physical, natural biological, and technological systems).

Multimedia Web information systems (MWIS) provide an important class of systems that can be used profitably in all the above (and many other) applications of multimedia. A MWIS that includes a multimedia PC, a set of other cooperating computers, and a server is shown in Fig. 5.18.

Two comprehensive books on research and theory of multimedia learning and multimedia on the Web are [109, 110]. These books provide complete studies of multimedia systems characteristics (computer control, integration, interactive, etc.) and the challenges of MWIS.

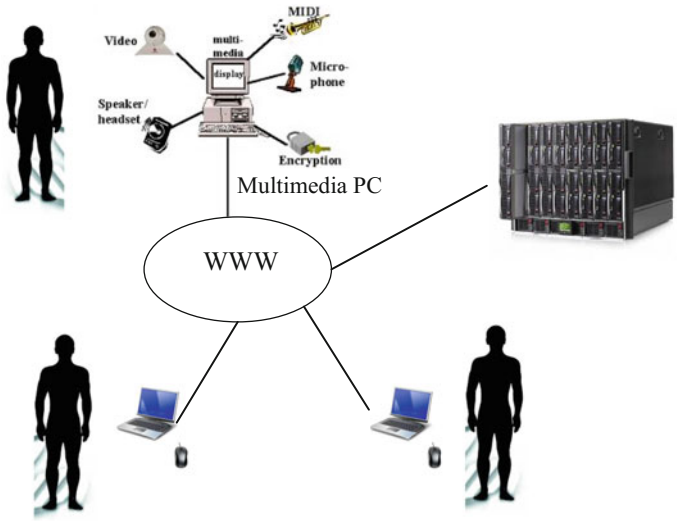



**Fig. 5.18** An example of multimedia Web information system

## 5.4 Information Systems

### *5.4.1 General Issues*

The *information systems* field has a history of about five decades and from its beginning has contributed very much to expand business and industry into global markets. Actually, the term "*information system*" involves the interaction between processes and technology both within the boundaries of an organization (*intra-organizational*) or across organization (*inter-organizational*) boundaries. Information systems should not be confused with information technology, since an information system involves information technology elements that cooperate with the organization's processes and components. In our time, the cornerstone of information systems is the *Internet* and *World Wide Web* (WWW) for business all over the globe, or a *Local Area Network* for local business. During the 1970s, the standard information transfer tool was TELEX, while the mainframe computer became the standard for the development and utilization of databases. At the same time, the development of *Management Information System* (**MIS**) programs emerged in order to meet the growing requirements of information-system managers. One can easily see from the above discussion that an information system does not involve only the information and other technology an organization uses, but also the interaction of this technology with the organization (humans) and the way in which the technology meshes with the business and economic processes of the organization. To facilitate their operations, today's large companies have the following Officer positions [179]: **CIO** (Chief Information Officer), **CEO** (Chief

Executive Officer), **CFO** (Chief Financial Officer), **COO** (Chief Operating Officer), and **CTO** (Chief Technical Officer). The Chief Information Security Officer (**CISO**) takes care of the information security within the organization and normally reports to the CIO. During the mid-1980s, the majority of manufacturing companies started to apply techniques from the IS field in order to meet their needs to forecast sales, take and implement orders, and distribute products. Today these operations are mainly implemented using the World Wide Web developed by *Berners-Lee* in 1989, based on the HTML protocol. This has opened new avenues in electronic data processing (**EDP**) and electronic data interchange (**EDI**).

### *5.4.2 General Structure and Types of Information Systems*

In a general way, an information system can be defined as a mix of interrelated subsystems that retrieve (or collect), process, store, and distribute information to support decision making and control in organizations. The outcome of an information system is addressed and used by those who desire to use it, e.g., clients, staff members, and generally citizens. Actually, an information system is a human social activity, which may not necessarily include the employment of computers.

The three principal activities of an information system are [111]:

- Input
- Processing
- Output

Thus, the general functional structure of an information system has the form shown in Fig. 5.19 [111]:

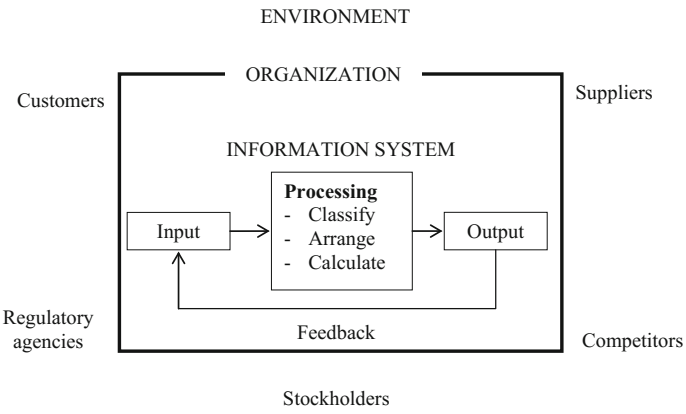



**Fig. 5.19** Structure and functions of information systems

The *input* (i.e., the raw-data acquired by the system) is *processed* and converted into more meaningful and usable packets of information and then passed to the users or activities (the *Output* in Fig. 5.21). This output is analyzed and, in case of any shortcomings, is fed back to the system for evaluation, correction, and refinement by the organization's relevant members. To be able to do this, these members must have good knowledge of management science and operational-research concepts and techniques in addition to information technology and computer-science techniques.

The operation of an organization is implemented in six hierarchical levels, which, in top-down sequence, are the following:

- Strategic level.
- Management level.
- Decision level.
- Knowledge level.
- Knowledge working level.
- Operational level

and are implemented by the appropriate managers and experts.

For each of these levels, the organization needs an appropriate information system. These systems' types in the same top-down hierarchy given above are as follows:

- Executive support systems (ESS).
- Management information systems (MIS).
- Decision-support systems (DSS).
- Office-automation systems (OAS).
- Knowledge work systems (KWS).
- Transaction-processing systems (TPS).

The above hierarchical operational levels and the corresponding information-system types are pictorially shown in the pyramid diagram of Fig. 5.20 [111].

Figure 5.21 shows a possible architecture example of MIS.

Besides these types of information systems, today we have available for use several other types of information systems that do not directly belong to the above system hierarchy. These new types of information systems include, but are not restricted to, the following [112, 113]:

- Data warehouses (DW).
- Enterprise resource planning (ERP).
- Enterprise systems (ES).
- Expert/knowledge-based systems (EKBS).
- Global information systems (GLIS).
- Geographic information systems (GIS).

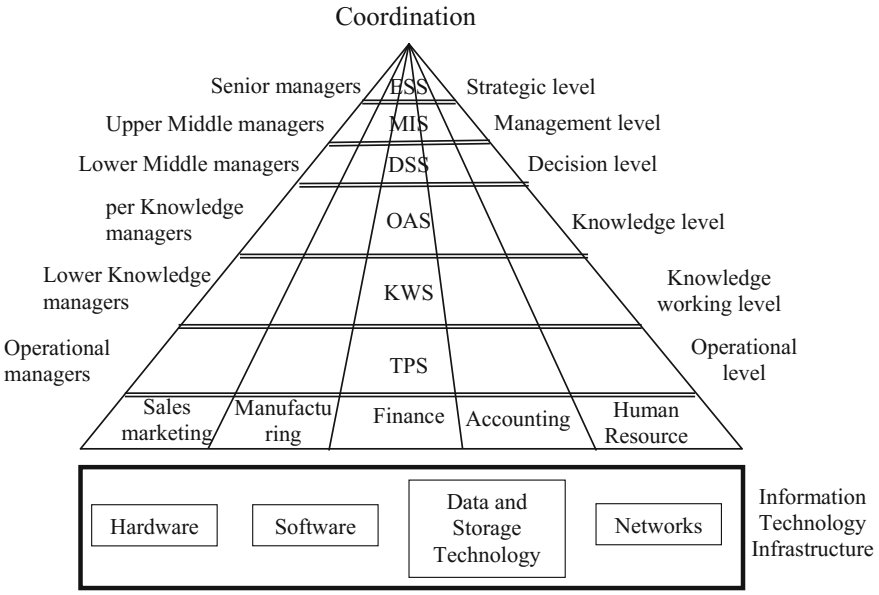



**Fig. 5.20** Information architecture of an organization

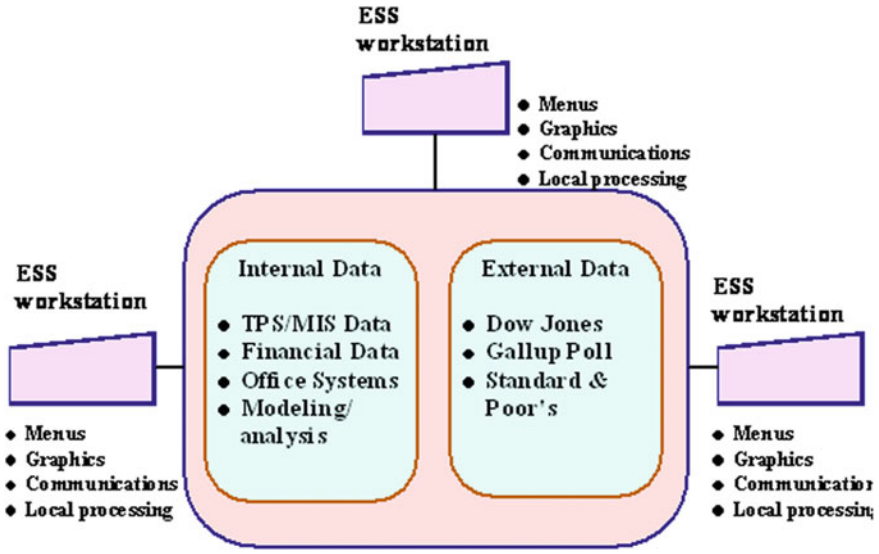



**Fig. 5.21** Structure of a management information system example (http://aiu.edu/publications/student/english/imgs/Management-Information-System_clip_image100.gif)

Regarding the information security issue, the organization should have the following components [114]:

- *Repositories* for permanent or temporary data storage (buffers, RAM, HD, etc.).
- *Interfaces* for supporting the interaction between computers and humans (keyboards, printers, scanners, etc.).
- *Channels* for connecting the repositories (cables, routers, etc.).

Modern organizations face the following challenges:

- **Strategic-business challenge**. This can be met with efficient information systems and information technology achievements.
- **Globalization challenge**. This can be met with the development of integrated multinational information systems suitable for trans-border dataflow and operations in many countries, as per the relevant laws.

### *5.4.3 Development of Information Systems*

The design and development of information systems is centered around the users, i.e., around their requirements, their performance features, and many other specifications. Therefore, the success or not of an information system is evaluated by its users within the organization on the basis of their satisfaction in using it. In order to be able to implement the desired integration of functions and departments of the organization, each department and subsystem must have access to a unique and unified database. In this way, each department obtains better knowledge and can appreciate better how its plans, decisions, and actions affect the operation of the other departments.

According to [115], any *information-systems development methodology* (**ISDM**) involves a collection of issues such as:

- Approach and philosophy
- Techniques
- Rules
- Procedures
- Tools
- Documentation
- Testing and refinement
- Training

If, as it is usually the case, the organization has an existing system, the process of ISDM is actually a system conversion, which involves the following major steps:

- Description of the system.
- Documentation of the input(s).
- Documentation of the output(s).

- Design of files and system processes.
- Flow charting of the program.
- Implementation of the program.
- Verification/refinement of the system
- Overall documentation.

Issues that must be addressed during system development include the following [111]:

- **System integrity** (subsystems integration, flexibility of the system, and expandability of the system).
- **Operational integrity** (skills of the systems operators and back-up for preventing breakdown when an equipment failure occurs or some key personnel are lost).
- **Internal integrity** (success of the system to perform its prescribed tasks, validity of system outputs, and system robustness to human and equipment errors).
- **Procedural integrity** (quality of documentation at all levels, appropriateness and human-friendliness of the system's procedures, and controllability of the procedures).
- **Change analysis** (i.e.,: suitable changes in the organization's activities in a concrete situation, good formulation of the problems before starting to solve them and before data-oriented work is carried out).

Process design is actually the first phase of classical system design. Its basic stages are:

- Requirements determination
- Analysis
- Logical design (flow charts)
- Physical design
- Implementation and validation
- Refinement(s)
- Maintenance

Today, knowledge-based/expert systems are embodied as an integral part of larger information systems and decision-support systems. The integration means that, instead of having separated hardware, software, and communications for each individual system, these systems are integrated into one facility. Integration can be at the development-tools level or at the application-system level. The two general forms of integration are:

- Functional integration
- Physical integration

*Functional integration* means that several different functions (tasks) are provided as a single system. For example, use of electronic mail, use of a spreadsheet, communication with databases, creation of graphics, etc., can all be performed at

the same workstation. The user can work with the aid of a unique consistent interface and can switch from one function or task to another and back again.

*Physical integration* is implemented through packaging of the hardware, software, and middleware features needed to accomplish functional integration. One can notice that physical integration involves several components, and it can have several configurations. The major approaches to physical integration are the following:

- **Access approaches** (Here, expert system development tools and/or application programs can access standard applications or development software. The three classes here are: single processor, multiprocessor, and networking).
- **Embedded approaches** (Here, the expert system is embedded in a conventional information system program, and so it embeds capabilities for value-added expert systems into standard programs. From the user's perspective, there is no distinction between the expert system and the parts of the conventional program).

The main areas of integration of expert systems are [116–119]:

- Integration with databases
- Integration with the interface subsystems
- Integration with decision-support systems
- Integration with factory automated systems
- Integration with error detection and recovery systems

A generic model for the design of information systems involves the following basic components:

- **Schemas** (Data model, fully normalized relations)
- **Programs** (Process model, fully normalized relations)
- **Data processing** (Data/process matrix entity life)

This model interacts with the *data dictionary* employed.

Historically, the information system development methods have evolved and have been developed in three generations:

- **First-generation methods** (structured programming, modular design and structure results, and programming style).
- **Second-generation methods** (diagram-oriented modeling, construction of checking of models, smoother path from requirements analysis to design and implementation stages, analysis techniques based on real-world needs on events and storage of information, and use of computer-aided software engineering (CASE) methods)
- **Third-generation methods** (high-level methods dealing with how individual analysis and design match together and interact, i.e., the philosophy of these methods is more concerned with the whole rather than the parts).

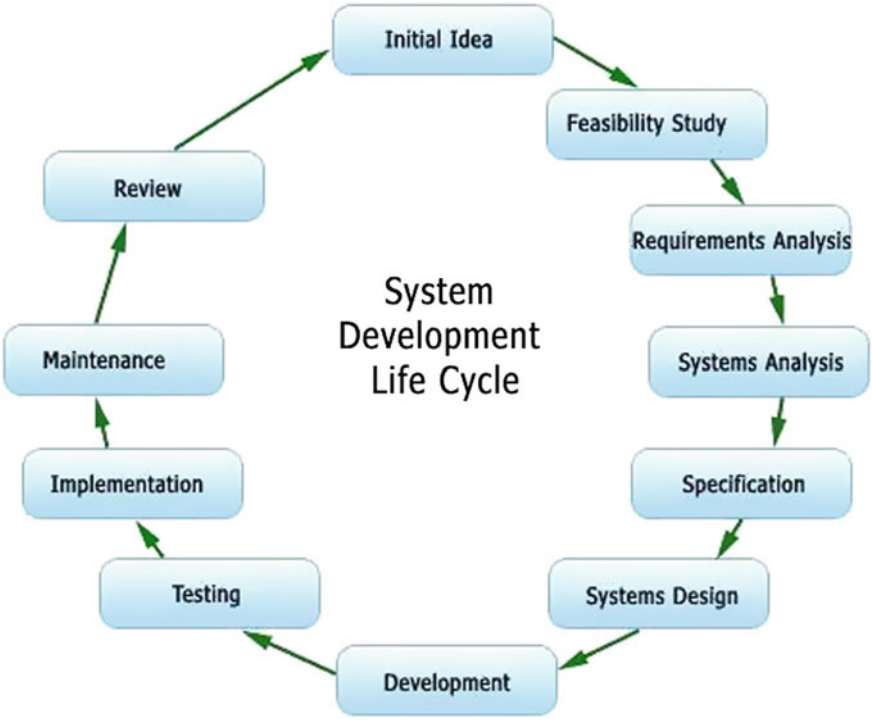



**Fig. 5.22** Flow chart of the system design/development cycle (http://shaparo-mis.wikispaces.com/file/view/system-development-life-cycle.png/187072855/405x342/system-development-life-cycle.png)

Figure 5.22 shows a self-explained pictorial/flow chart diagram of the information system design and development cycle.

## 5.5 Conclusions

This chapter, the second on information, was devoted to three modern axes of the information namely: *information science*, *information technology*, and *information systems*. These fields have had an enormous impact on human's current society and economic development. They are principal partners in all human activities from the education sector, science sector, agricultural sector up to the industrial and production sector, both on a stand-alone and an integrated basis.

Specifically, information science is mainly concerned with the documentation and dissemination of information and knowledge; information technology (or infotech) embraces all the methodologies and technologies that are used for the production, storage, processing, transmission, and dissemination of information;

and information systems apply information science and information technology concepts and tools in enterprises and organizations' everyday operation that require the symbiosis of technology with human-controlled actions and processes. The impact of information technology, with specific examples, on human life and society will be discussed in Part II of this book (Chap. 11).

## References


1. Information Science-Britannica Online Encyclopedia 2009, http://www.britannica.com/EBchecked/topic/287881/information-science/
2. B.R. Boyce, D.H. Kraft, Principles and theories in information science. Ann. Rev. Inf. Sci. Technol. **20**, 153–158 (1985)
3. J.L. Milstead (ed.), *ASIS Thesaurus of Information Science and Librarianship* (Information Today, Medford, NJ, 1998)
4. C. Zins, Redefining information science: From information science to knowledge science. J. Doc. **62**(4), 447–461 (2006)
5. C. Zins, Conceptions of information science: research articles. J. Am. Soc. Inf. Sci. Technol. **58**(3), 335–350 (2007)
6. C. Zins, Conceptual approaches for defining data, information and knowledge: research articles. J. Am. Soc. Inf. Sci. Technol. **58**(4), 479–493 (2007)
7. C. Zins, Knowledge map of information science: research articles. J. Am. Soc. Inf. Sci. Technol. **58**(4), 526–535 (2007)
8. C. Zins, Classification schemes of information science: twenty-eight scholars map the field. J. Am. Soc. Inf. Sci. Technol. **58**(5), 645–672 (2007)
9. S. Baruchson-Arbib, J. Bronstein, A view to the future of the library and information science profession: a Delphi study. J. Am. Soc. Inf. Sci. **53**(5), 397–408 (2002) (Available Online). http://onlinelibrary.wiley.com/doi/10.1002/asi.10051/abstract
10. Delphi method: principia cybernetica web, http://pespmcl.rub.ac.be/ASC/Delphi_metho.html
11. Information Technology Association of America (ITAA), http://www.itaa.org/es/docs/information%20Technology%20Definitions.pdf
12. M.A. Arbib, A.J. Kfoury, R.N. Moll, *A Basis for Theoretical Computer Science* (Springer-Verlag, Berlin, 1981)
13. M. Davis, E.J. Weyuker, *Computability, Complexity, and Languages: Fundamentals of Theoretical Computer Science* (Academic Press, New York, 1983)
14. R.L. Constable, *Computer Science: Achievements and Challenges, CIRCA* (2000). http://www.cs.cornell.edu/cis-dean/bgu.pdf
15. H. Abelson, G.I. Sussman, J. Sussman, *Structure and Interpretation of Computer Programs* (MIT Press, Cambridge, MA, 1996)
16. P.J. Denning, *Computer Science: The Discipline*. http://web.archive.org/web/20060525195404
17. Computer Sciences Accreditation Board (1997). http://www.csab.org/comp_sci_profession.html
18. Clay Mathematics Institute: P = NP, http://www.claymath.org/millenium/P_vsNP
19. L. Fortnow, The status of P versus NP problem. Commun. ACM **52**(9), 78–86 (2009)
20. S. Cook, in *The Complexity-Proving Procedures*. Proceedings of 3rd Annual ACM Symposium on Theory of Computing (Height, Ohio, May 1971) pp. 151–158
21. J. Van Leeuwen (ed.), *Handbook of Theoretical Computer Science* (Elsevier, Amsterdam, 1990)

22. M.R. Garey, D.S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness* (W.H. Freeman, New York, 1979)
23. C.H. Papadimitriou, *Computational Complexity* (Addison-Wesley, Reading, MA, 1993)
24. X.S. Yang, *Introduction to Computational Mathematics* (World Scientific Publishing, Singapore-London, 2008)
25. W.H. Press, *Numerical Recipes: The Art of Scientific Computing* (Cambridge University Press, Cambridge, 2007)
26. M.T. Heath, *Scientific Computing: An Introductory Survey* (McGraw-Hill, New York, 2002)
27. R.E. Crandall, *Topics in Advanced Scientific Computation* (Springer-Verlag, New York, Inc. (TELOS), Santa Clara, California, 1996)
28. Interactive Educational Modules in Scientific Computing (Prepared by M. Heath et.al). http://www.cse.illinois.edu/iem/
29. Computational Tools links, http://norma.mas.ecp.fr/wikimas/. Scientific Computing Software
30. Matlab: Parallel Computing Toolbox. http://www.mathworks.com/, http://www.mathworks.com/moler/
31. Mathematica 6, Scientific Computing World (2007). http://www.scientific-computing.com/products/review_details.php?review_id=17/
32. Brockport State College, http:www.brockport.edu/cps
33. Course Software-Scientific Computing with Matlab, http://dmpeli.mcmaster.ca/Matlab/Math1J03/Math1J03EntryPage.html
34. V. Eijkhout, *Introduction to High-Performance Scientific Computing* (e-book) (Victor Eijkhout, Austin, Texas, 2010). http://freecomputerbooks/com/introduction-to-High-Performance-Scientific-Computing.html
35. D. Metha, S. Sanhi, *Handbook of Data Structures and Applications* (Chapman and Hall, Boca Raton, FL, 2007)
36. M.T. Goodrich, R. Tamassia, Data Structures and Algorithms in Java, 4th edn. http://www.java4.dastructures.net/contents/contents/html
37. Oracle database 11 g, http://www.oracle.com/us/products/database/index.html
38. My SQL, http://www.mysql.com/why-mysql/marketshare/
39. Data Base Systems—Free books, http://www.database-books.us/db2.php
40. http://www.database-books.us/informix.php
41. A. Silberschatz, H.F. Korth, S. Sudarshan, *Database System Concepts* (McGraw-Hill, New York, 2010)
42. Free Books on Database Management and Training, http://www.techbooksforfree.com/database.shtml
43. Milo Free Textbook, *Computer Programming*, http:www.OSdata.com http://www.engin.umd.umich.edu/CIS/course.des/cis400/index.html
44. B. Hayes, The Semicolon Wars, American Scientists, http://www.americanscientists.org/template/AssetDetail/assetid/51982#52116
45. J. McCarthy, *What is Artificial Intelligence?* http://www-formal.stanford.edu/jmc/whatisai/whatisai.html
46. E.A. Feigenbaum, P. McCorduck, *The Fifth Generation* (Addison-Wesley, Reading, MA, 1983). http://www.worldcat.org/oclc/9324691
47. A. Barr, E.A. Feigenbaum, *Handbook of Artificial Intelligence* (Pitman, London, 1971)
48. E. Rich, *Artificial Intelligence* (McGraw-Hill, New York, 1984)
49. D. Popovic, V.P. Bhatkart, *Methods and Tools for Applied Artificial Intelligence* (Marcel Dekker Inc, New York, Basel, 1994)
50. T. Gruber, *What is an Ontology?* (2001). http://www-ksl.stanford.edu/kst/what-is-an-ontology.html
51. T. Gruber, Toward Principles for the Design of Ontologies Used for Knowledge Sharing. Int. J. Hum.-Comput. Stud. **43**(5–6), 907–928 (1995)
52. J. Stolze, D. Suter, *Quantum Computing* (Wiley, New York/Chichester, 2004)

53. E. Coelho, *Method Ontology* (1996). http://ksi.cpsc.ucalgary.ca/KAW/KAW96/coelho/node5.html
54. R. Forsyth, *Expert Systems* (Chapman and Hall, Boca Raton, FL, 1984)
55. R. Bowerman, P. Glover, *Putting Expert Systems into Practice* (Van Nostrand Reinhold, New York, 1988)
56. P. Harmon, R. Maus, W. Morrissey, *Expert Systems: Tools and Applications* (Wiley, New York/Chichester, 1988)
57. M. Mano, C. Kime, *Logic and Computer Design Fundamentals* (Pearson Prentice Hall, Upper Saddle River, N.J, 2008)
58. M. Mano, M. Ciletti, *Digital Design* (Pearson Prentice Hall, Upper Saddle River, NJ, 2007)
59. M. Rafiquzzaman, *Fundamentals of Digital Logic and Microcomputer Design (with CDROM)* (Wiley-Interscience, New York, 2005)
60. T. Moto-Oua, K. Fuchi, in *The Architectures of the Fifth Generation Computers*. Proceedings of 1983 IFIP Congress, (Paris, France, 1983), pp. 589–902
61. Webopedia: The Five Generations of Computers, http://www.webopedia.com
62. J.L. Baer, *Computer Systems Architecture* (Computer Science Press, New York, 1980)
63. D.M. Harris, S.L. Harris, *Digital Design and Computer Architecture* (Elsevier/Kaufmann, Amsterdam, 2007)
64. P.J. Koopman Jr., *Stack Computers*, http://www.ece.cmu.edu/~koopman/stack_computers/
65. Pipelining Concept in Computer Architecture, http://discuss.itacuments.com/index.php?topic=7635.0
66. P.M. Koge, *The Architecture of Pipelined Computers* (McGraw-Hill, New York, 1981)
67. L.D. Fosdick, C.J.C. Schauble, E.R. Jessup, *Tutorial: General Architecture of Vector Processors* (HPSC Group, University of Colorado), http://www-ugrad.cs.colorado.edu/~csci4576/VectorArch/Vector
68. M. Nielsen, I. Chuang, *Quantum Computation and Quantum Information* (Cambridge University Press, Cambridge, 2000)
69. D. Culler, J.P. Singh, A. Gupta, *Parallel Computer Architecture: A Hardware/Software Approach* (Morgan Kaufmann, San Francisco, 1998)
70. S.H. Roosta, *Parallel Processing and Parallel Algorithms: Theory and Computation* (Springer, Berlin, 2000)
71. Foster, *Designing and Building Parallel Programs* (Addison-Wesley Reading, MA, 1995). http://www.freetechbooks.com/designing-and-building-parallel-programs.html
72. Mitchell, Computing Power into the 21st Century: Moore's Law and Beyond. http://citeseer.ist.psu.edu/mitchell98computing.html
73. M.D. Hill, M.R. Marty, in Amdahl's Law in the Multicore Era. IEEE Computer Society, pp. 33–38, July 2008. http://www.cs.wisc.edu/multifaceted/amdahl/
74. B. Randel, The 1968/69 NATO Software Engineering Reports, Newcastle University. http://homepages.cs.ncl.ac.uk/brian.randel/NATO/NATOReports/index.html
75. J. Spence, I. Spence, Why we need a theory for software engineering. Dr. Dobbs J. 2 Oct 2009. http://www.drdobbs.com/open-source/224000375
76. R.H. Thayer (ed.), *Tutorial: Software Engineering Project Management* (IEEE Computer Society Press, Los Alamitos, Calif., 1999)
77. Computer Operating Systems, http://www.computerhope.com/os.htm
78. A.S. Tanenbaum, *Modern Operating Systems*. http://homepages.ucl.ac.uk/u0222323/Structure%20of%20operating
79. S. Heath, *Embedded Systems Design,* Paper back, 2nd edn. (Elsevier Science, Reprint Technical Science and Engineering, Amsterdam, 2008)
80. Embedded Computer Systems: CMU, http://www.ece.cmu.edu/~koopman/embedded.html
81. G.S. Brown, D.P. Campbell, Instrument engineering: growth and promise in process—control problems. Mech. Eng. **72**(2), 124 (1950)
82. R. Zurawski (ed.), *Embedded Systems Handbook: Embedded Systems Design and Verification* (CRC Press, Boca Raton, FL, 2009)

83. C.-M. Huang, J.-L. Chen, Y.-C. Chang, *Telematics Communication Technologies and Vehicular Networks: Wireless Architectures and Application* (Igi Global, Hershey, PA, 2009)
84. P. Nijkamp, G. Pepping, D. Banister, *Telematics and Transport Behaviour* (Springer-Verlag, Berlin/London, 1996)
85. M. Gott, *Telematics for Health: The Role of Telehealth and Telemedicine in Homes and Communities* (European Communities/Union (EUR-OP/OOPEC/OPOCE), Brussels, 1994)
86. Telematics Education and Learning, http://www.eadis.eu
87. Cellular Communications: Web ProForum Tutorials, The International Engineering Consortium, http://www.iec.org
88. J.E. Flood, *Telecommunications Networks* (IEE, London, U.K., 1997)
89. D.L. Shinder, *Computer Networking Essentials* (Cisco Press, Indianapolis, IN, 2002)
90. A.S. Tanenbaum, *Computer Networks* (Pearson Prentice Hall, Indianapolis, IN, 2002)
91. Network Types, http://www.wifinotes.com/types-of-network.html
92. S. Hekmat, *Communication Networks,* Pragsoft Corporation, Free Book Centre.Net. http://www.pragsoft.com/books/CommNetwork.pdf
93. S. Kiesler (ed.), *The Culture of the Internet* (Lawrence Erlbaum Associates, Mahwah, N.J, 1997)
94. K. Hartman, E. Ackerman, *Searching and Researching on the Internet and the World Wide Web* (Franklin, Beedle and Associates Inc., Wilsonville, OR, 2004)
95. M. Castells, *The Internet Galaxy: Reflections on the Internet, Business and Society* (Oxford University Press, Oxford, 2001)
96. R. Stout, *The World Wide Web Complete Reference* (Mc Graw-Hill, New York, 1996)
97. http://www.altavista.com
98. http://www.aboutus.org/InfoSeek.com, http://www.alexa.com/siteinfo/infoseek.com
99. http://download.cnet.com/Netscape-Navigator/3000-2356_4-10145004.htm
100. http://update.microsoft.com/windowsupdate/v6/thanks.aspx?ln=en&thankspage=5
101. http://www.internetworldstats.com/stats.htm
102. http://www.e-consultancy.com/
103. G. Gallego, *WBT: State-of-the Art in Web-Based Training*. http://www.gracespace.com/weblearn/state.htm#critical
104. S.G. Tzafestas (ed.), *Web-Based Control and Robotics Education* (Springer, Dordrecht/Heidelberg, 2009)
105. A. Banerji, A.M. Ghosh, *Multimedia Technologies* (Tata McGraw-Hill, New Delhi, India, 2010). http://www.mhhe.com/baneji/fmt
106. L. Makris, M.G. Strintzis, Multimedia Web Information Systems-World Wide Web History and Multimedia on the World Wide Web. http://encyclopedia.jrank.org/articles/pages/6841/Multimedia-Web-Information-Systems.html. Multimedia Web Information Systems—World Wide Web History, Multimedia on the World Wide Web (Net Industries and its Licensors)
107. D. Marshall, *Components of a Multimedia System* (2001). http://www.cs.cf.ac.uk/Dave/Multimedia/node16.html#SECTION0
108. R.E. Mayer, *The Cambridge Handbook of Multimedia Learning* (Cambridge University Press, Cambridge, 2005)
109. T.-P. Garrand, *Writing for Multimedia and the Web: A Practical Guide to Content Development* (Elsevier, Amsterdam, 2006)
110. Information Systems, http://www.uh.edu/∼mrana/try.htm#MMIS
111. K.C. Laudon, J.P. Laudon, *Management Information Systems* (MacMillan, London, 1988)
112. B. Langefors, *Theoretical Analysis of Information Systems* (Auerbach/Taylor and Francis Group, London, 1973)
113. D. Treek, R. Trobec, N. Pavesic, J.F. Tasic, Information systems security and human behaviour. Behav. Inf. Technol. **26**(2), 113–118 (2007). http://www.ingentaconnect.com/content/tandf/bit/2007/00000026/00000002/art00003
114. D.E. Avison, G. Fitzgerald, *Methodologies in Information Systems Development* (Blackwell Scientific Publications, Oxford, 1988)

115. R.W. Blanning (ed.), *Foundations of Expert Systems for Management* (Verlag TÜV Rheinland, Köln, 1990)
116. P.R. Watking, L.B. Eliot, *Expert Systems in Business and Finance: Issues and Applications* (Wiley, Chichester/New York, 1992)
117. S.G. Tzafestas, *Expert Systems in Engineering Applications* (Springer, Berlin, 1993)
118. A. Beerel, *Expert Systems in Business: Real World Applications* (Ellis Horwood, New York/London, 1993)
119. Modern Programming Languages, http://www.onesmartclick.com/engineering/programming-languages

# Chapter 6
# Feedback and Control I: History and Classical Methodologies

*Evolution is chaos with feedback.*

Joseph Ford

*Real meditation is not about mastering a technique; it's about letting go of control.*

Adyashanti

**Abstract** Feedback and control, the third fundamental element of life and society, is inherent in any stable and successfully operating system in the natural, biological, technological, or societal world. It is the fundamental mechanism that assures the achievement of system equilibrium and homeostasis. Very broadly speaking, we can say that feedback is any response or information about the result of a process that is achieved via available sensing elements. This chapter starts with an outline of the 'feedback' concept, illustrated by a set of biological examples, and followed by an exposition of the historical landmarks of feedback and control, including the achievements made from ancient times to the present. Then, an overview of the classical control methodologies is provided in a conveniently simple and coherent way. Specifically, the following concepts and methods are discussed with minimum mathematical detail: basic negative-feedback loop, stability, time-domain specifications, root locus, Nyquist, Bode and Nichols plots, frequency-domain specifications and stability criteria, compensator design in the time and frequency domains, and nonlinear systems analysis via the describing functions and phase-plane concepts. Actually, the chapter offers a good review of the field to enable the reader to see the role of feedback as a pillar of life and society, and it can be used as a quick reference source for all scientists interested in the field of feedback and classical control.

## 6.1 Introduction

This chapter is devoted to the third pillar of human life and society considered in the present book, i.e., to the concept or principle of "*feedback*". After *energy* (the food) and i*nformation* (communication and processing), *feedback* (control) is a must for



S. G. Tzafestas, *Energy, Information, Feedback, Adaptation, and Self-organization*, Intelligent Systems, Control and Automation: Science and Engineering 90, https://doi.org/10.1007/978-3-319-66999-1_6

the survival and equilibrium of any system in the natural, biological, technological, or societal world. Feedback control is the fundamental mechanism by which all systems of any kind sustain their equilibrium and homeostasis. Feedback over long periods of time is the means of adaptation and evolution of species and societies and the means for the balance in all manifestations of the life and ecosystem. *Norbert Wiener* in his 1948 book on *Cybernetics and Control* [1] explains that in voluntary action "feedback is the use of the difference (error) between a desired motion and the actual motion, as a new input (excitation) to bring the motion closer to the desired pattern". To make an effective action on the environment, we must have good effectors. These effectors must be properly monitored back to the central nervous system, and the readings of these monitors must be properly combined with other information sent by the sense organs to produce a suitably proportioned output to the effectors.

*Feedback is actually the means of designing any stable and successful system in technology and society.* The chapter is organized as follows. Section 6.2 provides an outline of the "*feedback*" concept (positive feedback, negative feedback) and a set of biological examples in which feedback is an inherent property. Then, in Sect. 6.3, the concept of *feedback control system* is discussed via operational diagrams. Section 6.4 presents the principal historical eras of feedback and control studies describing the achievements made, with the names of the respective inventors and investigators. Then, Sects. 6.5–6.11 provide an overview of the classical control methodologies with mathematical detail sufficient for the purposes of this book. Full accounts of this field are available in the literature, which includes books of length ranging from 300 to 400 pages to more than 1000 pages [2–6]. Here, the following absolutely necessary topics for the purposes of this book are discussed, namely: the basic-control loop, system stability, system performance specifications, second-order systems, root-locus, the Nyquist method, Bode plots, Nyquist stability criterion, Nichols plots, discrete-time systems, classical continuous-time and discrete-time compensator design (phase-lag, phase lead), Ziegler-Nichols, PID controller tuning, and nonlinear control-system analysis (describing functions, phase plane). The next chapter will be devoted to modern control methodologies.

## 6.2 The Concept of Feedback

### *6.2.1 General Definition*

According to the *Columbia Encyclopedia* (2008 edition): "feedback is the arrangement for the automatic self-regulation of an electrical, mechanical, or biological system by returning part of its output as input". Extending this definition, we can state that "*feedback* is any response or information about the result of a process". In all cases "feedback" is used to achieve and maintain a desired performance of the system or process at hand (natural or man-made, biological, etc.).

For example, in a management control system, "feedback" is a concept equivalent to "information about actual performance" in comparison to the planned performance, i.e., feedback is the process by which a given system or model is tested to see if it is performing as planned. Via timely feedback, it is possible to exert quickly corrective action(s) when the situation goes out of hand.

Other simple examples of feedback are the following:

- If the speed of an engine exceeds a preset value, then the governor of the engine reduces the supply of the fuel, thus decreasing the speed.
- In a thermostat, the actual temperature is compared with the desired temperature and the feedback enables reduction of the difference.
- In car-direction (driving) control, the driver senses the actual direction of the car and compares it with the desired direction on the road. If the actual direction is to the left of the desired one, the driver turns it a little to the right. In the opposite case, the turn is to the left.
- The temperature of a healthy human body is maintained in a very narrow range through the use of feedback (homeostasis).

### 6.2.2 *Positive and Negative Feedback*

*Feedback* can be divided *positive feedback* and *negative feedback*. To introduce formally the concept of feedback (positive and negative), we start with the notion of a *system*. A system is any *transformation of inputs to outputs*, as shown in Fig. 6.1.

The inputs are usually the outcome of the system's environment influence on the system, and the outputs are the results (product) generated by the system due to these inputs. In other words, the inputs are the "*cause*" and the outputs are the "*effect*", which implies that there is a time separation between inputs and outputs. The inputs are applied "before", and the outputs are obtained "after" the lapse of some time.

A system with feedback (known as *feedback system*) is pictorially shown in Fig. 6.2.

In Fig. 6.2, we can see that there is a closed loop between input-system-output through the feedback, which is also known as *closed-loop control*. The feedback is realized through a *sensor* that measures the output and sends this information back as the input of the system.

*Positive feedback* occurs when a change in the output is fed back to the system input such that the output changes even more in the same direction, which results in a continuing spiral of change (rheostasis). This implies that each pass around the

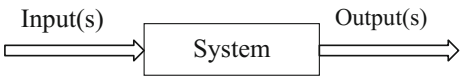



**Fig. 6.1** The general concept of "system"

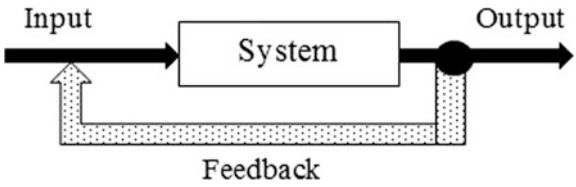



**Fig. 6.2** Schematic of a feedback system

feedback cycle magnifies the change instead of diminishing it. Positive feedback is used only in systems in which the amplification of a signal is needed. For example, in electronic systems positive feedback (of a suitable magnitude) is needed in order to obtain an *oscillator,* which produces a sinusoidal signal of constant magnitude.

*Negative feedback* occurs when a corrective action is applied in order to "*damp*" or "*reduce*" the magnitude of an "*error*", so that constant conditions are maintained inside the system, e.g., biological body (known as *homeostasis*). All the examples given in Sect. 6.2.1 are of *negative feedback* (*or deviation reducing feedback*).

In general (except from the cases where positive feedback is balanced by an equivalent amount of negative feedback, as in the oscillator example), *positive feedback* leads to *divergent behavior,* and, if a positive feedback loop is left to itself, it can lead to indefinite expansion or an explosion, which may result in the destruction of the system. Some examples of positive feedback are nuclear chain reactions, proliferation of cancer cells, industrial growth, and population explosion.

*Negative feedback* leads to the desired regulation (i.e., adaptive/goal-seeking behavior) maintaining the level of the output variable concerned (direction, speed, temperature, fluid level, material concentration, voltage level, etc.).

The goal (reference) to be sought may be:

- **Self-determined** (i.e., with no human involvement), as for example in the case of the *ecosystem,* which maintains the composition of the air, or the ocean, or the level maintenance of glucose in the *human blood*.
- **Human-determined** where the goals of the technological (man-made) systems are set by humans in order to satisfy their needs and preferences. This is actually the case of all servomechanisms and man-made automatic control systems (industrial, social, enterprise, economic).

Some more biological examples of positive feedback are the following [7, 8]:

- **Generation of nerve signals** The membrane of a nerve fiber causes a small leakage of sodium ions through sodium channels which produces a change in the potential of the membrane. This, in turn, causes more opening of channels, and so on. Thus, a small initial leakage leads to a sodium-leakage explosion which produces the *nerve action potential*.
- **Blood clotting** An injured tissue releases signal chemicals that activate platelets in the blood. Then, the activated platelets release chemicals to activate further platelets which cause a fast cascade and the formation of a blood clot.

- **Morphogenesis and growth** In general, most positive feedback in an organism initiates the fast auto-excitation of endocrine and nervous-systems element and plays a fundamental role in morphogenesis, growth, and the development of organs.
- **Economic systems** (e.g., stock markets) have both positive- and negative-feedback loops (mechanisms) which are based on cognitive and emotional factors that fall within the framework of *behavioral finance*.
- **Education** *Feedback* is also endogenous in *education* and is usually termed *reinforcement*. A correct answer or understanding of a concept by the student is reinforced. An educational curriculum is monitored by the relevant *Accreditation Board*, which gives feedback to the institute concerned in order to modify and improve the curriculum (on the basis of this feedback).
- **World-system development** *Positive feedback* was the mechanism that has caused the demographic growth and technological advancement in the world-system development. Technological development increases the carrying capacity of land for human demographic growth, which leads to more potential inventors and more technological progress, which results in further growth of the Earth's carrying capacity for people, and so on.

## 6.3 The Concept of Control

*Control* is tightly connected and inseparable from *feedback*, although in some special cases we use *feedforward (non-feedback) control* with great attention and increased care. Norbert Wiener said that "*the present time is the age of Communication and Control*" [1].

Control, like feedback, is present in one or the other form in all physical, biological, and man-made systems. A control system always involves two subsystems:

- The subsystem or process to be *controlled* (symbolically **P**).
- The subsystem that controls the first, called *controller* (symbolically **C**).

A control system contains always two processes:

- Action
- Perception

The controller **C** acts on **P** and changes its state in a desired (conscious or unconscious) way. The process under control **P** informs **C** by providing to it a *perception* of the state of **P**.

Therefore, control is always a kind of *action-perception* (i.e., *feedback*) *cycle* as shown in Fig. 6.3.

To illustrate in some more detail how a feedback control system operates, we will work on the operational diagram shown in Fig. 6.4.

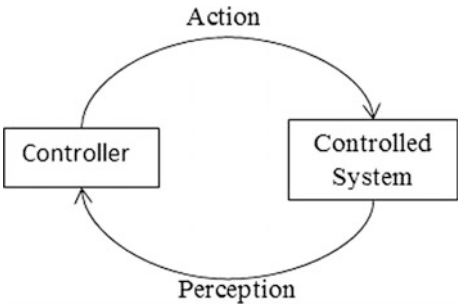



**Fig. 6.3** A feedback control (action-perception) cycle

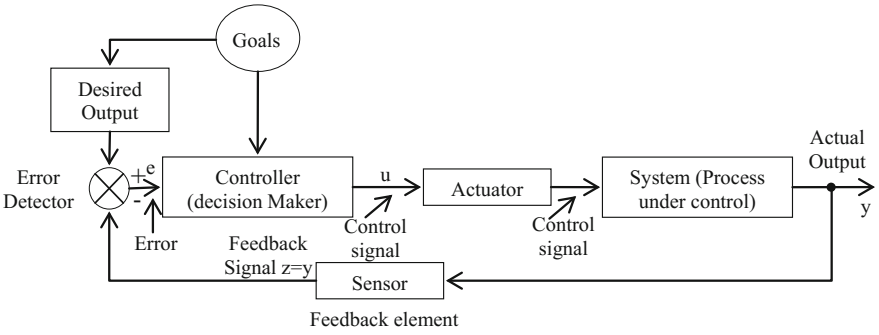



**Fig. 6.4** Operational diagram of a typical feedback-control system

- The *perception* about the state of the system (process) under control is provided by one or more appropriate *sensors* which constitute the *feedback elements*.
- The controller receives and processes the error between the desired and the actual output (provided by an *error detector*) and sends the *control action* (signal) to the system through a suitable *actuator*.
- The *actuator* or *effector* produces the control action in accordance with the control signal and imposes it upon the system.

The goals are set by the owner, the designer, or the user of the system, or by nature in the case of biological systems. In order for the control to be effective, the decisions and control actions must be exerted without or with a very small time delay.

Otherwise special care is needed. The fact that the above system involves *negative feedback* is indicated by the negative sign in the feedback path, which produces the error signal $\varepsilon = x - y$. If the actual output $y$ is greater than the desired output $x$, then the error $\varepsilon$ is negative, and the controller-actuator pair exerts a negative effect on the system, forcing it to decrease the actual output $y$ towards the desired output and reduce the error. If $y < x$, then $\varepsilon > 0$ and the action imposed to the system is positive so as to increase $y$ and again reduce the error. The above type of controller, which produces a control signal proportional to the actual error, is

called *Proportional Controller* (P). Other types of control use either the integral of the error (*Integral Controller:* I) or the derivative of the error (*Derivative Controller:* **D**). In practice, we usually use combinations of the above controllers namely: **PI, PD,** or **PID**. A controller which is also used frequently in practice is the two-valued controller. Here, the control signal $u$ takes a value in accordance with the *signum* of the error, which changes in a desired sequence of time instants, called *switching times*. If the two values of $u$ are 0 and 1, the controller is called an *on-off controller:* if they are −1 and +1, the controller is called a *bang-bang controller*.

A good representative example of a negative-feedback control system is the speed control of an electric motor shown in Fig. 6.5.

The input potentiometer is of the rotating type. The motor axis is unloaded (or loaded). The operational (block) diagram of the system (see Fig. 6.4) has the form shown in Fig. 6.6.

Specific mathematical developments at a suitable minimal level for the purposes of the present book will be provided in Sects. 6.5–6.11. The big question at the present is the same as in the past: "*Who is to control whom in regards to what aspects of life?*" This question will be discussed in Chap. 12.

## 6.4 Historical Landmarks of Feedback and Control

Feedback and control are inherent and endogenous in the life Earth, and, during the centuries up to a certain time, they have been in use by humans in their activities, but unconsciously. According to the literature [9–14], the history of automatic control is divided into the following principal periods:

- Prehistoric and early control period: until 1900
- Preclassical control period: 1900–1935
- Classical control period: 1936–1955
- Modern control period: 1956–Present.

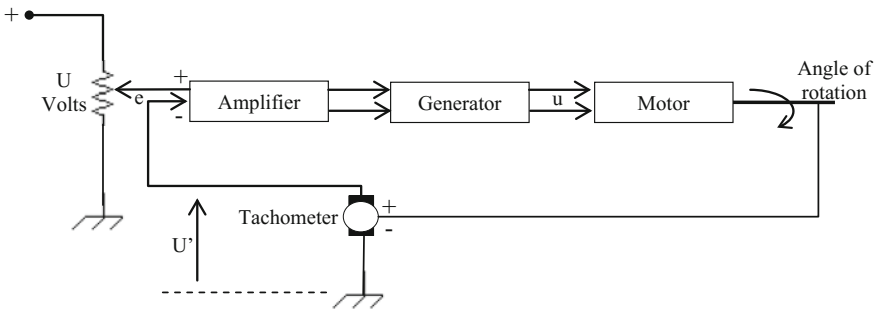



**Fig. 6.5** A negative feedback system controlling the speed of a motor

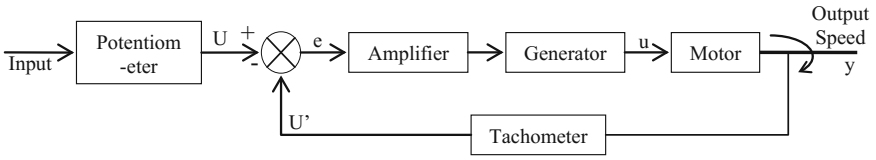



**Fig. 6.6** Operational diagram of the motor-speed control system

### *6.4.1 Prehistoric and Early Control Period*

This period essentially starts with the work of the Greek engineer Ktesibios and involves the following.

- **Ktesibios of Alexandria** (Ctesibius 285–222 B.C.) He used flow-rate control in the famous *water clock* (about 250 B.C.). This was achieved by using a floating valve to maintain the water level in a vessel at a constant depth. Constantly flowing water at the base of the vessel flows into a second tank, the level of which was proportional to the time elapsed. Ktesibios work marks the beginning of intentional control-system design, and, although he left no written documents, the water clock has been reconstructed via the accounts of Vitruvius, a Roman engineer.
- **Heron of Alexandria** (1st century B.C.) in his work "Pneumatica" described several primitive control systems or automata for entertainment purposes, water supply in bathhouses, automatic gate opening in temples, the automatic dispensing of wine, etc.
- **Pseudoarchimedes** (9th century A.C.). This name was given to an anonymous writer who left the first written texts on feedback-control devices.
- **Al-Sacati, Al-Azari** and **Musa** (13th century A.C.). Three brothers who wrote two documented books on feedback, where "on-off" control mechanisms were also included. The use of floating-valve control disappeared after the $13^{th}$ century and only reappeared during the industrial revolution.
- **William Salmon** and **Sutton Thomas Wood** (1570–1790). The revival of floating-valve control was done by William Salmon, who constructed a steam engine which used a floating valve to control the water level in the boiler. In 1784, Sutton Thomas Wood constructed in England a similar engine based on the ideas of Heron. In this way, the floating valve again found application, today constituting one of the most popular techniques.
- **Cornelius Drebbel** and **Bonnemain**. Cornelius Drebbel (1572–1663), a German engineer, constructed for the first time a *thermostat* based on the measurement of the temperature by the expansion of a liquid contained in a vessel connected to a U tube containing mercury. A float in the mercury controlled an arm which in turn controlled the draft to a furnace and hence the rate of combustion and heat output. Drebbel did not leave any documentation, but his work was published by Francis Bacon. In 1783, the French engineer Bonnemain (1743–1828) constructed improved temperature- control systems, especially his "*regulateur de feu*", independently of Drebbel's ideas.

- **Edmund Lee.** He built in 1745 the first automatic windmill that pointed continuously into the wind, using a small fan mounted at right angles to the main wheel of the windmill.
- **James Watt** and **Matthew Boulton.** Matthew Boulton (1728–1809) and James Watt (1736–1819) constructed a large mill to demonstrate the capabilities of the new rotating machine of Watt. The first person to design a centrifugal pendulum for negative-feedback control of the speed of a wind mill was Thomas Mead. The well- known *Watt's governor* (also called *fly-ball governor*) was designed and used in 1789. This governor has been improved in several ways during the first three quarters of the 19th century, e.g., William Siemens in 1846 and 1853, Charles Porter in 1858, Thomas Pickering in 1862, and William Hartnell in 1872.
- **James Clerk Maxwell** (1831–1879). Maxwell studied the stability of Watt's governor and published in 1868 his celebrated paper entitled "On Governors". He derived the linear differential equations (through linearization) for a number of governor systems and determined their characteristic polynomials. He showed that the stability of second-, third-, and fourth- order systems can be determined by examining the coefficients of the differential equations (or characteristic polynomials), deriving the necessary and sufficient conditions.
- **Edward Routh** and **Adolf Hurwitz.** Routh (1831–1907) studied deeply the stability problem formulated by Maxwell, and published his first results in 1874, and, in 1877, he derived the algebraic stability criterion, known today as the *Routh criterion*. In 1895, Adolf Hurwitz (1859–1919) derived the stability criterion independently in a different manner *(*the *Hurwitz criterion)*. The two alternative formulations are equivalent and represent what today is called *Routh–Hurwitz* criterion. The development of control theory that followed this stability study was facilitated by inventions based on electricity, such as the electric relay (a high-gain power amplifier) and the spring-based solenoid (a simple, proportional control device).
- **A.M. Lyapunov** (1857–1918). He studied the stability of nonlinear differential equations using a generalized concept of energy and formulated what is now called *Lyapunov stability criterion* (in 1892). His work was known and applied only in Russia until the early 1960s, when its importance was appreciated in the West.
- **O. Heaviside**. During the period 1892–1898 he developed the *operational calculus* and studied the transient behavior of systems through a concept similar to what we presently call the *transfer function*.

### *6.4.2 Pre-classical Control Period*

Pre-classical control period began in about 1900 with the application of feedback controllers to electrical, mechanical, hydraulic, pneumatic, thermal, and other man-made systems. However, during the first two decades of the 20th century,

these devices were designed without firm knowledge of the dynamics and stability problems involved. An exception to this was the automatic ship-steering system designed by E. Sperry in 1910.

**Elmer Sperry.** He invented the *gyroscope,* which originally (1911) was used for ship-steering feedback control and later in aircraft control.

**N. Minorsky** (1885–1970). In 1922, he made one of the first applications of nonlinear control for steering ships through the use of proportional-plus-integral-plus-derivative (PID) control (3-term control), which he introduced for the first time. He developed this control law by observing how a helmsman steered a ship. Minorsky's work became widely-known after his publications in *The Engineer* during the late 1930s.

**Bell Telephone Labs:** During the 1920s and 1930s, the Bell Telephone Laboratories investigated and applied frequency-domain methods of P.-S. de Laplace (1749–1827), J. Fourier (1768–1830), and A.L. Cauchy (1789–1857) in communication systems (as mentioned in Sect. 4.3).

**Harold Stephen Black** (1898–1983): A serious problem in the development and use of long-distance communication (telephony) during the early 1900s was the so-called amplification problem. Although improvements in cables and impedance loading had extended the distance for reliable transmission without amplification, the transcontinental transmission still required periodic amplification to amplify the voice signal, which however was also amplifying the noise, thus reducing the quality of service. H.S. Black was the first who showed and demonstrated (in 1927) that, by using negative-feedback amplifiers, noise amplification could be reduced. This technique was first implemented within AT&T in 1931.

**Harry Nyquist** (1889–1976): A collaborator of Black, in 1932 he published a paper titles "*Regeneration Theory*", which presented the frequency-domain stability criterion, called now *Nyquist criterion,* based on the polar plot of a complex function (transfer function).

- **Clesson E. Mason:** In 1928, Mason started experiments with feedback using the flapper-nozzle amplifier invented by Edgar H. Bristol at Foxboro Company. He developed in 1930 a feedback circuit, which linearized the valve operation enabling the introduction of an integral (reset) component in the controller.
- **Harold Locke Hazen** (1901–1980): Hazen designed a complete servosystem and studied deeply the performance of servomechanisms. His 1934 publication "*Theory of Servomechanisms*" is considered to be the starting point for the next period of control, called "*classical control period*". He coined the term "*servomechanism*" from the words "servant" and "mechanism".

### *6.4.3 Classical Control Period*

The previous control period was marked and ended by the three developments already mentioned, i.e.:

- Negative-feedback electronic amplifier
- Linearized pneumatic controller
- Design and study of servomechanisms

The first brief segment of the classical control period (1935–1940) was marked by intensive work and new advances in the USA and Europe. The key contributors of this initial period are the following:

- **Hendrick Bode:** In 1938, he used magnitude and phase-frequency plots (in terms of the frequency logarithm) and investigated the stability of the closed-loop system using the concepts of *gain margin* and *phase margin*. He adopted the point $(-1, 0)$ as the critical point for measuring these margins.
- **J.G. Ziegler** and **N.B. Nichols:** They developed the now called "*Ziegler–Nichols*" tuning methods for selecting the optimal parameters in PI and PID controllers. Their results were published in their 1942 book "*Network Analysis and Feedback Amplifier Design*".
- **Gordon S. Brown** and **A.C. Hall:** Assisted by the members of their group at MIT, they used constant-amplitude (M) and constant-phase (N) circles to determine estimates of the closed-loop behavior in the time domain. A.C. Hall, while working in the Radiation Laboratory of MIT, employed the frequency-domain technology developed at Bell Labs to tackle noise and design a control system for an airborne radar (1946). Nichols in 1947 developed this *Nichols Chart* for the design of feedback-control systems.
- **W.R. Evans:** In 1948, developed his *root-locus* technique, which provided a direct way to determine the closed-loop pole locations in the s-plane. During the 1950s, control-design theories were mainly based on the s-plane approach and the transient step response specifications of the rise time, overshoot, settling time, IAE, ISE, ITAE, etc.
- **Norbert Wiener:** While working at MIT, he studied stochastic systems through the frequency-domain stochastic analysis and developed the celebrated *Wiener statistical optimal filter* (in 1949) for the case of continuous-time signals, which improved the signal-to-noise ratio of communication systems. The case of discrete-time stochastic processes was studied by the Russian A.N. Kolgomorov (1941).
- **Donald McDonald:** His 1950 book "Non-Linear Techniques for Improving Servo Performance" inspired during the 1950s extensive research on the time-optimal control problem of single-input single-output systems using saturating control. The work pursued on this topic was briefly described by Oldenburger in his 1966 book entitled "Optimal Control". Other contributors in this area included Bushaw (1952), Bellman (1956), and J.P. LaSalle (1960), who showed that the *bang-bang* type is the time-optimal control.

The achievements of the classical control period (1935–1955) were disseminated through many textbooks. These include, in chronological order, the books of the following authors: Bode (1940); Kolmogorov (1941); Smith (1942); Gardner and Barnes (1942); Bode (1945); MacColl (1945); James, Nichols and Philips (1947);

Laner, Lesnik, and Matson (1947); Brown and Campbell (1948); Wiener (1949); Porter (1950); Chestnut and Mayer (1951, 1955); Tustin (1952); Thaler and Brown (1953); Nixon (1953); Evans (1954); Bruns and Saunders (1954); Truxal (1954); Thaler (1955). More details concerning the developments made in the classical control period (1930–1955) are given in [15], and for the period 1800–1930 in [16].

### *6.4.4 Modern Control Period*

The *Modern Control Period* starts with the works of R. Bellmann, L.S. Pontryagin, R. Kalman, and M. Athans.

**Richard Bellman:** During the period 1948–1953, Bellman developed "*dynamic programming*" via his "*principle of optimality*". The word "programming" was selected as a more accurate term than "planning". In 1957, he applied dynamic programming to the optimal control problem of discrete-time (sampled-data) systems and derived the optimal-control algorithm in backward time as a multistage decision making process [17]. The main drawback of dynamic programming is the so-called "*curse of dimensionality*", which was partially faced by Bellman and his co-worker Stuart Dreyfus via the development of carefully designed numerical-solution computer programs. The dimensionality problem is still an obstacle despite the considerably increased computer power of present-day supercomputers. The work of Bellman was carried out in the RAND Corporation while working on the allocation of missiles to targets to achieve maximum damage.

**Lev S. Pontryagin:** By 1958, Pontryagin has developed his *maximum principle* for solving optimal-control problems on the basis of the *calculus of variations* developed in the middle seventies by L. Euler. The optimal controller of Pontryagin was of the on/off relay type [18].

**Rudolf E. Kalman:** At the Moscow IFAC Conference (1960), Kalman presented his paper: "On the General Theory of Control Systems" in which he introduced the duality between multivariable feedback control and multivariable feedback filtering, thus showing that, having solved the control problem, one has also solved the filtering problem, and vice versa [19]. In the same year (1960), he published his seminal paper on the optimal-filtering problem for discrete-time systems in the state space [20] and his joint paper with J.E. Bertram on the analysis and design of control systems via Lyapunov's second method [21]. The Lyapunov method was extended in 1960 by J.P. Lasalle [22]. The continuous-time version of the Kalman filter was published in 1961 jointly with R. Bucy [23].

**Michael Athans:** He published in 1966 one of the first comprehensive books on optimal control jointly with P. Falb, which played a major role in research and education in the field [24]. He has published since then numerous publications with important results on optimal and robust multivariable control, many of which are included, along with other new results, in [25].

**V.M. Popov:** In 1961, he developed the so-called *circle criterion* for the stability analysis of nonlinear systems [26]. His work was extended and applied in following

years by many researchers, including I.W. Sandberg in 1964 [27], K.S. Narendra in 1964 [28], C.A. Desoer in 1965 [29], et al.

**Charles Stark Draper:** In 1960, he developed an *inertial navigation system* for vehicles moving in space, such as ships, aircrafts, or spacecrafts. This system uses gyroscopes to give accurate estimates of the position of the moving body and was first used in the Apollo guidance system [30].

Further important developments on modern control theory since the 1970s were made by many scientists all over the world and can be found in the control literature. Some of these contributors in alphabetic order are the following:

Ackerman, J., Anderson, B.D.O., Astrom, K.J., Balakrishnan, Bensoussan, A, Bertsekas,D., Brockett, R.W., Butkovskii, A. Cruz, J.B., Francis, G.F., Haddad, A.H., Horowitz, I., Houpis, C.H., Jacobson, D.H., Kailath, T., Kwakernaak, H., Lainiotis, D., Landau, I.D., Larson, R., Lewis, F.L., Lions, J.L., Luenberger, D.G., MacFarlane, A.G.J., Mayne, D.Q., Meditch, J.S., Mendel, J.M., Meystel, A., Narendra, K.S., Sage, A.P., Saridis, G.N., Seinfeld, J., Sheridan, T.B., Singh, M., Slotine, J.J., Spong, M.W.,, Titli, A., Tou, J.T., Unbehauen, H., Utkin, V.L., Vamos, T., Yurkovich, S., Wolovich, W.A., Wonham, L. Zadeh, W.A., Zames, G., Ziegler, B.P.

Looking at this literature, one can see that what is collectively called "*modern control*" has evolved along two main avenues:

- Optimal, stochastic, and adaptive control.
- Algebraic and frequency-domain control.

All techniques of these avenues are based on the assumption that a mathematical model of the system to be controlled is available, and all together they are called "*model-based*" control techniques. Another family of control techniques does not use any mathematical model; instead they employ what are called "*intelligent*" processes to specify how the system works and this bypass the lack of a model. These techniques are named "*model-free*" control techniques [31–38]. An important key concept in the whole modern control theory is "*observability* and *controllability*" which reveal the exact relationships between transfer functions and state-variable representations. The basic tool in optimal-control theory is the celebrated matrix *Riccati* differential equation which provides the time-varying feedback gains in a linear-quadratic control-system cell.

The fundamental concepts upon which the multivariable frequency methodology is based are the "*return ratio*" and "*return difference*" quantities of Bode. These concepts, together with the "*system matrix*" concept of Rosenbrock, were used to develop the multivariable (vectorial) Nyquist-theory, as well as the algebraic modal (pole shifting or placement) control theory. Rosenbrock introduced several new concepts, such as the diagonal dominance, characteristic locus, and inverse Nyquist array [39, 40].

The theory of sampled-data control systems was firmly established by G. Franklin, J. Ragazzini, and E. Jury who formularized the "*Jury*" *stability criterion* [41–44].

The field of robust control for systems with uncertainties was initiated by I. Horowitz [45–47], leading to the so-called "*quantitative feedback theory*" (QFT). Important contributions in this area were made by C. Houpis, A. MacFarlane, I. Postlethwaite, M. Safonov [48–55], et al.

The stochastic control area, after the seminal work of Kalman, was sustained by K. Astrom, B. Wittenmark, H. Kwakernaak, and J. Meditch [56–65]. The adaptive control area, especially the model-reference approach, was initiated by I. Landau [66–68].

Modern non-linear control theory has evolved, after the work of Popov and Lyapunov, through the work of G. Zames, K. Narendra, J.-J. Slotine [69–75], and others [76–78]. In the following sections, the classical analysis and design methodologies will be briefly discussed for both continuous-time and discrete-time (sampled-data) systems.

## 6.5 Classical Control

### *6.5.1 Introductory Issues*

The *classical-control* methodology is primarily restricted to linear time-invariant *single-input/single output* **(SISO),** or *monovariable,* systems which are described by linear scalar differential equations with constant coefficients. In classical-control methodology also belongs the study of SISO nonlinear systems through the *describing function* and *phase-plane* techniques. Classical control theory has naturally evolved using the concepts of *transfer function*, (harmonic) *frequency response*, and the methods of Nyquist and Bode. The identification of the system (i.e., the determination of the constant coefficients in the differential equation or the transfer function representation) was possible to be done by experimentally measuring the frequency (amplitude and phase) response or the parameters of the step response (rise time, overshoot, settling time, etc.). Actually, all these methods, as well as the *root locus* method, were based on *hand-using graphical techniques,* requiring particular skills and intuition. The design of robust controllers against internal or external disturbances and measurement noise was performed using the concepts of *gain* and *phase margin*. The Nyquist, Bode, and root-locus techniques are offering efficient ways for describing closed-loop performance in terms of open-loop parameters and features. Naturally, a direct application of SISO classical control techniques to *multi-input/multi-output* **(MIMO)** or *multivariable* systems was very difficult due to the interaction of the multitude control loops existing in these systems. Actually, a trial-and-error approach was required with many iterations but without guaranteeing overall satisfactory results and stability. This fact has motivated the development of the multivariable frequency-domain and quantitative feedback theory as mentioned in Sect. 6.4.

### 6.5.2 The Basic Feedback Control Loop

A *linear time-invariant* **(LTI)** SISO physical system is generally described by a differential equation (DE) of the form:

$$\sum_{k=0}^{n} a_k \frac{\mathrm{d}^k y(t)}{\mathrm{d}t^k} = \sum_{k=0}^{m} b_k \frac{\mathrm{d}^k u(t)}{\mathrm{d}t^k} \quad (m \le n)$$

where $a_k$ and $b_k (k = 0, 1, \ldots)$ are constant coefficients and $a_n \neq 0$ (which may have the value $a_n = 1$).

Application of the Laplace transform (see Table 6.1):

$$\bar{y}(s) = \int_0^\infty y(t)\mathrm{e}^{-st}\mathrm{d}t,$$

where $s = a + j\omega$ is the Laplace complex variable, on both sides of the above DE, and assuming zero initial conditions for $u$ (the input) and $y$ (the output) gives:

$$A_n(s)\bar{y}(s) = B_m(s)\bar{u}(s)$$

where:

$$A_n(s) = \sum_{k=0}^{n} a_k s^k = a_n s^n + a_{n-1}s^{n-1} + \cdots + a_1 s + a_0$$

$$B_n(s) = \sum_{k=0}^{m} b_k s^k = b_m s^m + b_{m-1}s^{m-1} + \cdots + b_1 s + b_0$$

**Table 6.1** Laplace transforms $\bar{x}(s)$ of some continuous-time functions $x(t)$

| $x(t)$ | $\bar{x}(s)$ | $x(t)$ | $\bar{x}(s)$ |
|---|---|---|---|
| $\delta(t)$ | 1 | $te^{-at}$ | $1/(s+a)^2$ |
| $\delta(t-nT)$ | $\mathrm{e}^{-nsT}$ | $1-\mathrm{e}^{-at}$ | $a/s(s+a)$ |
| $a^{t/T}$ | $1/[s-(1/T)\ln a]$ | $t-(1-\mathrm{e}^{-at})/a$ | $a/s^2(s+a)$ |
| $u(t)$ | $1/s$ | $\sin\omega_0 t$ | $\omega_0/\left(s^2+\omega_0^2\right)$ |
| $t$ | $1/s^2$ | $\cos\omega_0 t$ | $s/\left(s^2+\omega_0^2\right)$ |
| $t^2/2$ | $1/s^3$ | $e^{-a}\sin\omega_0 t$ | $\omega_0/\left[(s+a)^2+\omega_0^2\right]$ |
| $t^{m-1}/(m-1)$ | $1/s^m$ | $\sinh\omega_0 t$ | $\omega/\left(s^2-\omega_0^2\right)$ |
| $e^{-at}$ | $1/(s+a)$ | $\cosh\omega_0 t$ | $s/\left(s^2-\omega_0^2\right)$ |

This is an algebraic equation which, is solved for $\bar{y}(s)$, gives:

$$y(s) = G(s)\bar{u}(s)$$

where:

$$G(s) = \frac{\bar{y}(s)}{\bar{u}(s)} = \frac{B_m(s)}{A_n(s)}$$

is the *transfer function* of the system and graphically yields the block-diagram representation of Fig. 6.7.

The polynomial $B_m(s)$ is called the *input polynomial,* and $A_n(s)$ is called the *characteristic polynomial*. The roots of the equation $B_m(s) = 0$ are called the *zeros* of the system, and the roots of the equation $A_n(s) = 0$ (the *characteristic equation*) are the *poles* of the system.

Using the block diagram representation, the basic feedback SISO control loop has the form shown in Fig. 6.8. The symbols $G_c(s)$,$G_a(s)$, $G_p(s)$, and $F(s)$ represent, respectively, the transfer functions of the controller, the actuator, the system (process, plant) under control, and the feedback element (which usually is a sensor or measurement device).

The symbols $\bar{c}(s)$, $\bar{e}(s)$, $\bar{y}_f(s)$, $\bar{d}(s)$ and $\bar{y}(s)$ represent the *Laplace transforms* of the command signal $c(t)$, the error signal $e(t) = c(t) - y_f(t)$, the feedback signal $y_f(t)$, the disturbance signal $d(t)$, and the output signal $y(t)$.

Assuming for the moment that no disturbance exists ($\bar{d}(s) = 0$), combing the controller, actuator and system as:

$G(s) = G_c(s)G_a(s)G_p(s)$ (Forward transfer function),
and using the equations:

$$\bar{y}_c(s) = G(s)\bar{e}(s), \quad \bar{e}(s) = \bar{c}(s) - \bar{y}_f(s) = \bar{c}(s) - F(s)\bar{y}_c(s)$$

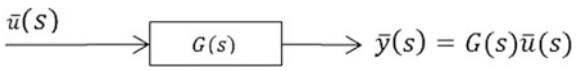



**Fig. 6.7** The block diagram of a linear SISO system

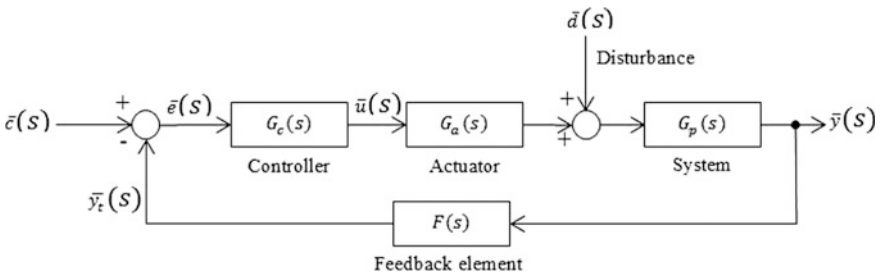



**Fig. 6.8** Basic control loop for a negative -feedback controlled system $G(s)$

where $\bar{y}_c(s)$ is the component of $\bar{y}(s)$ due only to $\bar{c}(s)$, we find the closed-loop transfer function:

$$H_c(s) = \frac{\bar{y}_c(s)}{\bar{c}(s)} = \frac{G(s)}{1 + G(s)F(s)}$$

In the same way, assuming that $\bar{c}(s) = 0$, we find:

$$H_d(s) = \frac{\bar{y}_d(s)}{\bar{d}(s)} = \frac{G_p(s)}{1 + G(s)F(s)}$$

where $\bar{y}_d(s)$ is the component of $\bar{y}(s)$ due only to $\bar{d}(s)$. The total output is: $\bar{y}(s) = \bar{y}_c(s) + \bar{y}_d(s)$ (Superposition principle).

Suppose that $\bar{c}(s) = 1/s$, which is the Laplace transform of the unit step signal $c(t) = 1$, $t \geq 0$, $c(t) = 0$, $t < 0$. Then, the goal of the control design is to select the controller $G_c(s)$ so that $y_c(t) \to 1$ and $y_d(t) \to 0$ as $t \to \infty$ (for various types of disturbance) with *acceptable overshoot* and *acceptable steady state error* $e_{ss} = \lim_{t\to\infty} e(t) = \lim_{s\to\infty} s\bar{e}(s)$.

The available methods for achieving this goal are: Evans *root locus method,* Nyquist *plots method,* Bode *diagrams method,* and Nichols diagram *method*.

A closed-loop (feedback) system with an overall *forward transfer function* $G(s)$ and a *feedback transfer function* $F(s)$ is called a feedback system in the *canonical form*. Defining $G(s)F(s)$ (the so-called *open loop transfer function*), in terms of its zeros $\mu_k$ and poles $\pi_k$ we get:

$$G(s)F(s) = \frac{KQ_1(s)}{s^p Q_2(s)} = \frac{K(s-\mu_1)\cdots(s-\mu_m)}{s^p(s-\pi_{p+1})\cdots(s-\pi_n)}$$

A system with $p$ pure integrations (i.e., with $\pi_0 = \pi_1 = \cdots = \pi_p = 0$) is said to be a *type-p* system. The *steady-state error* $e_{ss}$ of a unity feedback system (under the assumption that it is a stable system) is given by:

$$e_{ss} = e(\infty) = \lim_{s\to 0} s\bar{e}(s) = \lim_{s\to 0} \frac{s\bar{x}(s)}{1 + G(s)F(s)}$$

where $\bar{x}(s)$ is the applied input.

For the steady-state error, we have the following:

- **Position error $e_p$**: This is the steady-state error obtained when the input is a unit step function $(x(t) = 1 \quad \text{for } t > 0, x(t) = 0 \quad \text{for } t \leq 0)$, i.e.:

$$e_p = \lim_{s\to 0} \frac{s}{1 + G(s)F(s)} \frac{1}{s} = \frac{1}{1 + K_p}$$

where $K_p$, called the *position-error constant*, is equal to:

$K_p = \lim_{s\to 0} G(s)F(s) = KQ_1(0)/Q_2(0)$ For $p = 0$, and $K_p = \infty$ for $p \geq 1$.

- **Velocity error $e_\upsilon$**: This is the steady-state error obtained when the input is a unit ramp $(x(t) = t,\ \ t \geq 0)$, i.e.:

$$e_\upsilon = \lim_{s\to 0} \frac{s}{1 + G(s)F(s)} \frac{1}{s^2} = \frac{1}{K_\upsilon}$$

where $K_\upsilon$, called the *velocity-error constant*, is given by:

$$K_\upsilon = \lim_{s\to 0} sG(s)F(s) = \lim \frac{K}{s^{p-1}} \frac{Q_1(s)}{Q_2(s)} = 0$$

for $p = 0$, $K_\upsilon = KQ_1(0)/Q_2(0)$ for $p = 1$, and $K_\upsilon = \infty$ for $p \geq 2$.

- **Acceleration error $e_a$**: This is the steady-state error for a unit parabolic input $(x(t) = t^2/2, t \geq 0)$, i.e.:

$$e_a = \lim_{s\to 0} \frac{s}{1 + G(s)F(s)} \frac{1}{s^3} = \frac{1}{K_a}$$

where $K_a$, called the *acceleration constant,* is given by:

$$K_a = \lim_{s\to 0} s^2 G(s)F(s) = \lim \frac{K}{s^{p-2}} \frac{Q_1(s)}{Q_2(s)} = 0$$

for $p = 0, 1$, $K = KQ_1(0)/Q_2(0)$ for $p = 2$, and $K_\upsilon = \infty$ for $p \geq 3$.

### 6.5.3 System Stability

Stability is one of the most fundamental concepts of control. A system is considered *absolutely stable* if, for any initial conditions, the output of the *system*:

- Is bounded for all time $0 < t < \infty$.
- Returns to its equilibrium position (point, state) when the time tends to infinity $(t \to \infty)$.

A system is said to be "*asymptotically stable*" if its *impulse response* (i.e., the response to an impulsive input) has finite absolute value (norm) and goes to zero as $t \to \infty$, i.e.

1. $|h(t)| < \infty$ for any $t \geq 0$
2. $\lim_{t \to \infty} |h(t)| = 0$

where $h(t) = \mathcal{L}^{-1}H(s)$, $H(s)$ being the system transfer function, and $\mathcal{L}^{-1}(\cdot)$ the inverse Laplace transform.

For a system with input $u(t)$, irrespectively of whether it is an external input, a control input or a disturbance, the stability concept is changed to *bounded input–bounded output* (BIBO) stability. A system is BIBO stable if it, for any *bounded input,* produces a bounded output, i.e.:

$$\begin{aligned} &\text{If} |u(t)| \leq u_{\max} < +\infty \quad \text{for any } t > 0 \\ &\text{Then} |y(t)| \leq y_{\max} < +\infty \quad \text{for any } t > 0. \end{aligned}$$

Then $|y(t)| \leq y_{\max} < +\infty$ for any $t > 0$.

The two stability concepts, viz. asymptotic stability and BIBO stability, are equivalent. The consequence of the above two stability definitions is the position of the poles of a stable system, for which we have the following property:

*A linear continuous-time system is BIBO stable if and only if the poles of its transfer function $H(s)$ belong strictly to the left hand semi plane s, i.e., if all of its poles have negative real parts.*

To check in a direct way if a system is absolutely stable on the basis of the above definitions and property, we need to determine the impulse response (i.e., the inverse Laplace transform of the transfer function), which needs the knowledge of the system and is a time consuming process for high-dimensional systems.

The algebraic stability criteria of *Routh* and *Hurwitz* provide ways to check the stability of a system using directly the real coefficients (parameters) of the characteristic polynomial:

$$\chi(s) = a_n s^n + a_{n-1} s^{n-1} + \cdots + a_1 s + a_0$$

by constructing the so-called *Routh table* and the *Hurwitz determinants* [5, 6].

### 6.5.4 System Performance Specifications

The goal of designing a feedback-control system is to achieve a desired overall performance in the time and frequency domains. The system performance can be categorized into two kinds:

- Steady-state performance
- Transient-response performance.

Steady-state performance is measured by the *steady-state errors* of the system, when its output is desired to follow a constant-position (step) reference input, a velocity (ramp) input or a constant acceleration (parabolic) input. Small steady-state

error requires a large gain, but this implies that the system in nearer to the maximum allowable gain for guaranteed stability or equivalently that the *gain margin* is small.

The transient response performance is typically expressed by the step-response parameters, such as the overshoot $h$, the delay time $t_d$, the rise time $t_{rise}$, the dominant time constant $\tau_{dom}$, the response time $t_{resp}$, and the settling time $T_s$ (Fig. 6.9).

The *overshoot* $h$ is defined as:

$$h = 100 \times \frac{y_{max} - y_{ss}}{y_{ss}} \%$$

where $y_{max}$ is the maximum value of the response and $y_{ss}$ the steady-state value.

The *rise time* $t_{rise}$ is the time needed for the response $y(t)$ to go from 10 to 90% of $t_{ss}$.

The *response time* $t_{resp}$ is the time at which the step response goes back to $y_{ss}$ after its overshoot.

The *dominant time constant* $t_{dom}$ is defined as the time needed by the exponential envelope to reach 63% of its steady state value.

The *settling time* $T_s$ is defined to be the minimum time required for the response to enter a window $\eta = \pm 2\%$ or $\pm 5\%$, centered at the steady-state value.

In the case of first-order and second-order systems, the step response has the forms:

$$y(t) = Ce^{-at} \text{and } y(t) = Ce^{-at} \sin(\omega_0 t + \phi), \quad a > 0,$$

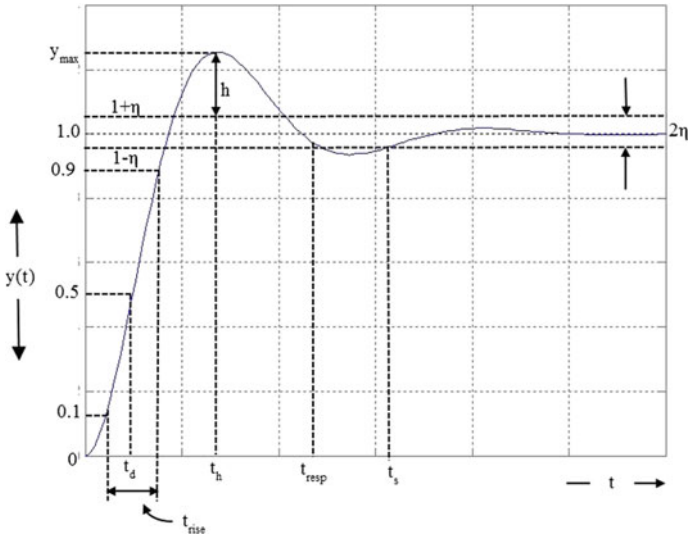



**Fig. 6.9** Typical form of step response with the main parameters

respectively. In both cases, the exponential factor $e^{-at}$ determines the time constant $\tau_{\text{dom}}$ as the time $t$ at which $at = 1$, i.e.:

$$\tau_{\text{dom}} = 1/a$$

At $\tau_{\text{dom}}$, we have $\mathrm{e}^{-a\tau_{dom}} = \mathrm{e}^{-1} = 1/e = 0.37 = 1 - 0.63$. Thus, the definition $\tau_{\text{dom}} = 1/\alpha$ follows.

The steady-state and transient performances can be expressed in a unified way by the so-called *generalized performance* indexes (criteria) **IAE** *(Integrated Absolute Error)*, **ISE** *(Integrated Squared Error)*, and **ITAE** *(Integrated Time by Absolute Error)* defined as:

$$\textbf{IAE}: V_1 = \int_0^\infty |e(t)|\mathrm{d}t; \quad \textbf{ISE}: V_2 = \int_0^\infty \mathrm{e}^2(t)\mathrm{d}t; \quad \textbf{ITAE}: V_3 = \int_0^\infty t|e(t)|\mathrm{d}t$$

### 6.5.5 Second-Order Systems

Second-order systems, although very simple, play a dominant role in automatic control theory because they involve all three elements of nature, namely *inertia, friction,* and *elasticity*. For example, a mechanical system with mass $m$, linear friction coefficient $\beta$, and a spring with elastic coefficient $k$ (see Fig. 6.10), is described by

$$\left(mD^2 + \beta D + k\right)y(t) = f(t), D = \mathrm{d}/\mathrm{d}t$$

where $y(t)$ is the mass displacement, and $f(t)$ the external force. The transfer function of this system is

$$\bar{y}(s)/\bar{f}(s) = 1/\left(ms^2 + \beta s + k\right)$$

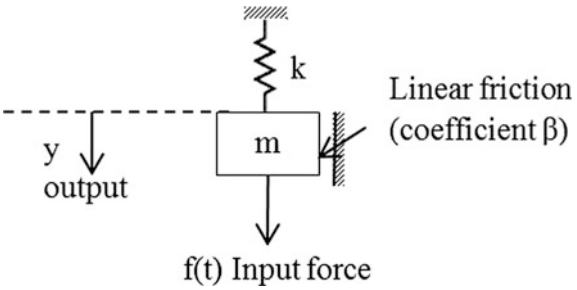



**Fig. 6.10** A mass m attached to a spring and moving with linear friction (proportional to velocity)

The characteristic polynomial of the system can be written in the canonical form:

$$ms^2 + \beta s + k = m\left(s^2 + 2\zeta\omega_n s + \omega_n^2\right)$$

where $\zeta$ (the damping ratio) and $\omega_n$ (the undamped natural cyclic frequency) are given by $2\zeta\omega_n = \beta/m$ and $\omega_n^2 = k/m$.

Assuming that $\zeta = \beta/2m\sqrt{k/m} = \beta/2\sqrt{km} < 1$ (underdamping), the poles of the system, i.e., the roots of the characteristic equation, $s^2 + 2\zeta\omega_n s + \omega_n^2 = 0$, are:

$$\pi_1 = -a + j\omega_0, \quad \text{and} \quad \pi_2 = -a - j\omega_0$$

where:

$$a = \zeta\omega_n \text{ (System damping)}$$
$$\omega_0 = \omega_n\sqrt{1-\zeta^2} \text{ (Natural frequency under damping)}$$

If the force $f(t)$ is a step function $u(t)$, then $\bar{y}(s)$ is given by:

$$\bar{y}_u(s) = 1/s(s-\pi_1)(s-\pi_2)$$

The step response is found to be:

$$y_u(t) = \mathsf{L}^{-1}y_u(s) = 1 - (\omega_n/\omega_0)e^{-at}\sin(\omega_0 t + \Psi)$$

where $\tan\Psi = \omega_0/a$. Graphically, the two conjugate poles $\pi_1$ and $\pi_2$ on the s-plane have the positions shown in Fig. 6.11.

From Fig. 6.11 we find that:

$$\cos\Psi = a/\omega_n = \zeta$$

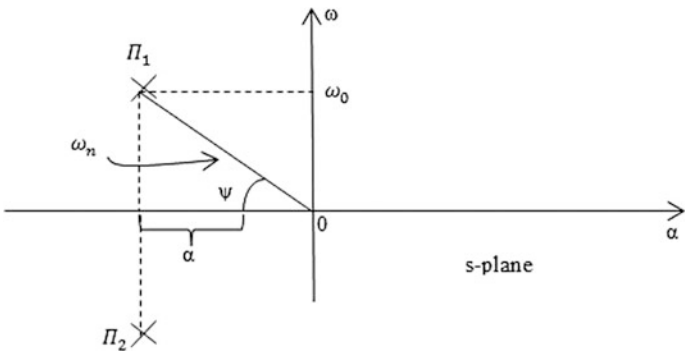



**Fig. 6.11** A pair of conjugate complex poles

Working on the expression of $y_u(t)$, we find that the (first) overshoot $h$ and the time $t_h$ (see Fig. 6.9) are given by:

$$h = e^{-\zeta\pi/\sqrt{1-\zeta^2}}, \quad t_h = \pi/\omega_n\sqrt{1-\zeta^2}$$

The settling time $T_s$ for accuracy $\eta = \pm 2\% = \pm 0.02$ can be found by equating the damping (envelope) waveform $e^{-at}$ at $t = T_s$ with 0.02, i.e. $e^{-\zeta\omega_n T_s} = 0.02$. This gives the relation $\zeta\omega_n T_s = 4$.

These expressions give the values of $h$, $t_h$, and $T_s$ in terms of the system parameters $\zeta$ and $\omega_n$ (or equivalently in terms of the original system parameters $m$, $\beta$, and $k$). In practice, a high-order system (higher than order two) can be approximated by the second-order system that has the dominant pair of conjugate poles (i.e., the poles with the smaller damping $\alpha_d = \zeta\omega_n$ which lie nearer to the imaginary axis). The above formulas are now applied to the dominant second-order system. This approximation is very good if the damping $\alpha$ of all the other poles satisfies the inequality:

$$|\alpha| \geq 10|\alpha_d| = 10|\zeta\omega_n|$$

The dominant time constant $\tau_d$ is equal to $\tau = 1/|\alpha|$.

## 6.6 The Root-Locus Method

The *root-locus* method of Evans provides an algebraic-geometric way for determining the positions of the closed-loop poles from the positions of the open-loop poles for different values of the open-loop gain $K$, i.e., of the parameter $K$ in the expression:

$$G(s)F(s) = K\frac{Q_1(s)}{Q_2(s)} = K\frac{(s-\mu_1)\cdots(s-\mu_m)}{(s-\pi_1)\cdots(s-\pi_n)}, \quad m \leq n$$

The closed-loop transfer function poles are the roots of the closed-loop characteristic equation $1 + G(s)F(s) = 0$, or:

$$Q_2(s) + KQ_1(s) = 0$$

Their position on the complex plane $s$ changes with the changes of $K$. The locus (line) on which the closed-loop poles lie (move), when $K$ changes from $K = 0$ to $K \to \infty$, is called the *root locus*. For $K = 0$, the closed-loop poles coincide with the open-loop poles (i.e., the roots of $Q_2(s) = 0$). For $K \to \infty$, the closed-loop poles tend to the positions of the open-loop zeros (i.e., the roots of $Q_1(s) = 0$).

In general, for the *root locus* to pass through a point $s_0$ of the complex plane $s$, $s_0$ must be a root of the characteristic equation: $Q_2(s_0) + KQ_1(s_0) = 0$, or equivalently:

$G(s_0)F(s_0) = KQ_1(s_0)/Q_2(s_0) = -1$. From this equation, we get the following two conditions:

$$|G(s_0)F(s_0)| = |K||Q_1(s_0)/Q_2(s_0)| = 1 \text{ (Magnitude condition)}$$

$$\underline{|G(s_0)F(s_0)} = \underline{|KQ_1(s_0)/Q_2(s_0)} = 180^\circ + p360^\circ \ , \ p = 0, \pm 1, \pm 2, \ldots \text{ (Phase condition)}$$

or, equivalently:

$$|Q_1(s_0)/Q_2 s_0| = 1/|K|$$

$$\underline{|Q_1(s_0)/Q_2(s_0)} = \begin{cases} (2p+1)180^\circ \textit{ for } K > 0 \\ 2p180^\circ \textit{ for } K < 0 \end{cases} \quad p = 0, \pm 1, \pm 2, \ldots$$

The *root locus* (i.e., the line of phase $(2p+1)180^\circ$ for $K > 0$) and the *line* (circle) of *constant gain* $K = 10$ of the feedback system with $G(s)F(s) = K/[s(s+1)(s+2)]$ are shown in Fig. 6.12.

The magnitude and phase of $G(s)F(s)$ can be determined graphically with the aid of a number of rules. These rules concern the following features of the root locus, which is symmetric about the real axis:

- Number of branches (each branch departs from an open-loop pole for $K = 0$ and arrives at an open-loop zero when $K = \infty$).
- Number, center, and angles of asymptotes.
- Branches lying on the real axis.
- Points at which two or more branches meet and separate their paths.

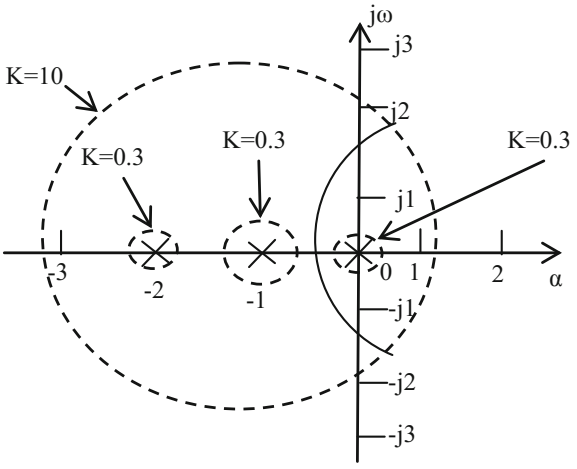



**Fig. 6.12** Root locus and constant-gain locus $K = 10$ of the system $GF(s) = K/s(s+1)(s+2)$

Today, there are excellent software packages that can be used to draw the root locus (e.g., MATRIX, Matrix-X, Control-C, MATLAB, etc.).

The *root-locus* helps to determine directly the following parameters and features of the closed-loop system:

- The *critical frequency* $\omega_c$ and *critical gain* $K_c$ for which the system passes from stability to instability.
- The system-transfer function (and the time response).
- The *gain margin* and *phase margin.*
- The gain $K$ required for the damping factor $\zeta$ to have a desired value, and vice versa. To this end we draw the straight line passing from the origin and having an angle $\Psi$, for which $\cos\Psi = \zeta$ (or $\Psi = \arccos\zeta$), with the negative real axis.

It is remarked that if the number of open-loop poles exceeds the number of (finite) open-loop zeros by three or more, then there always exists a critical value of the gain $K$ beyond which the root-locus enters the right-hand semi plane $s$, and so the system becomes unstable.

## 6.7 Frequency-Domain Methods

### *6.7.1 Nyquist Method*

The *Nyquist, Bode* and *Nichols methods* are frequency-domain methods. They are stronger than the algebraic methods of Routh and Hurwitz (because they provide information for the gain and phase margins, like the root-locus method).

Given a transfer function $G(s)$, we get the function $G(j\omega)$ by setting $s = j\omega$, i.e., by restricting the values of $s$ on the imaginary axis. The function $G(j\omega)$ is a complex function of the real variable $\omega$ and can be written in one of the following equivalent forms:

$$G(j\omega) = A(\omega) + jB(\omega) \quad \textit{Cartesian form}$$
$$G(j\omega) = |G(j\omega)|\underline{\phi(\omega)} \quad \textit{Polar form}$$
$$G(j\omega) = |G(j\omega)|[\cos\phi(\omega) + j\phi(\omega)] \textit{Euler form}$$

where $A(\omega) = \mathrm{Re}[G(j\omega)]$ (real part), $B(\omega) = \mathrm{Im}[G(j\omega)]$ (imaginary part), and $\tan\phi(\omega) = B(\omega)/A(\omega)$ (phase), $M(\omega) = |G(j\omega)| = \sqrt{A(\omega)^2 + B(\omega)^2}$ (magnitude).

*Direct polar diagram* (plot) or simply *polar plot* of $G(jw)$ is its graphical representation on the plane $G(j\omega)$ for values of $\omega$ from $\omega = 0$ up to $\omega = +\infty$ (i.e., when $\omega$ traverses the positive imaginary axis). The polar plot of $G(j\omega)$ is the same no matter which representation form is used (Cartesian, polar, Euler). The polar plot of a simple RL circuit with transfer function:

$$G(j\omega) = \frac{R}{R + j\omega L} = \frac{1}{1 + j\omega\tau}, \quad \tau = \frac{L}{R}$$

has the circular form shown in Fig. 6.13.

The form of polar plots of a general system with:

$$G(j\omega) = \frac{K(1 + j\omega\tau_a)\cdots(1 + j\omega\tau_p)}{(j\omega)^n(1 + j\omega\tau_1)\cdots(1 + j\omega\tau)}$$

for $n = 0$ (type 0), $n = 1$ (type 1), $n = 2$ (type 2), and $n = 3$ (type 3) is given in Fig. 6.14.

We now define the *Nyquist path* as any positive (clockwise) closed curve of the plane $s$ that covers the entire right-hand $s$ semi plane. The Nyquist path can have

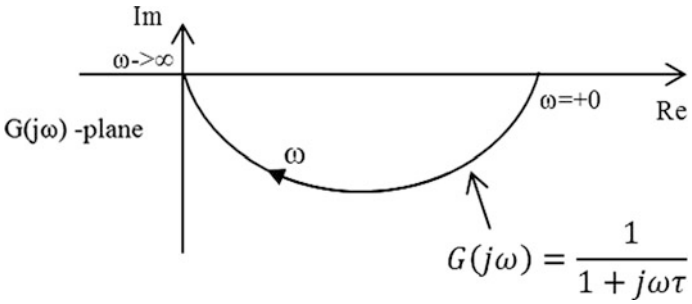



**Fig. 6.13** Polar plot of $1/(1 + j\omega\tau)$

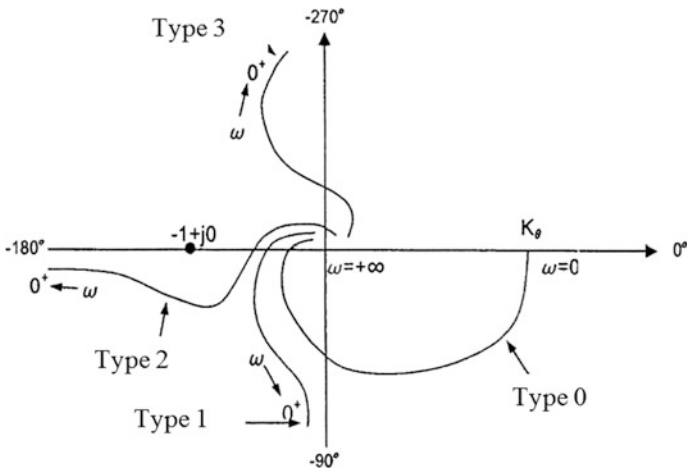



**Fig. 6.14** General form of polar plots for systems of type 0 up to type 3

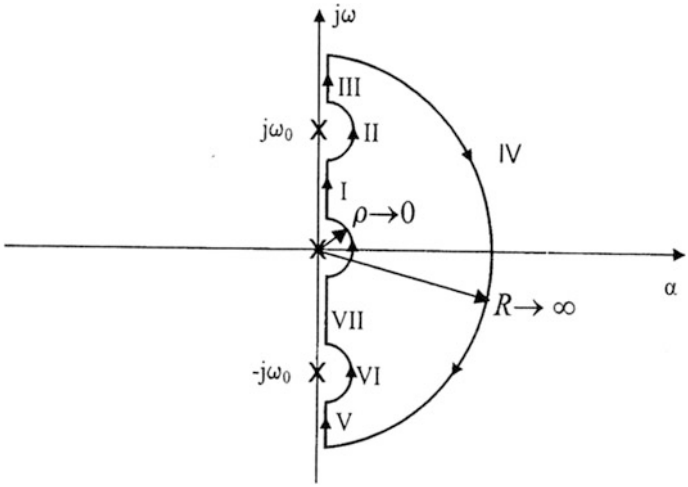



**Fig. 6.15** Typical form of the Nyquist path

any shape. The usual shape used in practice is the one shown in Fig. 6.15 which has a semi-circular form. Any poles of the system lying on the imaginary axis are avoided (not included in the Nyquist path) by drawing small semi-circumferences with radii $\rho \to 0$.

*Complete Nyquist plot* (or simply *Nyquist plot*) is called the mapping of the entire *Nyquist path* on the plane $G(s)$. It is noted that symmetric parts of the Nyquist path correspond to symmetric parts of the Nyquist plot on the plane $G(s)$. Thus it suffices to draw the Nyquist plot for the positive imaginary semi axis and then to draw its symmetrical part. Two examples of Nyquist plots are given in Fig. 6.16.

As already described, for a closed-loop control system to be (absolutely) stable, the poles of the closed-loop characteristic polynomial $\chi(s) = 1 + G(s)F(s)$ must lie on the *left half semi plane s*. Equivalently, no closed-loop poles should be enclosed by the Nyquist path.

### The Nyquist-Stability Criterion

A closed-loop system with open-loop transfer function $G(s)F(s)$ is stable if and only if

$$N_p = -P_r \leq 0$$

where $N_p$ is the number of positive (clock-wise) encirclements of the Nyquist-plot of $G(s)F(s)$ around the Nyquist point $(-1, 0)$, and $P_r$ is the number of poles of $G(s)F(s)$ on the right-hand semi plane $s$ [352, 353, 356].

- If $N_p \geq 0$, the system is *unstable*, and the number $Z_r$ of the roots of $\chi(s) = 1 + G(s)F(s)$ on the right-hand semi plane $s$, satisfies the equality:

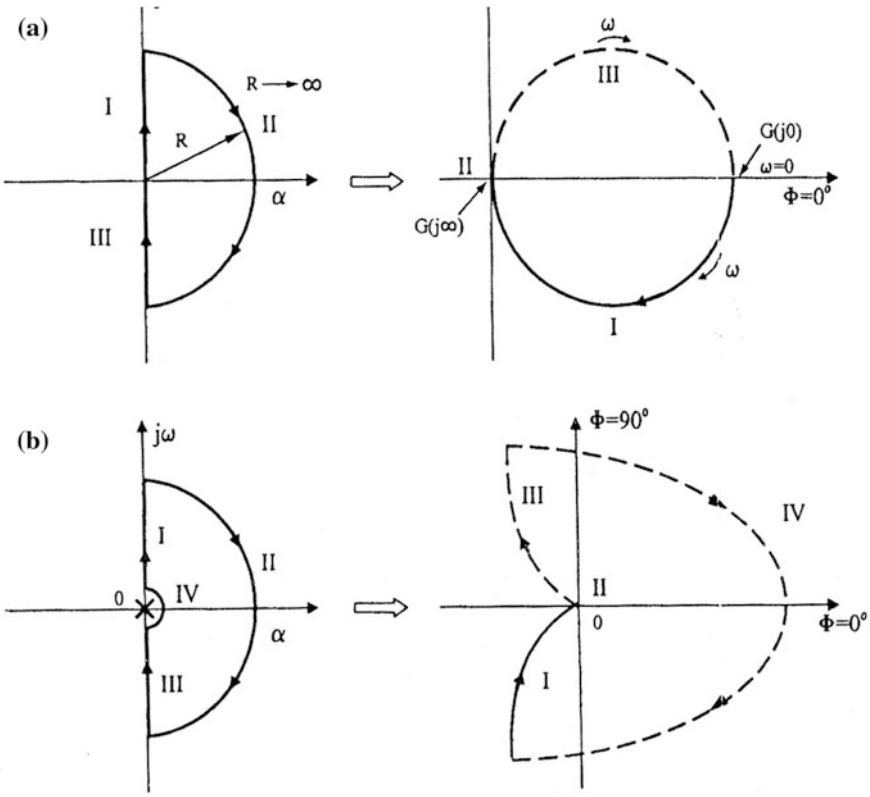



**Fig. 6.16** Examples of complete Nyquist plots **a** $G(s) = 1/(s+1)$, **b** $G(s) = 1/s(s+1)$

$$Z_\mathrm{r} = N_\mathrm{p} + P_\mathrm{r}$$

- If $N_\mathrm{p} \leq 0$, the Nyquist point $(-1, 0)$ is not inside (encircled by) the Nyquist plot, i.e., the Nyquist point lies to the left of the direction of traversing the Nyquist plot.
- If $P_\mathrm{r} = 0$, the system is stable if and only if $N_\mathrm{p} = 0$, i.e. again if the Nyquist point $(-1, 0)$ is not contained in the interior of the Nyquist plot.

Two examples of Nyquist plots are shown in Fig. 6.17. The first corresponds to the stable closed-loop system with $G(s)F(s) = 1/s(s+1)$ and the second to the unstable system $G(s)F(s) = 1/s(s-1)$.

The gain margin

$$K_\mathrm{margin} = 1/|G(j\omega)F(j\omega)|_{\omega=\omega_\phi}$$

and the phase margin

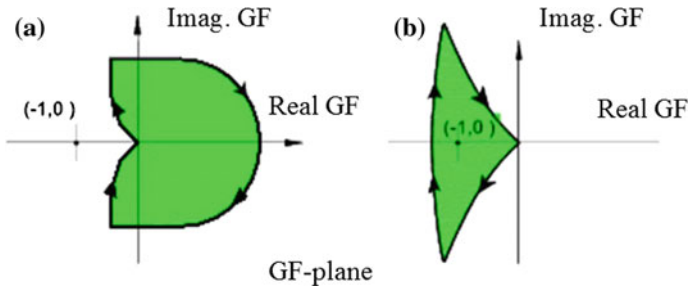



**Fig. 6.17** Nyquist plot of a stable system with (**a**) $G(s)F(s) = 1/s(s+1)$ and an unstable system (**b**) with $G(s)F(s) = 1/s(s-1)$

$$\phi_{\text{margin}} = 180^\circ + \underline{|G(j\omega)F(j\omega)|}_{\omega=\omega_a}$$

can be determined from the polar plot of $G(j\omega)F(j\omega)$ as shown in Fig. 6.18.

### 6.7.2 *Bode Method*

The Nyquist plot describes a control system (magnitude and phase) via the unique curve $G(j\omega)$ in the plane $G(j\omega)$ with parameter the frequency $\omega$. On the contrary, *Bode plots* describe the system via two distinct curves: the curve for the magnitude $|G(j\omega)|$ and the curve for the phase $\underline{|G(j\omega)|}$. Because the transfer function (can almost always) be expressed as products of simpler transfer functions, Bode has introduced the use of logarithms and the following definitions:

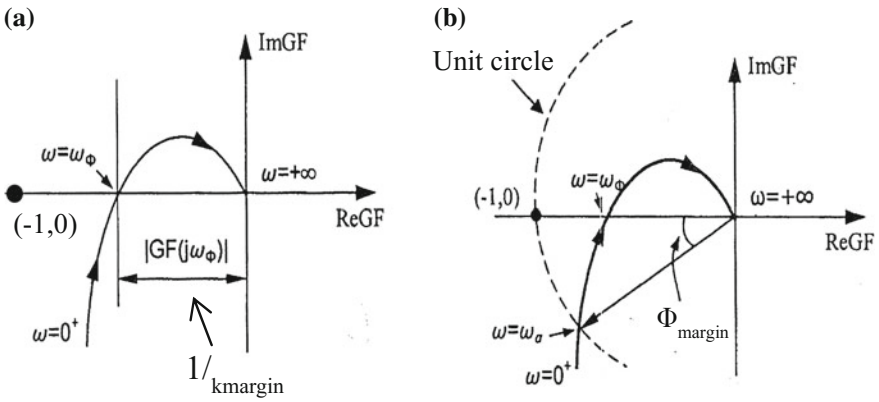



**Fig. 6.18** Definition of $|G(j\omega_\phi)F(j\omega_\phi)|$ and $\phi_{\text{margin}}$

- *Magnitude Bode plot* is the plot of $20\log_{10}|G(j\omega)|$db (decibels) in terms of $\log_{10}\omega$.
- *Phase Bode plot* is the plot of the phase $\angle G(j\omega)$ in terms of $\log_{10}\omega$.

In general, the following relation holds:

$$\begin{aligned}\log_{10}[G(j\omega)] &= \log_{10}|G(j\omega)|e^{j\phi(\omega)} = \log_{10}|G(j\omega)| + \log_{10} e^{j\phi(\omega)} \\ &= \log_{10}|G(j\omega)| + j0.434\phi(\omega)\end{aligned}$$

which implies that:

$$\begin{aligned}\mathrm{Re}[\log_{10} G(j\omega)] &= \log_{10}|G(j\omega)| \\ \mathrm{Im}[\log_{10} G(j\omega)] &= 0.434\phi(\omega)\end{aligned}$$

where $\mathrm{Re}[z]$ and $\mathrm{Im}[z]$ represent the *real part* and *imaginary part* of the complex number $z$.

Obviously, if:

$$G(j\omega) = F_1(j\omega)F_2(j\omega)/[H_1(j\omega)H_2(j\omega)]$$

then:

$$\begin{aligned}20\log_1|G(j\omega)| = {} & 20\log_{10}|F_1(j\omega)| + 20\log_{10}|F_2(j\omega)| \\ & - 20\log_{10}|H_1(j\omega)| - 20\log_{10}|H_2(j\omega)|\end{aligned}$$

This means that the Bode plot of the magnitude of $G(j\omega)$ can be found by the algebraic sum of the magnitude Bode plots of the factors $F_1(j\omega), F_2(j\omega)$, $H_1(j\omega)$ *and* $H_2(j\omega)$.

The same is true for the phase plot, i.e.:

$$\angle G(j\omega) = \angle F_1(j\omega) + \angle F_2(j\omega) - \angle H_1(j\omega) - \angle H_2(j\omega)$$

Therefore, in order to draw the Bode plots for the magnitude and phase of general linear system of the form:

$$G(j\omega) = \frac{K(1+j\omega\tau_1)(1+j\omega\tau_2)\cdots}{(j\omega)^p(1+j\omega_1 T_1)\left[1+(2\zeta/\omega_\mathrm{n})j\omega+\left(1/\omega_\mathrm{n}^2\right)(j\omega)^2\right]\cdots},$$

we plot separately and add the magnitude Bode plots of its factors in the numerator and denominator.

We say that a frequency $\omega_2$ is greater than the frequency $\omega_1$, by a *decade,* if $\omega_2/\omega_1 = 10$ or $\log_{10}(\omega_2/\omega_1) = 1$. Thus the number of frequency decades between $\omega_1$ and $\omega_2$ is given by $\log_{10}(\omega_2/\omega_1)$ decades.

Useful in the application of Bode's method are the so-called *asymptotic Bode plots*. For example, consider the following type-0 system:

$$G(j\omega) = K/(1 + j\omega\tau)$$

At low frequencies, $\omega \ll 1/\tau$, we have $20\log_{10}|G(j\omega)| = 20\log_{10} K$. The frequency $\omega_1 = 1/\tau$ is called the "*knee*" of the asymptotic Bode plot. At frequencies $\omega > \omega_1$, the slope of the asymptotic Bode plot is equal to −20 db/decade. Thus the asymptotic Bode plot of the above *0-type* system is as shown in Fig. 6.19a.

Working in the same way, for a *type-1* system of the form:

$$G(j\omega) = K/j\omega(1 + j\omega\tau),$$

we find the asymptotic diagrams shown in Fig. 6.19 in the two cases where (a) $\omega_1 > K$, (b) $\omega_1 < K$.

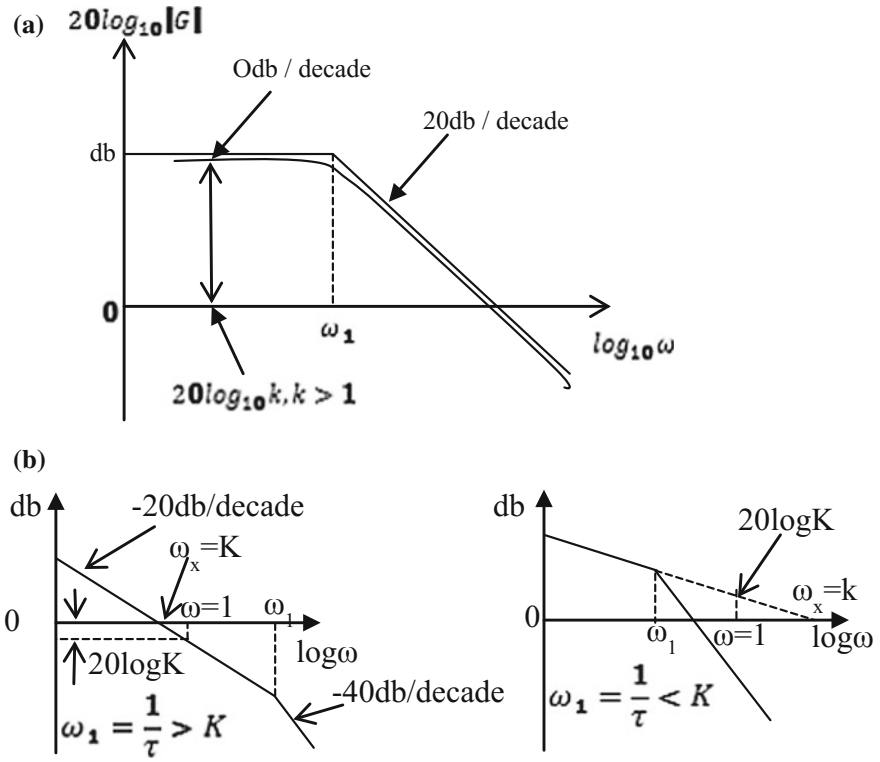



**Fig. 6.19** **a** Asymptotic Bode plot of the 0-type system $K/(1 + j\omega\tau)$. For comparison, the exact Bode plot (curved line) is also shown, **b** Asymptotic diagram of a type-1 system with knees $\omega_1 > 1$ and $\omega_1 < 1$, respectively

The above concepts regarding Bode's asymptotic diagrams can be extended to all systems of arbitrary type, with the remark that special care is needed when the system has pole or zeros on the right-hand semi plane $s$.

As we saw in Sect. 6.5.5, the second-order system possesses a central position in control-systems analysis and design because it provides simple formulas for $\zeta$ and $\omega_n$. In the frequency domain, the system specifications corresponding to $\zeta$ and $\omega_n$ are the *phase margin* and the *bandwidth* B of the system, which are approximately related to $\zeta$ and $\omega_n$ as:

$$\zeta = \phi^0_{\text{margin}}/100, \quad \text{for } \zeta < 1/\sqrt{2} = 0.707$$

$$B = \omega_n \sqrt{1 - 2\zeta^2 + \sqrt{2 - 4\zeta^2 + 4\zeta^4}} = \omega_n, \quad \text{for } \zeta = 1/\sqrt{2} = 0.707$$

Equivalent frequency domain parameters are the *peak frequency* $\omega_p$ and *peak amplitude value* $M_p$ (or maximum amplitude value) of the conventional frequency plot of the system's magnitude:

$$\omega_p = \omega_n \sqrt{1 - 2\zeta^2} \quad \left(0 < \zeta < 1/\sqrt{2}\right)$$

$$M_p = 1/2\zeta\sqrt{1 - \zeta^2} \quad \left(0 < \zeta < 1/\sqrt{2}\right)$$

The *bandwidth B* is related to the *rise time* $t_{\text{rise}}$ as:

$$t_{\text{rise}} = \pi/B$$

All the above relations can be properly used in the design of compensators and controllers as described in the following Sect. 6.9.

**Nyquist Criterion on the Bode plots** The Nyquist criterion can also be staled with the aid of the gain margin $K_{\text{margin}}$ and the phase margin $\phi_{\text{margin}}$ illustrated in Fig. 6.18. The frequency $\omega_\phi$ at which the *phase condition* holds (i.e., $\underline{G(j\omega_\phi)F(j\omega_\phi)} = 180^\circ$ ) is called "*phase-crossover frequency*" or "*frequency at which the gain margin is determined*", and the frequency $\omega_a$ at which the *magnitude condition* holds (i.e., $|G(j\omega_a)F(j\omega_a)| = 1$) is called "*gain crossover*" or "*frequency at which the phase margin is determined*". Clearly, referring to Fig. 6.20a it is seen that $K_{\text{margin}}$ is equal to:

$$K_{\text{margin}} = 1/\left|G(j\omega_\phi)F(j\omega_\phi)\right|$$

or:

$$20\log_{10} K_{\text{margin}} = -20\log_{10}\left|G(j\omega_\phi)F(j\omega_\phi)\right|$$

From Fig. 6.21b we find that:

$$\phi_{\text{margin}} = 180^\circ + \underline{|G(j\omega_a)F(j\omega_a)}$$

The relation for $20\log_{10} K_{\text{margin}}$ indicates that the *gain margin (positive)* in db is equal to the amount (in db) by which the quantity $20\log_{10} K_{\text{margin}}$ is below the 0db line (see Fig. 6.24a). The relation for $\phi_{\text{margin}}$ indicates that the *phase margin (positive)* is equal to the "*angle*" by which the angle $\underline{|G(j\omega_a)F(j\omega_a)}$ is above (greater than) 180° at the frequency where $|G(j\omega_a)F(j\omega_a)| = 1$ or, equivalently, $20\log_{10}|G(j\omega_a)F(j\omega_a)| = 0$ db (see Fig. 6.20a). Figure 6.20b shows the case where $K_{\text{margin}} < 0$ db and $\phi_{\text{margin}} < 0$.

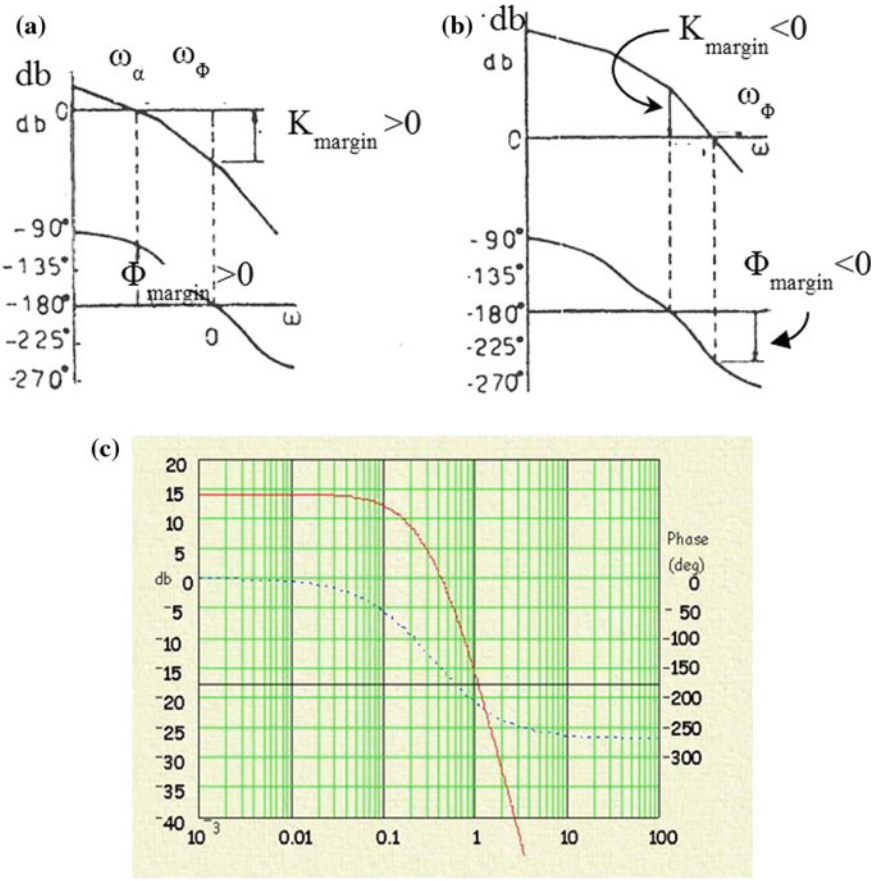



**Fig. 6.20** Gain and phase margins illustrated via Bode plots. **a** $K_{\text{margin}} > 0$, and $\phi_{\text{margin}} > 0$, **b** $K_{\text{margin}} < 0$, and $\phi_{\text{margin}} < 0$, **c** Detailed gain and phase Bode forms $\left(\phi_{\text{margin}} \simeq +30^\circ\right)$

From Fig. 6.20, we see that the **Nyquist stability criterion** can be stated via the **Bode plots** as:

"*A closed-loop system is stable if its open-loop amplitude Bode plot is below the 0 db line at* $\omega = \omega_\phi$*, or its open-loop phase plot is above the* $-180^\circ$ *line at* $\omega = \omega_a$*. In the opposite case, the system is unstable*".

*Remark*: It should be noted that the above formulation is valid when the system is a *minimum phase* system, i.e., when its Nyquist plot crosses the real axis and the unit circle only once. The response of a system is usually acceptable if $45^\circ < \phi^\circ_{\text{margin}} < 60^\circ$ (or, equivalently, if $0.45 < \zeta < 0.60$).

### 6.7.3 Nichols Method

The plot of the magnitude $|G(j\omega)F(j\omega)|$ (in db) with respect to the phase $\underline{|G(j\omega)F(j\omega)}$ (in degrees), in orthogonal axes, is called the system's *Nichols plot*.

The Nichols plot of the system:

$$G(j\omega)F(j\omega) = 4/[(j\omega)(1 + 0.125j\omega)(1 + 0.5j\omega)]$$

has the form shown in Fig. 6.21a, where the gain margin $K_{\text{margin}}$ and phase margin $\phi_{\text{margin}}$ are shown. We see that $K_{\text{margin}} > 0\,\text{db}$ and $\phi_{\text{margin}} > 0^\circ \phi_{\text{margin}} > 0^\circ$

Therefore, the system is *stable*.

A change of the system gain implies a translation of the plot *upwards* (if the gain is *increased*) or *downwards* (if the gain is *decreased*) without any change of the phase. Clearly, if the Nichols plot is moved *upwards* the gain margin is *decreased,* whereas if it is moved *downwards* the gain margin is *increased*.

In general, in terms of the Nichols (or L) plot, the Nyquist criterion is stated as:

> A closed-loop minimum phase system with open-loop transfer function $G(s)F(s)$ is **stable** if its Nichols plot $\left(20\log_{10}|G(j\omega)F(j\omega)|, \underline{|G(j\omega)F(j\omega)}\right)$ traversed in the direction of increasing $\omega$, leaves the Nichols Nyquist point $(0\,db,\ -180^\circ)$ on **its right**.

Figure 6.21b shows a particular stable system in which the Nichols plot is superimposed on the Nichols chart.

## 6.8 Discrete-Time Systems

### 6.8.1 General Issues

The concepts outlined in Sects 6.5–6.7 concern the case in which the control system is described by continuous-time models (i.e., by differential equations over

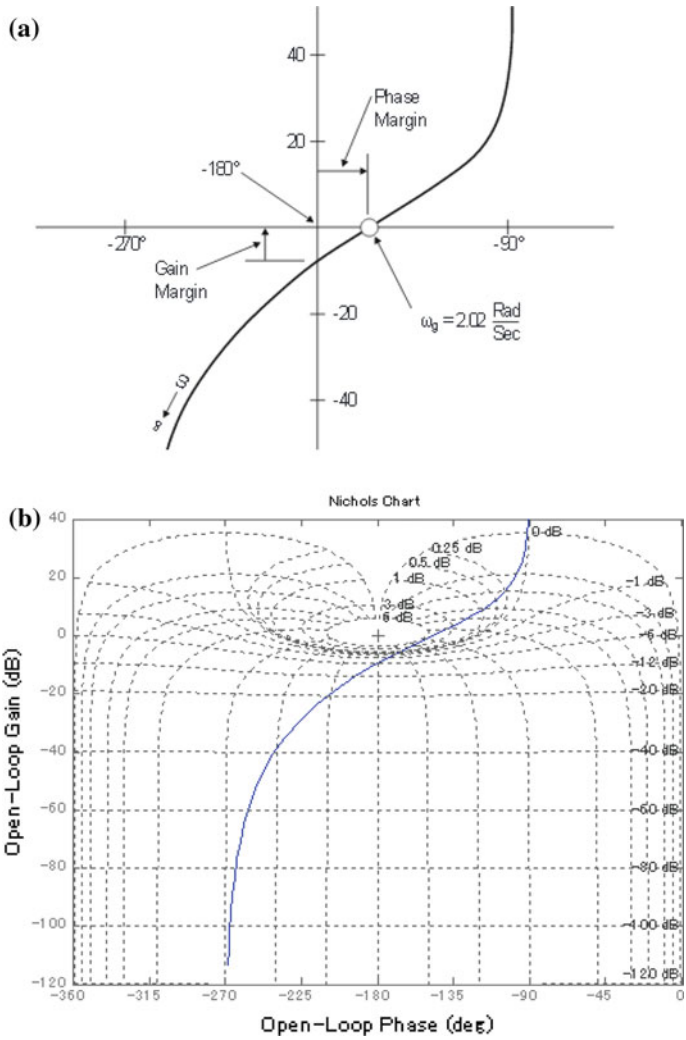



**Fig. 6.21** **a** Nichols plot of a stable system, **b** Nichols plot of a stable system, which is tangential to the 6 dB constant gain curve of the Nichols chart. Thus, since the Nichols Nyquist point is on the right of the plot, the system's gain margin is 6 dB

time. A whole body of similar concepts and results have also been developed for discrete-time systems.

*Discrete-time* (or *sampled-data*) systems are described by *difference equations* in the time domain and *pulse-transfer functions* in the frequency domain in terms of the complex variable:

$$z = e^{Ts}, \quad s = a + j\omega,$$

and the $\mathsf{Z}$ transform $F(z)$ of a given signal $f(t)$ which is defined as:

$$F(z) = \mathsf{Z}f(t) = \sum_{k=0}^{\infty} f(k)z^{-k}$$

where $f(k) = 0, 1, 2, \ldots$ are the sampled values of $f(t)$ at the sampling times $t_k = kT$, with $T$ being the *sampling period* (assumed here constant). The $\mathsf{Z}$ transform plays the same role as the Laplace transform. The input-output relation of a discrete-time system is

given by:

$$A(z)Y(z) = B(z)X(z)$$

where $X(z) = \mathsf{Z}x(k)$, $Y(z) = \mathsf{Z}y(k)$, and $A(z) = \sum_{k=0}^{n} a_k z^k$ $(a_n = 1)$, $B(z) = \sum_{k=1}^{n} b_k z^k$

From the above equation we obtain the *pulse* (or *discrete-time*) transfer function of the system:

$$G(z) = \frac{Y(z)}{X(z)} = \frac{B(z)}{A(z)}$$

One can observe that the pulse transfer function in the $z$ domain can be directly obtained from the difference equation (in the time domain) by simple replacement of the forward time operator $E$ with the complex variable $z = r + j\Omega$.

The block diagram of the system is analogous to the one given in Fig. 6.7 for continuous-time system and is shown in Fig. 6.22.

Again, $B(z)$ is the *input polynomial* and $A(z)$ the *output* or *characteristic* polynomial. The roots of $A(z) = 0$ are the poles of the discrete-time system (in the $z$ complex domain) and the roots of $B(z) = 0$ are the zeros of the system. The *sampled-data closed-loop transfer function* has the standard form (analogous to continuous-time systems):

$$\frac{Y(z)}{X(z)} = \frac{G(z)}{1 + G(z)F(z)}$$

only if there is a sampler at the *error path* (which samples both the input and feedback signals), and a sampler at the *output* of $\bar{g}(s)$.

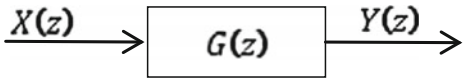



**Fig. 6.22** The block diagram of a linear SISO sampled-data system

A discrete-time transfer function $G(z)$ can be represented by its zeros and poles on the z plane. A system $\bar{g}(s)$ that has poles on the left-hand semi plane s (and so it is stable), in its discrete-time form $G(z)$ has poles in the interior of the unit circle (Fig. 6.23).

The map on the z-plane of a constant damping ratio $\zeta$ straight line of the s-plane is a logarithmic spiral as shown in Fig. 6.24

A discrete-time system with characteristic polynomial $A(z) = (z - 0.5)(z + 0.5)$ is stable since $|z_1| < 1$ and $|z_2| < 1$. A discrete-time system with poles $z_1 = +j$ and $z_2 = -j$ has the characteristic polynomial $(z - j)(z + j) = z^2 + 1$. The system is *marginally stable.*

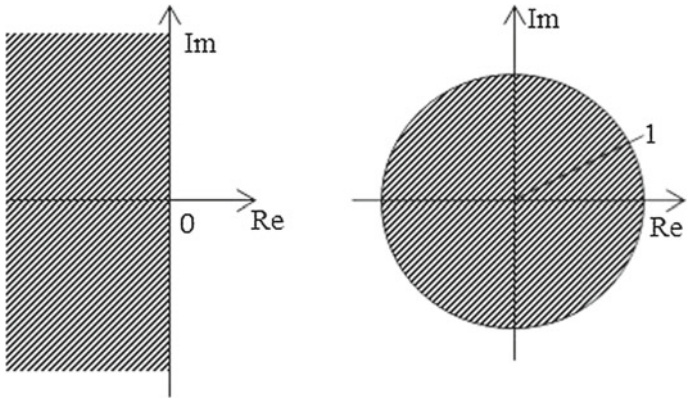



**Fig. 6.23** The interior of the unit circle is the stable region. The exterior is the unstable region

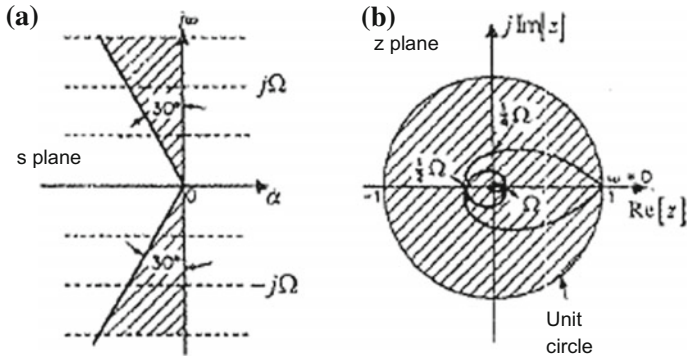



**Fig. 6.24** **a** Constant damping line $\zeta = |a|/\omega$ $(a = -x < 0)$ on the plane $s$ (straight line), **b** Constant damping line on the plane $z$ (spiral line)

The stability criteria of *Routh* and *Hurwitz* cannot be applied directly to discrete-time systems, but indirectly in a *new plane w*, which is related to the plane $z$ by the conformal mapping:

$$z = (w+1)/(w-1)$$

We can verify that, on the left semi plane $w$, we have $|w+1| < |w-1|$, i.e., $|z| < 1$ Therefore the interior of the unit circle on the z-plane is mapped to the left w-semi plane, and the exterior of this circle is mapped to the right w-semi plane. Thus, in order to check the stability of a discrete-time system with characteristic polynomial $A(z) = 0$, we use the polynomial:

$$A'(w) = A((w+1)/(w-1)) = 0$$

and apply the *Routh* or *Hurwitz* criterion to it.

Of course, there are also algebraic stability criteria that are directly applied in the $z$ domain. $|a_0| < a_2$.

### 6.8.2 Root Locus of Discrete-Time Systems

The *root-locus method* can be applied in a direct way to discrete-time systems with the basic difference that the role of the imaginary axis $j\omega$ of the $s$-plane, in the $z$-plane is played by the circumference of the unit circle. The construction of the root locus is made using the same rules as in the continuous-time case. Consider a closed-loop discrete time in the canonical form as shown in Fig. 6.25.

The characteristic equation of this system is $1 + G(z)F(z) = 0$, which is again represented by the *magnitude* and *phase conditions*:

$$|G(z)F(z)| = 1, \underline{|G(z)F(z)} = 180^\circ$$

For example, if:

$$G(z)F(z) = \frac{K(1-\mathrm{e}^{-T})z}{(z-1)(z-\mathrm{e}^{-T})},$$

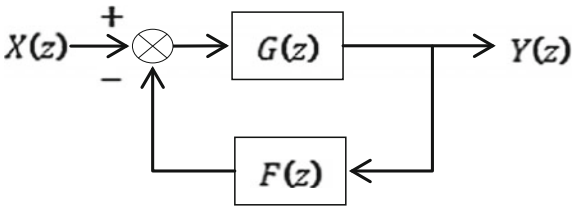



**Fig. 6.25** Block diagram of a sampled-data closed-loop system with forward- and feedback-transfer functions $G(z)$ and $F(z)$, respectively

the system characteristic equation is:

$$(z-1)\left(z-\mathrm{e}^{-T}\right)+K\left(1-\mathrm{e}^{-T}\right)z=0$$

and the root locus for $K$ from $K=0$ to $K=+\infty$ has the form shown in Fig. 6.26.

The system passes from stability to instability for the value of $K=K_{\mathrm{c}}$ (critical gain) at the point $z=-1$, where the root locus goes outside the unit circle.

### 6.8.3 *Nyquist Criterion for Discrete-Time Systems*

The *Nyquist stability criterion* is directly applicable to discrete-time systems. To this end, we draw the *Nyquist path* on the z-plane, leaving outside it all the system poles that lie on the unit circumference (via small semicircles) as shown in Fig. 6.27.

Then, we draw the *Nyquist plot* of the system $G(z)F(z)$ which corresponds to a single traversal of the unit circumference, (usually for $\omega=[-\Omega/2,\Omega/2]$ starting from the lower part that corresponds to $-\Omega/2\leq\omega\leq 0$ (or to $-180^{\circ}\leq\omega T\leq 0^{\circ}$).

The Nyquist criterion then states that:

> The closed-loop system that has the open-loop transfer function $G(z)F(z)$ is stable if and only if
>
> $$N_{\mathrm{p}}=-P_{\mathrm{r}}\leq 0$$
>
> where $P_{\mathrm{r}}$ is the number of poles of $G(z)F(z)$ outside the unit circle, and $N_{\mathrm{p}}$ is the number of clock-wise (positive) encirclements of the Nyquist plot around the Nyquist point (−1, 0) of $G(z)F(z)$ where we have $G(z)F(z)=-1$

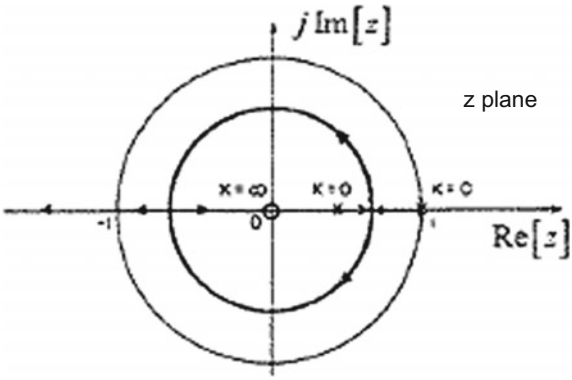



**Fig. 6.26** Root locus of $(z-1)(z-\mathrm{e}^{-T})+K(1-\mathrm{e}^{-T})z=0$ for $0\leq K<+\infty$

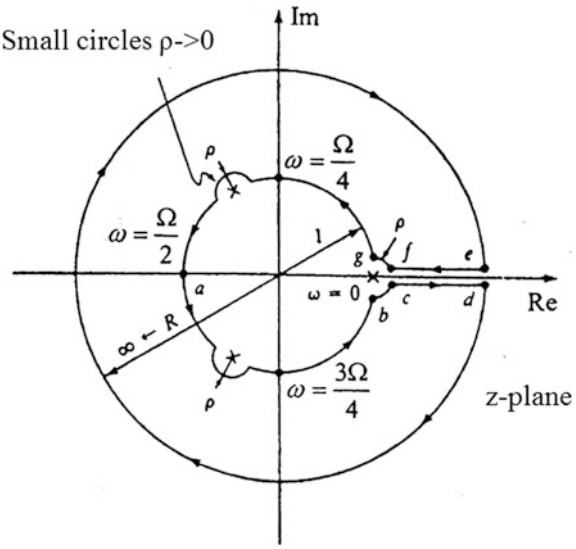



**Fig. 6.27** Nyquist path of a discrete-time system on the z-plane. The small circles around the poles that lie on the unit circle are drawn to avoid them

It is noted that the remarks given for the continuous-time systems are also valid here. For example, the number $Z_r$ of the roots of $\chi(z) = 1 + G(z)F(z)$ that lie outside the unit circle is equal to $Z_r = N_p + P_r$.

### 6.8.4 Discrete-Time Nyquist Criterion with the Bode and Nichols Plots

As it was done with the application of Routh and Hurwitz criteria to discrete-time systems, in order to use the Nyquist criterion, we introduce a new plane $w = a_w + j\omega_w$ which is related to the plane $z = e^{sT}$ through the mapping:

$$z = \frac{1+w}{1-w} \quad \text{or} \quad w = \frac{1-z^{-1}}{1+z^{-1}}$$

Thus, setting $z = e^{j\omega T}$ in the expression for $w$, we get:

$$w = \frac{1-e^{-j\omega T}}{1+e^{-j\omega T}} = j\tan\left(\frac{\omega T}{2}\right) \quad \text{i.e.,} \quad \omega_w = \tan\left(\frac{\omega T}{2}\right)$$

where $\omega_w$ is the frequency on the plane $w$, and has period $\Omega$. Inverting the relation $\omega_w = f(\omega)$ we get:

$$\omega = (2/T)\tan^{-1}\omega_{\mathrm{w}}$$

Thus, to draw the Bode (or Nichols) plot of a discrete-time system $G(z)$, we obtain the new (transformed) function:

$$\hat{G}(w) = G((1+w)/(1-w))$$

and work in the usual way as for the continuous-time case.

## 6.9 Design of Classical Compensators

### 6.9.1 General Issues

The design of a *compensator* (or controller) is the procedure by which we determine (select) a compensator/controller which, if applied to a given system, leads to an overall closed-loop system that possesses desired specifications of transient and steady-state performance. If the compensator is placed in the forward path, then it is called a *series compensator,* and, if it is placed in the feedback path, it is called a *feedback compensator*, as shown in Fig. 6.28a, b. In many cases, we design simultaneously a series compensator and a feedback compensator.

The compensator design can be performed using any one of the methods described thus far, i.e.:

- Design via root locus
- Design via Nyquist plots
- Design via Bode plots
- Design via Nichols plots

A special design method for series compensation is *Ziegler-Nichols* PID (Proportional plus Integral plus Derivative) compensator. Our purpose here is to provide a tour of all these design methods.

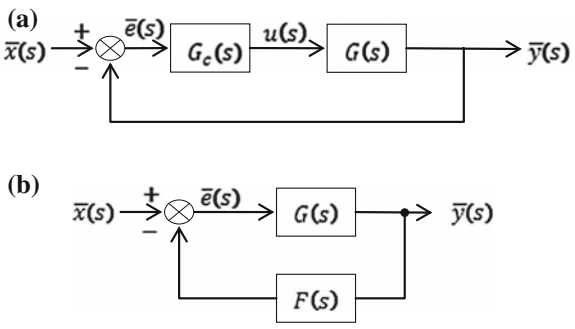



**Fig. 6.28** **a** A series compensator $G_{\mathrm{c}}(s)$, **b** A feedback compensator $F(s)$

### 6.9.2 Design via Root Locus

The root locus can be used for selecting the value of the gain $K$ such that the closed-loop poles are moved to positions that correspond to the desired specifications $\zeta$, $\omega_n$, $h$, $\tau_{dom}$, etc. In many cases, this is not possible by mere *gain-control.* In these cases, the root locus must be properly modified. The *gain* is selected to achieve the desired steady-state performance (errors). Here we distinguish the following cases:

- The transient response is acceptable, but the steady-error is large. In this case we can increase the gain without any change of the root locus (pure proportional control).
- The system is stable but the transient response is not acceptable. In this case, the root locus must be shifted to the left, i.e., at a greater distance from the imaginary axis.
- The system is stable, but both the steady-state performance (steady-state errors) and the transient performance are not acceptable. Here, both a gain increase and a shift of the root locus to the left are needed.
- The system is unstable for all the values of the gain. Thus the root locus should be modified such that some part of every branch lies on the left semi plane $s$. In this way, the system will become conditionally stable.

Actually, the use of a compensator introduces new poles and new zeros in the open-loop transfer function. Typically, the input used in all cases is the unit step function. If the desired response is *subcritical,* the compensator is usually selected such that to obtain a pair of dominant complex poles, so that the contribution of all the other poles is negligible. In general, one can apply the following compensation techniques.

**Phase-lag compensation** This type of compensation is done when the transient response is good, but the steady-state error is very high. We actually increase the type of the system with minimal change of the closed-loop poles. The ideal phase-lag compensator has a transfer function:

$$G_c(s) = (s - \mu)/s, \quad \mu < 0$$

i.e., a pole at $s = 0$ and a zero in the left semi plane $s$, but very near to $s = 0$ to secure that the actual positions of the closed-loop poles remain the same, and so the transient response remains practically unchanged. The above compensator is actually a *proportional-plus-integral* controller. This is verified by writing $G_c(s)$ as:

$$G_c(s) = 1 + K_i/s, \quad K_i = -\mu > 0$$

where $K_i$ is the gain of the *integral term.* In practice, the implementation of the pure integral term is difficult, and so a passive RC circuit is used.

**Phase-lead compensation** This type of compensation is done when the steady-state error is acceptable, but the transient response is not acceptable. To improve the transient response, the root locus should be moved to the left of its original position. The ideal phase lead compensator has transfer function ($\mu < 0$):

$$G_c(s) = s - \mu = 1 + K_d s, \quad K_d = -\mu > 0$$

This compensator adds a zero to the forward path, which in practice can be obtained by using a differentiator. This controller is actually a *proportional-plus-derivative controller*.

There are many methods for selecting the zero and pole of the compensator. These include the Dorf method, the bisection method, and the constant-phase circles method [46–48].

**Phase lead-lag compensation** This type of compensation is needed when both the steady-state and the transient performances are not satisfactory. We can separately design a lag compensator to achieve the steady-state specifications and a lead compensator for the transient specifications and then combine them in series with a buffer circuit between them (to avoid the loading of the first circuit by the second circuit).

### 6.9.3 Design via Frequency-Domain Methods

The compensator design in the frequency domain is performed using the parameters $M_p$, $\omega_p$, and $K$. Therefore, if the performance specifications are given by other parameters they should be converted to $M_p$, $\omega_p$, and $K$. The selection of the gain via the Nyquist plot is made using the constant $M$ circle, which has the desired value $M = M_p$ of $M$, as shown in Fig. 6.29b.

The Nyquist polar plot should be tangential to the $M_p$ circle. To this end, we first draw the straight line from the origin at an angle $\theta$ with $\sin\theta = 1/M_p$, and then we

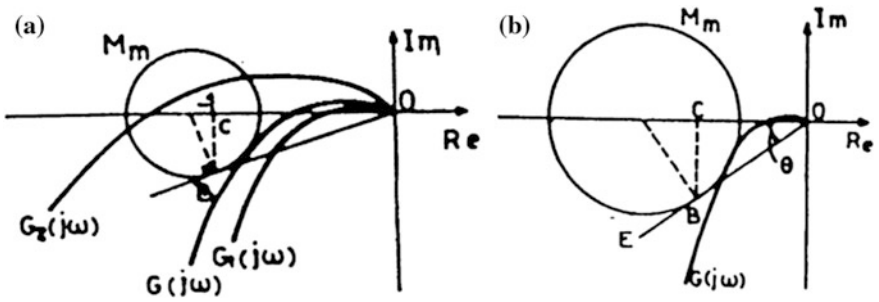



**Fig. 6.29** **a** Polar plot of $G(j\omega)F(j\omega)$ for several values of $K(K_1 < K < K_2)$, **b** Determination of $K_0$ that leads to $M_p$

draw a circle with center at the negative real axis, which is tangential to the $M_p$ circle. The required gain $K_0$ is given by the relationship:

$$K_0/K_1 = 1/(OC),$$

where $(OC)$ is the length of the line segment shown in Fig. 6.29b, and $K_1$ is the gain contained in $G(j\omega)F(j\omega)$. If $M_p$ and $\omega_p$ are satisfactory, but the steady-state error is very high, we have to increase the gain without changing the part of the frequency-domain plots in the vicinity of the Nyquist point $-1+j0$ or $-180^\circ/0$ db. This can be achieved by using a proper phase-lag compensator $G_c(s)$. Compensator design techniques using polar, Bode, and Nichols plots can be used as described in [41–44].

### 6.9.4 Discrete-Time Compensator Design via Root-Locus

The root-locus method is used for designing discrete-time compensators in exactly the same way as for continuous-time compensators, with the difference that we must also take into account the sampling period $T$.

In the z-plane, the following issues must be used in the design process:

- The constant damping ratio $\zeta$ line is a spiral (see Fig. 6.24). Poles with $\zeta \geq \zeta_0$ lie inside the spiral that corresponds to $\zeta_0$.
- The poles with constant real part $\mathrm{Re}[z] = \mathrm{e}^{-aT}$, $a > 0$, lie on the circumference centered at the origin and having a radius $\mathrm{e}^{-aT}$.
- A pole that lies at angle $\theta$ with respect to the real line corresponds to a number of sampling periods per cycle equal to $N = 2\pi/\theta$.
- The sampling period $T$ must satisfy Shannon's sampling theorem's condition.

The transfer function of a phase-lead or phase-lag discrete-time compensator has the form:

$$G_c(z) = K_c(z-\mu_c)/(z-\pi_c)$$

where $\mu_c$ is a real zero and $\pi_c$ is a real pole. If $\lim_{z\to 1} G_c(z) = 1$, then the compensator does not change the steady-state performance (errors).

## 6.10 Ziegler-Nichols Method for PID Controller Tuning

The **PID** (**P**roportional plus **I**ntegral plus **D**erivative) or *three-term controller* is the most popular controller in process control and involves three additive components as shown in Fig. 6.30.

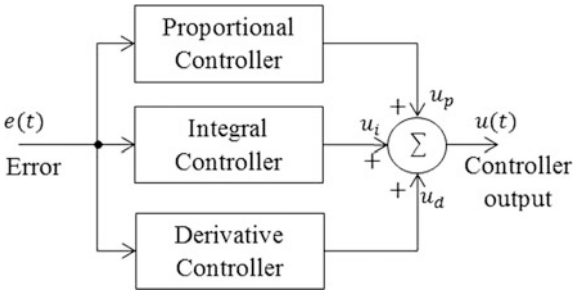



**Fig. 6.30** The three components ($u_p$, $u_i$, and $u_d$) of a PID controller

The output $u_p$ of the proportional term is given by:

$$u_p = K_a e(t)$$

where $K_a$ is the so-called "proportional gain", and the outputs of the integral and derivative terms are, respectively:

$$u_i = K_i \int_0^t e(t')dt'$$

$$u_d = K_d \frac{de(t)}{dt}$$

where $K_i$ and $K_d$ are called *integral-controller gain* and *derivative-controller gain,* respectively.

The overall output $u(t)$ of the PID controller, i.e., the overall control signal provided by the PID controller is given by:

$$u(t) = u_p + u_i + u_d = K_a e(t) + K_i \int_0^t e(t')dt' + K_d \frac{de(t)}{dt}$$

and is usually written as:

$$u(t) = K_a \left\{ e(t) + \tau_d \frac{de(t)}{dt} + \frac{1}{\tau_i} \int_0^t e(t')dt' \right\}$$

where:

$$\tau_d = K_d/K_a, \quad \tau_i = K_a/K_i$$

are called the "*time constant*" of the derivative term and the integral term, respectively. In this case, $K_a$ is called the *gain of the PID controller*.

The transfer function of the PID controller is found from the above time-domain integrodifferential equation, as:

$$G_c(s) = \frac{\bar{u}(s)}{\bar{e}(s)} = K_a\left(1 + s\tau_d + \frac{1}{s\tau_i}\right)$$

The time constant of the integral term is usually called "*reset time*". If the error $e(t)$ has the form of a unit ramp function (see Fig. 6.31a), then the control signal $u(t)$ (i.e., the output of the PID controller) has the form pictured in Fig. 6.31b.

Since here the controller has a fixed structure, our design task is reduced to that of selecting the three parameters $K_a$, $\tau_d$, and $\tau_i$. The procedure of selecting these parameters is known in the literature as *PID controller tuning*.

The most popular PID parameter tuning method is the *Ziegler-Nichols* method (1942). Among the various existing variants of this method, we describe here the one which is based on the *stability limits of the closed-loop system.* This includes the following steps, which are performed by a human operator.

Step 1: Disconnect the derivative and integral terms (i.e., use only the proportional term).

Step 2: Increase the gain $K_a$ until the stability limit is reached and oscillations are started. Let $T_0$ be the oscillations' period and $K_c$ the critical value of the gain.

Step 3: Select the parameters $K_a$, $\tau_i$, and $\tau_d$ as follows:

- For proportional control: $K_a = K_c/2$
- For PI control: $K_a = 0.45K_c,\ \tau_i = 0.8T_0$
- For PID control: $K_a = 0.6K_c,\ \tau_i = T_0/2,\ \tau_d = \tau_i/4$

The performance achieved by these values in typical systems is acceptable (giving about 10–20% overshoot).

The discrete-time form of the PID controller can be found using the so-called orthogonal integration:

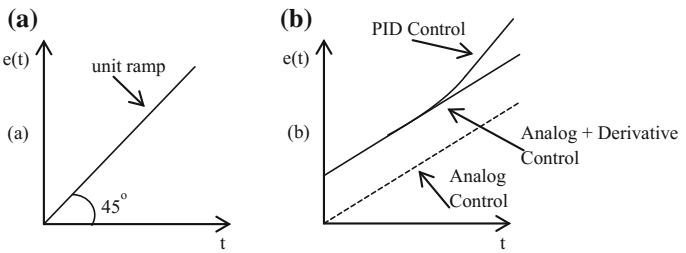



**Fig. 6.31** Unit ramp response of PID controller: **a** Unit ramp error, **b** corresponding PID control signal

$$s \to \frac{1}{T}(z-1)$$

or the *trapezoidal integration (Tustin's approximation), which is more accurate:*

$$s \to \frac{2}{T}\left(\frac{z-1}{z+1}\right)$$

The orthogonal integration leads to the discrete-time PID controller form:

$$u(k) = K_{\mathrm{a}}\left\{e(k) + \frac{\tau_{\mathrm{d}}}{T}[e(k) - e(k-1)] + \frac{T}{\tau_{\mathrm{i}}}\sum_{i=0}^{k-1} e(i)\right\}$$

This controller is of *non-recursive* type. The *recursive equivalent* is obtained by applying it at time $k-1$ and subtracting to find:

$$\Delta u(k) = u(k) - u(k-1) = \beta_0 e(k) + \beta_1 e(k-1) + \beta_2 e(k-2)$$

where:

$$\beta_0 = K_0\left(1 + \frac{\tau_{\mathrm{d}}}{T}\right), \beta_1 = -K_{\mathrm{a}}\left(1 + \frac{2\tau_d}{T} - \frac{T}{\tau_i}\right), \beta_2 = K_{\mathrm{a}}\frac{\tau_{\mathrm{d}}}{T}$$

The present value $u(k)$ of the control signal is now computed as:

$$u(k) = u(k-1) + \Delta u(k)$$

(i.e., adding the correction $\Delta u(k)$, to the previous value $u(k-1)$). The z-transfer function of the above recursive PID controller is found to be:

$$G_{\mathrm{c}}(z) = \frac{U(z)}{E(z)} = \frac{\beta_0 + \beta_1 z^{-1} + \beta_2 z^{-2}}{1 - z^{-1}},$$

which can be written in the typical parallel three-term form:

$$G_{\mathrm{c}}(z) = K_{\mathrm{a}}\left[1 + c_{\mathrm{d}}\left(1 - z^{-1}\right) + c_{\mathrm{i}}\frac{z^{-1}}{1 - z^{-1}}\right]$$

where $K_{\mathrm{a}}$ is the proportional term, $K_{\mathrm{a}}c_{\mathrm{d}}(1 - z^{-1})$ is the derivative term, and $K_{\mathrm{a}}c_{\mathrm{i}}z^{-1}/(1 - z^{-1})$ is the integral term. The tuning procedure is now the proper selection of $K_{\mathrm{a}}$, $c_{\mathrm{d}}$, and $c_{\mathrm{i}}$.

The above PID parameter tuning procedure (Ziegler–Nichols) is a practical experimental procedure that is applied in industrial systems by the control operators.

Clearly the PID controller is a phase lead-lag controller, where only three parameters, $K_a$, $\tau_d$, and $\tau_i$ have to be selected. Therefore, any other series compensator design procedure can also be applied.

## 6.11 Nonlinear Systems: Describing Functions and Phase-Plane Methods

The two alternative classical control methods developed for nonlinear control-systems analysis and design are the describing functions and phase-plane methods.

### *6.11.1 Describing Functions*

*Nonlinear systems* do not satisfy the *superposition principle* like the linear systems, i.e., the response $y(t)$ of them to the sum of two inputs $x_1(t)+x_2(t)$ is not equal to the sum of its responses $y_1(t)$ and $y_2(t)$ to the inputs $x_1(t)$ and $x_2(t)$, separately. Actually, most physical systems, for inputs of very large magnitude, are not linear, due to the existence of nonlinearities such as the *saturation* of actuators and measurement devices/sensors, *dead zone,* hysteresis, nonlinear friction, and other nonlinear characteristics. A simple example of nonlinear systems is the *inverse pendulum* vehicle, which is often used as a *benchmark* for testing nonlinear controllers.

The *describing function technique,* which is a frequency domain technique, can be illustrated by the nonlinear feedback control system shown in Fig. 6.32.

In this system, the *nonlinearity N* is due to the saturation of the amplifier that drives the linear motor $G_2(j\omega)$. If the input $e_i$ of the nonlinearity $N$ is *sinusoidal* (harmonic), its output $e_0$ will be *periodic* (but not harmonic), which can be expanded in a Fourier series as:

$$e_0(t) = \sum_{n=1}^{\infty} [a_n \cos(n\omega_i t) + b_n \sin(n\omega_i t)]$$

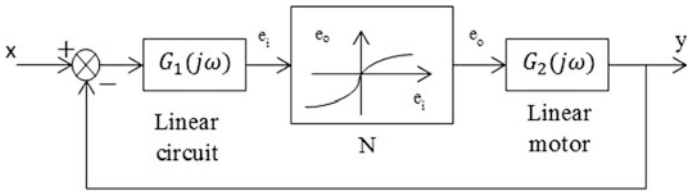



**Fig. 6.32** A nonlinear system involving the linear parts $G_1(j\omega)$ and $G_2(j\omega)$ and a nonlinearity $N$

where $\omega_i$ is the cyclic frequency of the *harmonic input* $e_i$. This series consists of the *fundamental* (basic) *component* $e_b = a_1 \cos(\omega_i t) + b_1 \cos(\omega_i t)$ and the sum of the other higher order components that have frequencies $n\omega_i (n = 2, 3, \ldots)$.

**Describing function** *(or nonlinear gain) is called the ratio of the fundamental component* $e_b$ *of the output over the input* $e_i$.

This means that, in calculating the describing function, all higher-order components are neglected, a fact that is usually justified by the following two observations:

- Usually, control systems involve low-pass subsystems (e.g., circuits), and so the higher-frequency components are typically subject to strong attenuation, over the basic frequency, and can be omitted.
- The amplitudes of the high-order components are initially very much smaller than the amplitudes of the fundamental component.

As an example, consider the system of Fig. 6.33, where the input $e_i \sin(\omega t)$ is applied to an ideal *switch* $(-V, +V)$.

The Fourier series of the output $e_0$ of the switch is:

$$e_0 = (4V/\pi)[\sin \omega t + (1/3) \sin 3\omega t + (1/5) \sin 5\omega t + \ldots]$$

and so the system output $y$ is given by:

$$y = -\frac{4V}{\pi(\omega\tau)^2}\left[\sin \omega t + \frac{1}{3} \cdot \frac{1}{9} \sin 3\omega t + \frac{1}{5} \cdot \frac{1}{115} \sin 5\omega t + \ldots\right]$$

We observe that:

- Amplitude of 3rd harmonic <4% of the amplitude of the fundamental harmonic.
- Amplitude of 5th harmonic <1% of the amplitude of the fundamental harmonic.

Therefore the third, fifth, and all higher harmonics can indeed be neglected without significant error. The describing function of the ideal switch is:

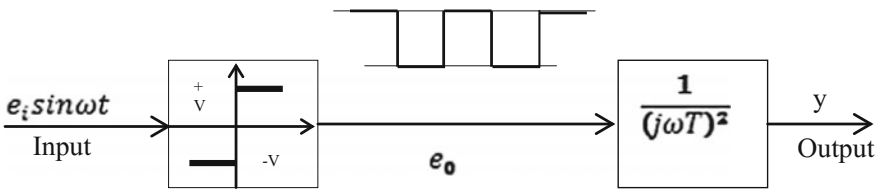



**Fig. 6.33** Ideal switch followed by a double integrator

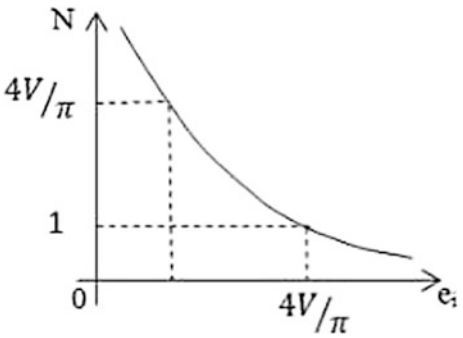



**Fig. 6.34** Describing function of the ideal switch

$$N = \frac{\text{Amplitude of fundamental harmonic}}{\text{Amplitude of input}} = \frac{4V}{\pi e_{\text{i}}}$$

The diagram of $N$ with respect to $e_{\text{i}}$ has the form shown in Fig. 6.34.

We observe that the describing function of the ideal switch is *independent of the frequency* and does not introduce any phase shift. Such a describing function is called a *simple describing function*. However, here $N$ is a function of the amplitude $e_{\text{i}}$ of the harmonic input.

A describing function that introduces a phase shift but is not dependent on the frequency is called *complex.* There are also cases where a complex-describing function depends on the frequency. This is usually due to energy storage in the nonlinearity.

### 6.11.2 Oscillations Condition

The *oscillations condition,* which is the most basic condition of feedback control systems, has the same general form in both linear and nonlinear systems. Thus, for the systems shown in Fig. 6.35, the *oscillations condition* has the following form:

*Linear system (a):* $G(j\omega) = -1/K$

*Nonlinear system (b):* $K_1K_2G_1(j\omega)G_2(j\omega) = -1/N$

The oscillations condition is obtained from the closed-loop characteristic equation for $s = j\omega$, where:

$$\frac{y(j\omega)}{x(j\omega)} = \frac{KG(j\omega)}{1 + KG(j\omega)} \quad \text{(Linear System)}$$

$$\frac{y(j\omega)}{x(j\omega)} = \frac{K_1K_2G_1(j\omega)G_2(j\omega)N}{1 + K_1K_2G_1(j\omega)G_2(j\omega)N} \quad \text{(Nonlinear System)}$$

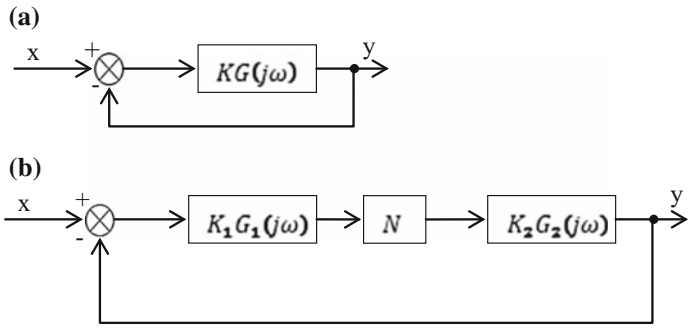



**Fig. 6.35** Unity feedback control systems **a** Linear system, **b** Nonlinear system

We observe that, in the nonlinear system, the oscillations condition consists of the linear part, which is independent of the amplitude and the nonlinear part $-1/N$, which is a function of the input amplitude. In general, the oscillation (stability) condition can be studied on a polar plot (as in linear systems) with the difference that there are two functions, namely, the linear and nonlinear functions. As a result, we have two differences between the Nyquist plots of linear and nonlinear systems:

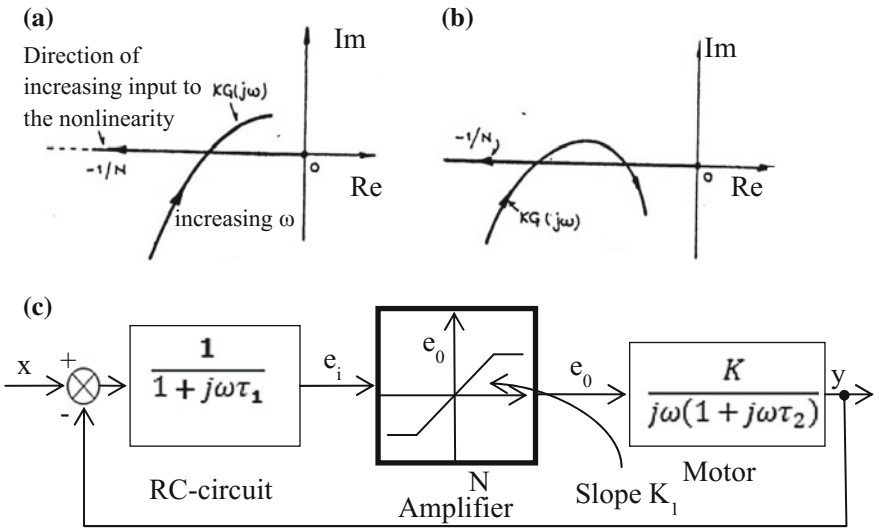



**Fig. 6.36** Two examples of nonlinear Nyquist plots: **a** the curves $KG(j\omega)$ and $-1/N$ intersect at a single point, **b** the curves $KG(j\omega)$ and $-1/N$ intersect at two points, **c** Closed-loop system with a 'saturation' element

- We have to examine an entire gain region which depends on the amplitude of the signal, whereas in linear systems only one gain value needs to be considered.
- There is a relationship between the amplitude and the frequency of the oscillations, a fact that does not occur in linear systems.

Two typical forms of nonlinear Nyquist plots are shown in Fig. 6.36a, b.

### 6.11.3 *Stability Investigation of Nonlinear Systems via Describing Functions and Nyquist Plots*

We will study the conditions for stability of nonlinear systems using the describing-function technique. To this end, the closed-loop system of Fig. 6.36c, which involves an element with a *saturation characteristic,* will be examined.

The condition for *sustained oscillations* is:

$$KG(j\omega) = \frac{K}{j\omega(1+j\omega\tau_1)(1+j\omega\tau_2)} = -\frac{1}{N}$$

The Nyquist plot for several values of $K$ is shown in Fig. 6.37.

For very small values of $K$ (the gain of the linear part $G(j\omega)$), the two curves $-1/N$ and $KG(j\omega)$ do not intersect, and so there is no chance to have sustained oscillations (Fig. 6.37a). For a certain value of $K$, the two loci just touch each other (Fig. 6.37b). This is a minimum value of $K$ for which we have sustained oscillations. For greater values of $K$ (Fig. 6.37c), the two loci intersect, and we can have sustained oscillations (i.e., oscillations of constant amplitude). The frequency of oscillations depends on K, because the gain $K$ affects the point at which the two loci $-1/N$ and $KG(j\omega)$ intersect.

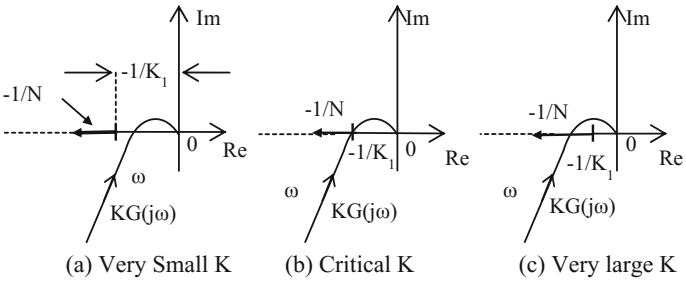



**Fig. 6.37** Nyquist plot and describing function. Three different values of $K$ are used: very small (**a**), critical (**b**), and large (**c**)

### 6.11.4 Application of Root Locus to Nonlinear Systems

The root locus technique can also be applied to nonlinear systems represented by describing functions. The basic difference from the linear system case is that here the gain varies with the amplitude of the signals, which leads to closed-loop poles that are shifted when the amplitude of the input-signal changes. The difficulty is that the standard exponential form of the responses that correspond to "moving (changing) poles" is no longer valid. As a consequence, the power that the root-locus technique has in linear systems is lost, and the root locus can only give information about the existence of limit cycles and about whether *sustained oscillations* are possible, as the describing function method does.

For systems with *real describing* functions (without phase shift), the poles are moving on the 180° lines (root locus) of their linear parts. As an example, consider the case shown in Fig. 6.38, with a *saturation nonlinearity N*.

Suppose that the initial value of the gain gives closed-loop poles lying on the left-hand semi-plane *s*. By increasing the amplitude of signals we have the start of saturation, and so the value of $K$ is reduced. As a result, the closed-loop poles are moving towards the imaginary axis. When the poles arrive at the imaginary axis (for the critical value $K_c$ of $K$), the amplitude of oscillations increases suddenly, and the operational point moves to the position $C$, where the system as at *limit cycle*. In the general case, the critical (oscillation) condition is:

$$KG(j\omega) = -1/N$$

and the procedure is as follows:

Step 1: We draw the lines of constant amplitude and constant phase of the linear part $KG(j\omega)$.
Step 2: We draw the describing function of the nonlinear part $-1/N$.
Step 3: We determine on the constant-amplitude and constant-phase lines the points that satisfy the relations:

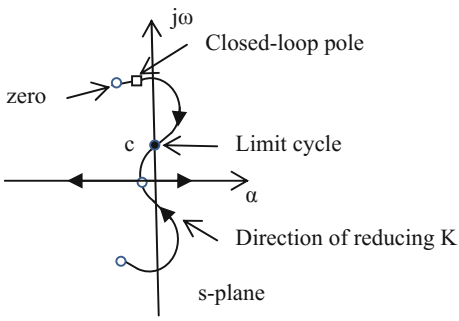



**Fig. 6.38** Study of a nonlinear system via the root locus. The system can oscillate in a 'limit cycle'

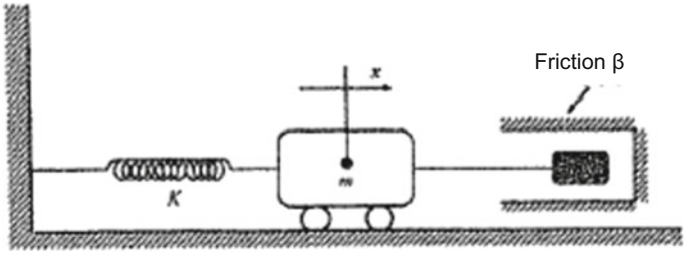



**Fig. 6.39** A linear mass-spring system

$$|KG(j\omega)| = |1/N|, \underline{|G(j\omega)} = 180^\circ - \underline{|N}$$

### 6.11.5 *Phase Plane*

The phase-plane technique is applicable to second-order systems. With this method, we can study completely the step-response of linear and nonlinear systems. To illustrate the method, we start by working on the mass-spring system of Fig. 6.39.

The differential equation of this second-order system is $m\mathrm{d}^2x/\mathrm{d}t^2 + \beta \mathrm{d}x/\mathrm{d}t + kx = 0$. In terms of the velocity, $\upsilon = \mathrm{d}x/\mathrm{d}t$, the above equation can be written as:

$$m\mathrm{d}\upsilon/\mathrm{d}t + \beta\upsilon + kx = 0$$

Writing:

$$\mathrm{d}\upsilon/\mathrm{d}t = (\mathrm{d}\upsilon/\mathrm{d}x)(\mathrm{d}x/\mathrm{d}t = \upsilon\mathrm{d}\upsilon/\mathrm{d}x),$$

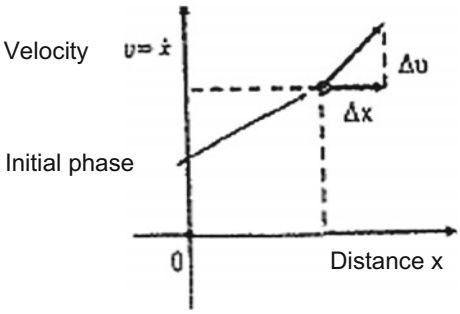



**Fig. 6.40** Phase plane $\{\Delta\upsilon/\Delta x = -(kx + \beta\upsilon)/m\upsilon\}$

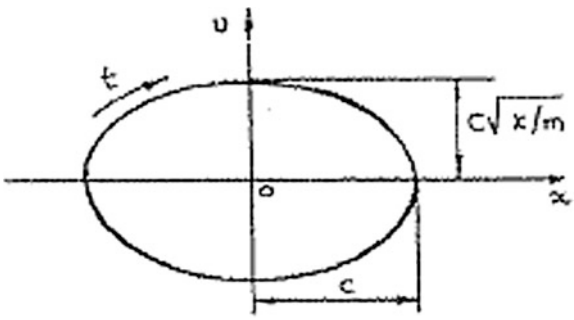



**Fig. 6.41** For $\beta = 0$, the trajectories are ellipses

we finally arrive at the equation:

$$\frac{\mathrm{d}\upsilon}{\mathrm{d}x} = -\left(\frac{kx + \beta\upsilon}{m\upsilon}\right)$$

This equation contains the variables $x$ (position) and $\upsilon$ (velocity), which are called "*phase-variables*" of the system, and can be studied on the plane $x - \upsilon$, which is called the *phase plane* (Fig. 6.40).

A given point of the $x - \upsilon$ plane represents a *state* or *phase* (position and velocity) of the system, and the rate of change of the velocity $\upsilon$ with respect to the position $x$ is described by the equation. The system passes from a continuous sequence of states that satisfy this equation. This sequence of states is called the *trajectory* of the system on the phase plane.

Time appears as a parameter along the trajectory. A set of many trajectories that correspond to different initial conditions is called a "*portrait*" of the system. The trajectory is found by integrating the equation for $\mathrm{d}\upsilon/\mathrm{d}x$. Thus, in the zero-friction case $(\beta = 0)$, we get $m\upsilon\mathrm{d}\upsilon = -kx\mathrm{d}x$, from which it follows that $\int (m/k)\upsilon\mathrm{d}\upsilon = -\int x\mathrm{d}x + C^2/2$, or $(m/k)\upsilon^2 + x^2 = C^2$.

This equation represents an *ellipse* as shown in Fig. 6.41.

From the velocity definition $\upsilon = \mathrm{d}x/\mathrm{d}t$, it follows that $\Delta x = \upsilon\Delta t$, and so, because in actual practice we have $\Delta t > 0$, for $\upsilon > 0$ we have $\Delta x > 0$, while for $\upsilon < 0$ it follows that $\Delta x < 0$. Thus the allowed direction for traversing the trajectory shown and below the x-axis is as shown in Fig. 6.42 (up). The trajectory of a system of the form shown in Fig. 6.39, when $\beta > 0$, has the general shape shown in Fig. 6.42 (down, a). The corresponding time response is as shown in Fig. 6.42 (down, b).

**Definition** *Limit cycle is any closed trajectory of the phase plane.*

Limit cycles are distinguished in *stable* and *unstable* limit cycles. In the case of a stable limit cycle, all the phase-plane trajectories converge finally to this limit cycle (Fig. 6.43a). On the contrary, in the case of an *unstable-limit cycle,* all the trajectories diverge from this limit cycle (Fig. 6.43b).

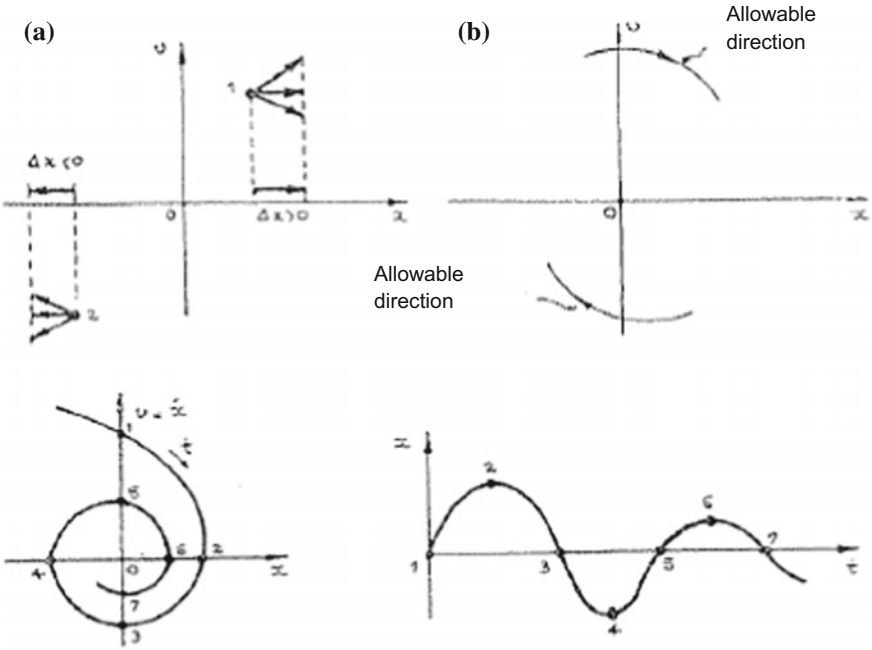



**Fig. 6.42** Up: The allowable direction of traversing the trajectories on the phase-plane is the clock-wise direction. Down: **a** Trajectory of the system $\mathrm{d}v/\mathrm{d}x = -(kx + \beta v)/mv$ for $\beta > 0$, **b** Corresponding time response

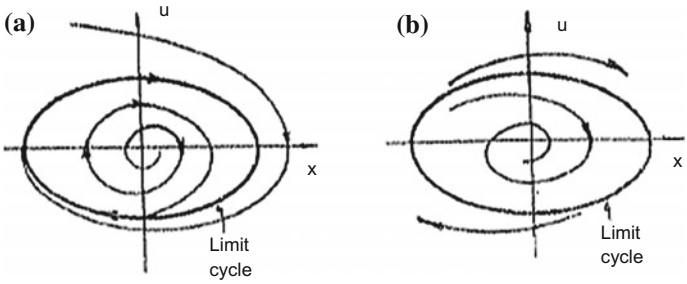



**Fig. 6.43** Limit cycle on the phase plane: **a** Stable limit cycle, **b** Unstable limit cycle

A classical stable limit cycle is the limit cycle of *Van der Pol equation*:

$$a\mathrm{d}^2x/\mathrm{d}t^2 - b\left(1 - x^2\right)\mathrm{d}x/\mathrm{d}t + cx = 0,$$

which has *negative damping* for $|x| < 1$ and *positive damping* for $|x| > 1$.

The general form of the Van der Pol limit cycle is shown in Fig. 6.44. Its actual shape depends on the values of the parameters $a$, $b$, and $c$.

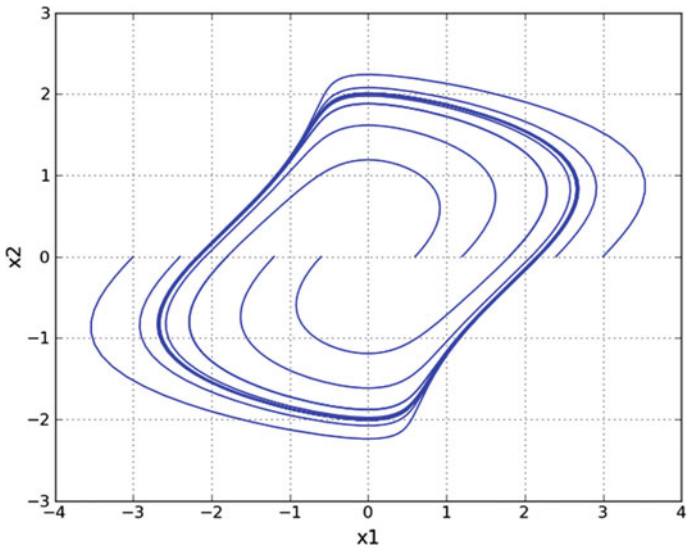



**Fig. 6.44** Typical form of the Van der Pol limit cycle. For $b = 0$, the limit cycle reduces to circles centered at the origin (Fig. 6.60)

## 6.12 Concluding Remarks

In this chapter, we have briefly summarized the basic methodologies of classical control, including a short description of positive and negative feedback and the principal historical landmarks in the study of feedback control systems. The chapter was not intended to cover all aspects and issues, but those necessary for an encyclopedic presentation which, however, by their nature cannot be purely descriptive and need some mathematics. Specifically, the following concepts and techniques were discussed in, the author hopes, a conveniently flowing manner: the basic negative feedback loop with both purposive and disturbance inputs, stability, time-domain performance specifications, root-locus, Nyquist plots and stability criterion, Bode amplitude/phase plots, frequency-domain performance specifications, Nichols plot, statement of the Nyquist criterion using the Bode and Nichols plots, discrete-time control systems analysis using the above methods, classical compensators, PID compensators, and analysis of nonlinear systems via the describing functions and phase-plane concepts. Naturally, the contents of the chapter are well-known to the specialists and professionals in control, but they are considered to offer a good, concise introduction for the non-expert or novice in the field, who has a sufficient knowledge of differential equations, Laplace transforms, and Fourier analysis. Full presentations can be found in the referenced books.

## References


1. N. Wiener, *Cybernetics or Control and Communication in the Animal and the Machine* (Paris, Herman–MIT Press, Cambridge, MA, 1948)
2. C.H. Houpis, G.B. Lamont, *Digital Control Systems: Theory, Hardware, Software* (McGraw-Hill, New York, 1992)
3. R. Isermann, *Digital Control Systems*, vol. 1: Fundamentals, Deterministic Control (Springer, Berlin, 1989)
4. P. Katz, *Digital Control Using Microprocessors* (Prentice-Hall, London, 1981)
5. R.C. Dorf, R.H. Bishop, *Modern Control Systems*, 7th edn. (Addison-Wesley, Reading, MA, 1995)
6. J.J. DiStefano III, A.R. Stubberud, I.J. Williams, *Theory and Problems of Feedback Control Systems Design* (McGraw-Hill, New York, 1990)
7. A.C. Guyton, *Textbook of Medical Physiology* (W.B. Saunders, Philadelphia, 1991)
8. A.D. Brown, *Feed or Feedback: Agriculture, Population Dynamics and the State of the Planet* (International Books, Tuross Head, NWS, Australia, 2003)
9. O. Mayr, *The Origins of Feedback Control* (MIT Press, Cambridge, MA, 1970)
10. S. Bennett, A brief history of automatic control, IEEE Control Syst. Mag., 17–25 (1996)
11. C.C. Bissel, Secondary sources for the history of control engineering: an annotated bibliography. Int. J. Control **54**, 517–528 (1991)
12. A.V. Khramoi, History of Automation in Russia before 1917, Moscow 1956, English Translation, Jerusalem, 1969
13. O. Mary, James Clerk Maxwell and the origins of cybernetics. ISIS **62**, 425–444 (1971)
14. F.L. Lewis, A brief history of feedback control, chapter 1, in: *Applied Optimal Control and Estimation* (Prentice-Hall, Upper Saddle River, NJ, 1992)
15. S. Benner, *A History of Control Engineering 1930-1955* (Peter Peregrinus, Stevenage, 1993)
16. S. Bennet, *A History of Control Engineering 1800-1930* (Peter Peregrinus, 1979) (Reprinted 1986)
17. R. Belmann, *Dynamic Programming* (Princeton University Press, New Jersey, 1957)
18. L.S. Pontryagin, V.G. Boltyansky, R.V. Gamkrelidze, E.F. Mishchenko, *The Mathematical Theory of Optimal Processes* (Wiley, New York, 1962)
19. R.E. Kalman, On the general theory of control systems, in *Proceeding of 1st IFAC Congress,* Moscow, vol. 1 (Butterworth, London, pp. 481–492, 1960)
20. R.E. Kalman, A new approach to linear filtering and prediction problems. ASME J. Basic Eng. **82**, 34–45 (1960)
21. R.E. Kalman, J.E. Bertram, Control system analysis and design via the "second method" of Lyapunov, (I) continuous-time systems, Trans. ASME J. Basic Eng., 371–393 (1960)
22. J.P. Lasalle, Some extensions of Lyapunov's second method. IRE Trans. Circ. Theory **7**, 520 (1960)
23. R.E. Kalman, R.S. Bucy, New results in linear filtering and prediction theory. ASME J. Basic Eng. **80**, 193–196 (1961)
24. M. Athans, P. Falb, *Optimal Control* (McGraw-Hill, New York, 1966)
25. F.S. Asl, M. Athans, A. Pascoal, Issues, progress and new results in robust adaptive control. J. Adapt. Control Signal Process. **20**, 519–579 (2006)
26. V.M. Popov, Absolute stability of nonlinear system of automatic control. Autom. Remote Control **22**(8), 857–875 (1961)
27. I.W. Sandberg, A frequency-domain condition for the stability of feedback systems containing a single time-varying nonlinear element. Bell. Syst. Tech. J. **43**(4), 1601–1608 (1964)
28. K.S. Narendra, A. Goldwyn, A geometrical criterion for the stability of certain nonlinear nonautonomous systems, IEEE Trans. Circ. Theory **CT-11**(3), 406–407 (1964)
29. C.A. Desoer, A generalization of the Popov criterion, IEEE Trans. Autom. Control **AC-10**(2), 182–185 (1965)

30. C.S. Draper, Bid for space 1961, MIT and project Apollo, MIT Institute Archives & Special Collections http://libraries.mit.edu/archives/exhibits/apollo/
31. K.S. Fu, Learning control systems-review and outlook, IEEE Trans. Autom. Control **AC-15**, 210–221 (1970)
32. K.S. Fu, Learning control systems and intelligent control systems: an intersection of artificial intelligence and automatic control, IEEE Trans Autom. Control **AC-16**(2), 70–72 (1971)
33. K.S. Fu, *Syntactic Pattern Recognition and Applications* (Prentice-Hall, Englewood Cliffs, 1982)
34. H. Haddad, G.C. Lee, GoS guided bandwidth management in differentiated time scales. J. Optimiz. Theory Appl. **115**, 517–547 (2002)
35. C.I. Harris, C.G. Moore, M. Brown, *Intelligent Control: Aspects of Fuzzy Logic and Neural Nets* (World Scientific, Singapore, 1993)
36. M. Brown, C.J. Harris, *Neuro-Fuzzy Adaptive Modeling and Control* (Prentice Hall, New Jersey, 1984)
37. G.N. Saridis, T.K. Dao, A learning approach to the parameter-adaptive self-organizing control problem. Automatica **8**, 589–597 (1972)
38. G.N. Saridis, Analytic formulation of the principle of increasing precision with decreasing intelligence for intelligent machines. Automatica **25**(3), 461–467 (1989)
39. H.H. Rosenbrock, G.E. Hayton, Dynamical indices of a transfer function matrix. Int. J. Control **20**, 177–189 (1974)
40. H.H. Rosenbrock, *State Space Multivariable Theory* (Wiley, New York, 1970)
41. G.F. Franklin, J.D. Powell, A. Emani-Naeini, *Feedback Control of Dynamic Systems* (Addison-Wesley, Reading, MA, 1994)
42. E.I. Jury, *Theory and Application of the Z-Transform, Method* (Wiley, New York, 1964)
43. E.I. Jury, *Sampled Data Control Systems* (Wiley, New York, 1958)
44. E.I. Jury, Stability of multidimensional systems and related problems, chapter 3, in: *Multidimensional Systems: Techniques and Applications* (S.G. Tzafestas, ed.) (Marcel Dekker, New York/Basel, 1986)
45. I. Horowitz, *Synthesis of Feedback Systems* (Academic Press, New York, 1963)
46. I. Horowitz, M. Sidi, Synthesis of feedback systems with large plant ignorance for prescribed time domain tolerances. Intl. J. Control **16**(2), 287–309 (1972)
47. I. Horowitz, Quantitative feedback theory, Proc. IEE, **129**(Part D, No. 6), 215–226 (1982)
48. C.H. Houpis, M. Pachter, Application of QFT to control system design: for engineers. Int. J. Robust Nonlinear Control **7**(6), 561–580 (1997)
49. C.H. Houpis, G.B. Lamont, *Digital Control Systems: Theory, Hardware, Software* (McGraw-Hill, New York, 1992)
50. C.H. Houpis, S.I. Rasmussen, *Quantitative Feedback Theory: Fundamentals and Applications* (Marcel Dekker New York, 1999)
51. A.G.J. MacFarlane, The development of frequency—response methods in automatic control, IEEE Trans. Autom. Control **AC-24**, 250–265 (1979)
52. A.G.J. MacFarlane, I. Postlethwaite, The generalized Nyquist stability criterion and multivariable root loci. Int. J. Control **25**, 81–127 (1977)
53. I. Postlethwaite, A.G.J. MacFarlane, *A Complex Variable Approach to the Analysis of Linear Multivariable Feedback Systems* (Springer, Berlin, 1979)
54. M.G. Safonov, Stability Margins of Diagonally Perturbed Multivariable Feedback Systems, IEE Proc. **129-D**, 251–256 (1982)
55. M.G. Safonov, J. Doyle, Minimizing Conservativeness of Robust Singular Values, in: *Multivariable Control* (S.G. Tzafestas, ed), (D. Reidel, Dordrecht, 1984)
56. K.J. Aström, B. Wittenmark, Self-tuning controllers based on pole-zero placement. Proc. IEE, Part D **127**, 120–130 (1980)
57. K.J. Aström, P. Hagander, J. Sternby, Zeros of sampled systems. Automatica **20**, 31–38 (1984)
58. K.J. Aström, B., Wittenmark, *Adaptive Control* (Addison-Wesley, Reading, MA, 1989)

59. K.J. Aström, B. Wittenmark, Problems of Identification and Control. J. Math Anal. Appl. **34**, 90–113 (1971)
60. H. Kwakernaak, Optimal Filtering in Linear Systems with Time Delays, IEEE Trans. Autom. Control **AC-12**, 169–173 (1967)
61. H. Kwakernaak, Robust control and $H_\infty$: optimization. Automatica **29**, 255–273 (1972)
62. H. Kwakernaak, R. Sivan, *Linear Optimal Control Systems* (Wiley, New York, 1972)
63. J.S. Meditch, Orthogonal projection and discrete optimal linear smoothing. SIAM J. Control **5**, 74–89 (1967)
64. J.S. Meditch, On optimal linear smoothing theory. Inf. Control **10**, 598–615 (1967)
65. J.S. Meditch, *Stochastic Optimal Linear Estimation and Control* (McGraw-Hill, New York, 1969)
66. I.D. Landau, *Adaptive Control: The Model Reference Approach* (Marcel-Dekker, New York, 1979)
67. I.D. Landau, From robust control to adaptive control. Control Eng. Pract. **7**(10), 1113–1124 (1997)
68. I.D. Landau, R. Lozano, M. M'Saad, *Adaptive Control* (Springer, New York, 1998)
69. K.S. Narendra, K. Parthasarathy, Identification and control of dynamical systems using neural networks. IEEE Trans. Neural Netw. **1**(1), 4–27 (1990)
70. K.S. Narendra, A. Annaswamy, *Stable Adaptive Systems* (Prentice-Hall, NJ, 1989)
71. J.-J.E. Slotine, J.A. Coetsee, Adaptive sliding controller synthesis for nonlinear systems. Int. J. Control **43**, 1631–1651 (1986)
72. J.-J.E. Slotine, W. Li, *Applied Nonlinear Control* (Prentice-Hall, Englewood Cliffs, N.J., 1991)
73. G. Zames, On the input-output stability of time-varying non-linear feedback systems, part I: conditions derived using concepts of loop gain, conicity, and positivity, IEEE Trans. Autom. Control, **AC-11**(2), 228–238 (1966), Part II: conditions involving circles in the frequency plane and sector nonlinearities, IEEE Trans. Autom. Control, **AC-11**(3), 465–476 (1966)
74. G. Zames, Feedback and optimal sensitivity: model reference transformations multiplicative semi-norms and approximate inverses, IEEE Trans. Autom. Control **AC-26**, 301–320 (1981)
75. G. Zames, N.A. Shneydor, Structural stabilization and quenching by dither in nonlinear systems, IEEE Trans. Autom. Control, **AC-22**(3), 352–361 (1977)
76. A.L. Fradkov, I.V. Miroshnik, V.O. Nikiforok, *Nonlinear and Adaptive Control of Complex Systems* (Springer, Berlin, 2007)
77. A. Astolfi, Towards Applied Nonlinear Adaptive Control
78. J.J. D'Azzo, C.H. Houpis, *Feedback Control System Analysis and Synthesis*, 2nd edn. (McGraw-Hill, New York, 1966)

# Chapter 7
# Feedback and Control II: Modern Methodologies

*Because we are self-controlling beings, we are also responsible for our actions. This is not a moral or ethical proposition, but simply a causal one.*

Butter Shaffer

*People have to be responsible for their thoughts, so they have to learn to control them. It may not be easy, but it can be done.*

Rolling Thunder

**Abstract** Modern control has decisively contributed to the human society development providing the means for successful control and efficient and safe operation of complex technological and non-technological systems such as computer-based systems, aircrafts, robots, automation systems, managerial systems, decision support systems, economic systems, etc. It is based on the concepts of "system state vector" and "state-space models" which are applicable to time-varying, multivariable, and nonlinear systems in both continuous-time and discrete-time representations. In this chapter, we present the fundamental concepts, principles, and methodologies covering most developments at an introductory level. Specifically, the following topics are considered: state-space modeling, Lyapunov stability, controllability and observability, optimal, stochastic, adaptive, predictive, robust, nonlinear, and intelligent control. Also, the following classes of dynamic models, that cover a wider range of natural and man-made systems, are briefly discussed: large-scale, distributed-parameter, time delay, finite state, and discrete event models. The field of modern control is still expanding offering new challenges in research and real-life bioengineering and technological applications.

**Keywords** Modern control · State space model · Canonical state-space model Lyapunov stability · Controllability · Observability · Duality principle Kalman decomposition · State-feedback control design · State-observer design Optimal control · Bellman principle of optimality · Linear-quadratic control (LQC) Pontryagin minimum principle · Gauss-Markov model · Kalman-Bucy filter Stochastic control · Separation principle · Model reference adaptive control Self-tuning control · Gain-scheduling control · Robust control · Sliding–mode



S. G. Tzafestas, *Energy, Information, Feedback, Adaptation, and Self-organization*, Intelligent Systems, Control and Automation: Science and Engineering 90, https://doi.org/10.1007/978-3-319-66999-1_7

control · Intelligent control · Neural/fuzzy control · DPS control
Time-delay control · Finite-state automata control · Discrete event systems control

## 7.1 Introduction

This chapter is a continuation of Chap. 6 on feedback and control and provides a guide to the fundamental concepts, principles, and methodologies of what is collectively known as "*modern control*". The fundamentals of the concept of feedback and the "*classical control*" including the history of control (early, preclassical, classical, and modern control) were reviewed in Chap. 6. Here, we will provide a global conceptual account to the following modern control topics that play a key role in the design and development of the advanced computer control and automation systems of modern society [1–153]:

- State-space modeling
- Lyapunov's direct stability method
- Controllability and observability
- State-feedback controllers
- Optimal and stochastic control
- Adaptive and predictive control
- Robust control
- Nonlinear control
- Intelligent control
- Large-scale systems (*LSS*) control
- Distributed-parameter systems (*DPS*) control
- Time-delay systems (*TDS*) control
- Finite-state automata/machines (*FSM*) control
- Discrete-event systems (*DES*) control.

## 7.2 The State-Space Model

### *7.2.1 General Issues*

Modeling control systems with their *state-space equations* is the basic cornerstone of modern control theory. It is applicable to time-varying, linear, nonlinear, time-delay, and distributed-parameter systems in both continuous-time and discrete-time representations. The main benefits of the state-space modeling are the following [1, 3, 9, 11]:

- Time-varying systems are described in the same way as the time-invariant systems.
- Multi-input/multi-output (**MIMO**) systems are analyzed as the single-input/single-output **(SISO)** systems.

- State-space equations are suitable for direct simulation/programming on the digital computer.
- Higher order systems can be studied without additional difficulties.

**State vector** $\mathbf{x}(\mathbf{t})$ of a system is the minimum-dimensionality vector, with components called *state variables*, the knowledge of which at $t = t_0$ together with the input (or inputs) $\mathbf{u}(\mathbf{t})$ for $t \geq t_0$ determines completely the behavior of the system for any time $t \geq t_0$.

This means that the state of a system at time $t$ is specified only by its initial value at $t = t_0$ and the input for $t \geq t_0$ and is independent of the system state and the inputs for times previous to $t = t_0$. It is noted that the state variables $x_1(t), \ldots, x_n(t)$ may not necessarily be measurable physical quantities. In practice, however, an effort is made to use, as much as possible, measurable state variables because the state-feedback control laws need all of them.

A system with a state vector $\mathbf{x}(t)$ that has $n$ components:

$$\mathbf{x}^T(t) = [x_1(t), x_2(t), \ldots, x_n(t)]$$

where $\mathbf{x}^T$ is the transpose of the column vector $\mathbf{x}$, is said to be an *n-dimensional system*. The expression of $\mathbf{x}(t)$ as a function of $t, t_0, \mathbf{x}(t_0) = \mathbf{x}_0$, and $\mathbf{u}(\tau) = [u_1(\tau), \ldots, u_m(\tau)]^T, \tau \geq t_0$, i.e.,:

$$\mathbf{x}(t) = \boldsymbol{\varphi}(t; t_0, \mathbf{x}_0, \mathbf{u}(\tau)),$$

is called the *system's trajectory*.

The output $\mathbf{y}(t)$ of the system is a similar function of $\mathbf{x}(t), \mathbf{u}(t)$ and $t$, i.e.:

$$\mathbf{y}(t) = \boldsymbol{\eta}(t; \boldsymbol{\varphi}(t, t_0, \mathbf{x}_0, \mathbf{u}(t)), \mathbf{u}(t))$$

for all $t \geq t_0$.

The trajectories satisfy the transition property:

$$\boldsymbol{\varphi}(t; t_0, \mathbf{x}(t_0), u(\tau)) = \boldsymbol{\varphi}(t; t_1, \mathbf{x}(t_1), u(\tau))$$

For all $t_0 < t_1 < t$, where $\mathbf{x}(t_1) = \boldsymbol{\varphi}(t; t_0, \mathbf{x}_0, \mathbf{u}(\tau))$. Two important properties of the trajectories are:

- $\lim_{t \to t_1} \boldsymbol{\varphi}(t; t_1, \mathbf{x}(t_1), \mathbf{u}(\tau)) = \mathbf{x}(t_1)$ for all $t_1 \geq t_0$
- $\boldsymbol{\varphi}(t; t_0, \mathbf{x}_0, \mathbf{u}(\tau))$ does not depend on $\mathbf{u}(\tau)$ for $\tau > t$.

Let two numbers $\alpha$ and $\beta$, two initial conditions $\mathbf{x}_1(t_0)$ and $\mathbf{x}_2(t_0)$, two inputs $\mathbf{u}_1(\tau)$ and $\mathbf{u}_2(\tau)$, and two outputs $\mathbf{y}_1(\tau)$ and $\mathbf{y}_2(\tau)$, for $\tau \geq t_0$. Then, the system is *linear* if the following conditions hold:

- The state $\mathbf{x}_3(t_0) = \alpha \mathbf{x}_1(t_0) + \beta \mathbf{x}_2(t_0)$, the output $\mathbf{y}_3(\tau) = \alpha \mathbf{y}_1(\tau) + \beta \mathbf{y}_2(\tau)$ and the input $\mathbf{u}_3(\tau) = \alpha \mathbf{u}_1(\tau) + \beta \mathbf{u}_2(\tau)$ are permissible.

- The output $\mathbf{y}_3(\tau)$ and state $\mathbf{x}_3(\tau)$ correspond to the initial state $\mathbf{x}_3(t_0)$ and the input $\mathbf{u}_3(\tau)$.

The linear time-varying systems constitute a very important class of systems because the linear approximations of nonlinear systems in the vicinity of nominal trajectories are always linear time-varying models.

### 7.2.2 Canonical Linear State-Space Models

The state-space equation of a continuous-time linear system has the form:

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}, \mathbf{y} = \mathbf{C}\mathbf{x} + \mathbf{D}\mathbf{u},$$

where **x** is the n-dimensional state vector, **u** is the m-dimensional input vector, **y** is the q-dimensional output vector, and **A, B, C, D** are time-invariant or time-varying matrices of proper dimensions. The block diagram of this state-space model has the form shown in Fig. 7.1.

Given a time-invariant **SISO** control system described by the *n*th-order differential equation:

$$\left(D^n + a_1 D^{n-1} + \cdots + a_{n-1} D + a_n\right) y(t) = \left(b_0 D^n + b_1 D^{n-1} + \cdots + b_n\right) u(t)$$

where $D = \mathrm{d}/\mathrm{d}t$ is the "derivative operator", one can find three alternative canonical state-space models, namely, (i) the observable canonical model, (ii) the controllable canonical model, and the Jordan (or modal) canonical model, with matrices as follows [15]:

**Observable Canonical Model**

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} -a_1 & 1 & 0 & \cdots & 0 \\ -a_2 & 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ -a_{n-1} & 0 & 0 & \cdots & 1 \\ -a_n & 0 & 0 & \cdots & 0 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} b_1 - a_1 b_0 \\ b_2 - a_2 b_0 \\ \cdots \\ b_n - a_n b_0 \end{bmatrix}$$

$$\mathbf{C} = [1 \quad 0 \quad \cdots \quad 0], \mathbf{D} = [b_0], \mathbf{u} = [u]$$

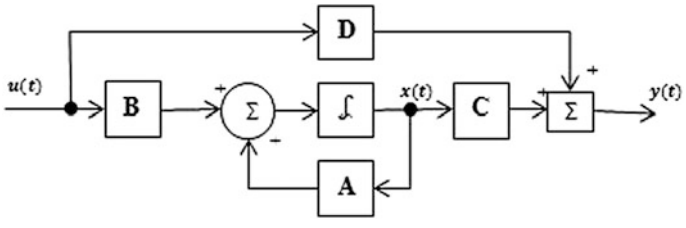



**Fig. 7.1** A block diagram of a general linear state-space model

**Controllable Canonical Model**

$$\mathbf{x}=\begin{bmatrix}x_1\\x_2\\\vdots\\x_n\end{bmatrix},\mathbf{A}=\begin{bmatrix}0&1&0&\cdots&\cdots&0\\0&0&1&&&0\\\cdots&\cdots&\cdots&\cdots&\cdots&\cdots\\0&0&0&\cdots&\cdots&1\\-a_n&-a_{n-1}&-a_{n-2}&\cdots&\cdots&-a_1\end{bmatrix},\mathbf{B}=\begin{bmatrix}0\\0\\\vdots\\0\\1\end{bmatrix},\mathbf{u}=u$$

**Modal Canonical Model**

$$\mathbf{A}=\begin{bmatrix}\lambda_1&&&\mathbf{0}\\&\lambda_2&&\\&&\ddots&\\\mathbf{0}&&&\lambda_n\end{bmatrix},\mathbf{B}=\begin{bmatrix}1\\1\\\vdots\\1\end{bmatrix},\mathbf{C}=[\rho_1,\rho_2,\ldots,\rho_n],\mathbf{D}=[b_0]$$

The observable (or first) canonical model is found by solving the system's differential equation for $y(t)$ in terms of the defining state variables $x_1, x_2, \ldots, x_n$ as:

$$x_1=\frac{1}{D}(-a_1y+b_1u)+x_2$$
$$x_2=\frac{1}{D}(-a_2y+b_2u)+x_3,\ldots,x_n=\frac{1}{D}(-a_ny+b_nu)$$

The controllable canonical model is found by defining the state variables as:

$$Dx_1=x_2, Dx_2=x_3,\ Dx_{n-1}=x_n$$

which leads to:

$$Dx_n=-a_1x_n-a_2x_{n-1}-\cdots-a_{n-1}x_2-a_nx_1+u$$
$$y=b_nx_1+b_{n-1}x_2+\cdots+b_1x_n+b_0(u-a_1x_n-\cdots-a_{n-1}x_2-a_nx_1)$$

The modal canonical form is found by expanding the system transfer function in partial fractions, namely:

$$y(s)=b_0\bar{u}(s)+\frac{\rho_1}{s-\lambda_1}\bar{u}(s)+\cdots+\frac{\rho_n}{s-\lambda_n}\bar{u}(s),\quad \lambda_1\neq\lambda_2\neq\cdots\neq\lambda_n$$

where the poles (eigenvalues) were assumed distinct ($\lambda_1\neq\lambda_2\neq\cdots\neq\lambda_n$), and defining the state variables as:

$$\bar{x}_1(s)=\frac{1}{s-\lambda_1}\bar{u}(s),\bar{x}_2(s)=\frac{1}{s-\lambda_2}\bar{u}(s),\ldots,\bar{x}_n(s)=\frac{1}{s-\lambda_n}\bar{u}(s)$$

We see that in this case the system matrix **A** is diagonal. If some eigenvalues are multiple, then the matrix **A** is block-diagonal with submatrices, the so-called *Jordan blocks*. Given a system with arbitrary matrices **A, B,** we can convert it to a desired canonical form by using a suitable state-similarity transformation.

### 7.2.3 Analytical Solution of the State Equations

The solution of the continuous-time state-space model equations for a time-invariant linear system:

$$\mathbf{dx}/\mathrm{d}t = \mathbf{Ax} + \mathbf{Bu}, \mathbf{y}(t) = \mathbf{Cx} + \mathbf{Du}$$

with initial condition $\mathbf{x}(0) = \mathbf{x}_0$ is given by:

$$\mathbf{x}(t) = \mathbf{\Phi}(t)\mathbf{x}_0 + \int_0^t \mathbf{\Phi}(t-\tau)\mathbf{Bu}(\tau)\mathrm{d}\tau$$

$$\mathbf{y}(t) = \mathbf{C\Phi}(t)\mathbf{x}_0 + \mathbf{C}\int_0^t \mathbf{\Phi}(t-\tau)\mathbf{Bu}(\tau)\mathrm{d}\tau + \mathbf{Du}(t)$$

where $\mathbf{\Phi}(t) = e^{\mathbf{A}t}$ is the *fundamental matrix* (or *transition matrix*) of the system that satisfies the condition $\mathbf{\Phi}(0) = I$ (unit matrix). The corresponding solution equations for a time-varying system:

$$\mathrm{d}\mathbf{x}/\mathrm{d}t = \mathbf{A}(t)\mathbf{x} + \mathbf{B}(t)\mathbf{u}, \quad \mathbf{y}(t) = \mathbf{C}(t)\mathbf{x} + \mathbf{D}(t)\mathbf{u}$$

are:

$$\mathbf{x}(t) = \mathbf{\Phi}(t, t_0)\mathbf{x}(t_0) + \int_{t_0}^t \mathbf{\Phi}(t,\tau)\mathbf{B}(\tau)\mathbf{u}(\tau)d\tau$$

$$\mathbf{y}(t) = \mathbf{C}(t)\mathbf{\Phi}(t, t_0)\mathbf{x}(t_0) + \mathbf{C}(t)\int_{t_0}^t \mathbf{\Phi}(t,\tau)\mathbf{B}(\tau)\mathbf{u}(\tau)\mathrm{d}t,$$

where $\mathbf{\Phi}(t, t_0) = e^{\int_{t_0}^t \mathbf{A}(\tau)\mathrm{d}\tau}$ subject to the condition $\mathbf{A}(t)\mathbf{A}(\tau) = \mathbf{A}(\tau)\mathbf{A}(t)$ for all $t$ and $\tau$.

Analogous equations can be found for discrete-time systems. For example, in the time-invariant case we have:

$$\mathbf{x}(k) = \mathbf{\Phi}(k)\mathbf{x}_0 + \sum_{i=0}^{k-1} \mathbf{\Phi}(k-i-1)\mathbf{B}\mathbf{u}(i)$$
$$\mathbf{y}(k) = \mathbf{C}\mathbf{x}(k) + \mathbf{D}\mathbf{u}(k)$$
$$\mathbf{\Phi}(k) = \mathbf{A}^k$$

The transition matrices $\mathbf{\Phi}(t) = e^{\mathbf{A}t}$ and $\mathbf{\Phi}(k) = \mathbf{A}^k$ are found by inverting their Laplace an Z transforms respectively, i.e., $\mathbf{\Phi}(t) = \mathscr{L}^{-1}(s\mathbf{I} - \mathbf{A})^{-1}, \mathbf{\Phi}(k) = \mathcal{Z}^{-1}\left\{z(z\mathbf{I} - \mathbf{A})^{-1}\right\}$ where $\mathscr{L}^{-1}$ and $\mathcal{Z}^{-1}$ are the inverse Laplace and $\mathcal{Z}$ transforms, respectively. The inverse matrices of $(s\mathbf{I} - \mathbf{A})$ and $(z\mathbf{I} - \mathbf{A})$ can be found by Cramer's rule, or using the Leverrier algorithm in more complex cases.

## 7.3 Lyapunov Stability

### *7.3.1 General Issues*

The Lyapunov stability theory has contributed substantially in the development of modern control. The algebraic stability criteria, as well as the stability criteria of Nyquist, Bode, and Nichols are applicable only to linear time-invariant systems. Lyapunov's stability method can also be applied to time-varying systems and to nonlinear systems. Lyapunov introduced the generalized notion of *energy* (called the *Lyapunov function*) and studied dynamic systems without external inputs. Combining Lyapunov's theory with the concept of **BIBO** stability (Sect. 6.5.3), we can derive stability conditions for *input-to-state* stability (**ISS**). Lyapunov's criterion can be applied to both continuous-time and discrete-time systems. Lyapunov introduced two stability methods. The first method requires the availability of the system's time response (i.e., the solution of the differential equations). The second method, also called *direct Lyapunov method,* does not require knowledge of the system's time response. A brief description of this method follows [4, 6, 17].

**Definition 1** The equilibrium state $\mathbf{x} = \mathbf{0}$ of the free system $\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x}$ is stable in the Lyapunov sense (*L-stable*) if, for every initial time $t_0$ and every real number $\varepsilon > 0$, there exists some number $\delta > 0$ as small as desired that depends on $t_0$ and $\varepsilon$, so that:

If $\|\mathbf{x}_0\| < \delta$ then $\|\mathbf{x}(t)\| < \varepsilon$ for all $t \geq t_0$, where $\|.\|$ denotes the norm of the vector $\mathbf{x}$, i.e., $\|\mathbf{x}\| = \left(x_1^2 + x_2^2 + \cdots + x_n^2\right)^{1/2}$.

**Theorem 1** *The transition matrix* $\mathbf{\Phi}(t, t_0)$ *of a linear system is bounded by* $\|\mathbf{\Phi}(t, t_0)\| < k(t_0)$ *for all* $t \geq t_0$ *if and only if the equilibrium state* $\mathbf{x} = 0$ *of* $\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x}$ *is L-stable.*

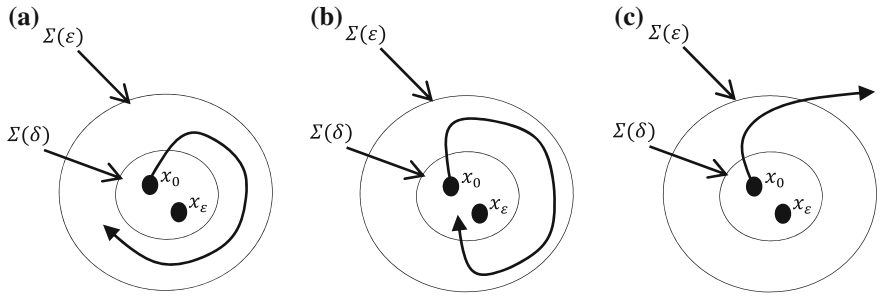



**Fig. 7.2** Illustration of L-stability (**a**) L-asymptotic stability (**b**), and instability (**c**). $\sum(\varepsilon)$ and $\sum(\delta)$ symbolize n-dimensional balls (spheres) with radii $\varepsilon$ and $\delta$, respectively

The bound of $\|\mathbf{x}(t)\|$ of a linear system does not depend on $\mathbf{x}_0$. In general, if the system stability (of any kind) does not depend on $\mathbf{x}_0$, we say that we have *global (total) stability* or *stability in the large*. If the stability depends on $\mathbf{x}_0$, then it is called *local stability*. Clearly, local stability of a linear system also implies total stability.

**Definition 2** The equilibrium state $\mathbf{x} = \mathbf{0}$ is asymptotically stable if:

(i) It is L-stable.
(ii) For every $t_0$ and $\mathbf{x}_0$ sufficiently near to $\mathbf{x} = \mathbf{0}$, the condition $\mathbf{x}(t) \to \mathbf{0}$, for $t \to \infty$, holds.

**Definition 3** If the parameter $\delta$ (in Definition 1) does not depend on $t_0$, then we have uniform L-stability.

**Definition 4** If the system $\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}$ is uniformly L-stable and for all $t_0$ and for arbitrarily large $\rho$, the relation $\|\mathbf{x}_0\| < \rho$ implies $\mathbf{x}(t) \to \mathbf{0}$ for $t \to \infty$, then the system is called *uniformly asymptotically stable*.

**Theorem 2** *The linear system* $\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x}$ *is uniformly asymptotically stable if and only if there exist two constant parameters* $k_1$ *and* $k_2$ *such that:* $\|\mathbf{\Phi}(t, t_0)\| \leq k_1 e^{-k_2(t-t_0)}$ *for all* $t_0$ *and all* $t \geq t_0$.

**Definition 5** The equilibrium state $\mathbf{x} = \mathbf{0}$ of $\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x}$ is said to be *unstable* if for some real number $\varepsilon > 0$, some $t_1 > t_0$, and any real number $\delta$ arbitrarily small, there always exists an initial stable $\|\mathbf{x}_0\| < \delta$ such that $\|\mathbf{x}(t)\| > \varepsilon$ for $t \geq t_1$.

Figure 7.2 illustrates geometrically the concepts of L-stability, L-fasymptotic stability, and instability.

### 7.3.2 Direct Lyapunov Method

Let $\mathrm{d}(\mathbf{x}(t), \mathbf{0})$ be the distance of the state $\mathbf{x}(t)$ from the origin $\mathbf{x} = \mathbf{0}$ (defined using any valid norm). If we find some distance $\mathrm{d}(\mathbf{x}(t), \mathbf{0})$ that tends to zero for $t \to \infty$,

then we conclude that the system is asymptotically stable. To show that a system is asymptotically stable using Lyapunov's direct method, we do not need to find such a *distance* (norm) but a *Lyapunov function* which is actually a generalized energy function.

**Definition 6** Any scalar function $V(\mathbf{x})$ of $\mathbf{x}$ that, for all $t \geq t_0$ and $\mathbf{x}$ in the vicinity of the origin, satisfies the following four conditions is called a *time-invariant Lyapunov function.*

(i) $V(\mathbf{x})$ is continuous and has continuous derivatives
(ii) $V(\mathbf{0}) = 0$
(iii) $V(\mathbf{x}) > 0$, for all $\mathbf{x} \neq 0$
(iv) $\frac{\mathrm{d}V(\mathbf{x})}{\mathrm{d}t} = \left[\frac{\partial V(\mathbf{x})}{\partial \mathbf{x}}\right]^T \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} < 0$, for $\mathbf{x} \neq \mathbf{0}$

**Theorem 3** (Lyapunov theorem) *If a Lyapunov function $V(\mathbf{x})$ can be found for the state of a nonlinear or linear system $\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), t)$ where $\mathbf{f}(\mathbf{0}, t) = \mathbf{0}$ ($\mathbf{f}$ is a general function), then the state $\mathbf{x} = \mathbf{0}$ is asymptotically stable.*

*Remarks*

(i) If Definition 5 holds for all $t_0$, then we have "uniformly asymptotic stability".
(ii) If the system is linear, or we replace in Definition 5 the condition "$\mathbf{x}$ *in the vicinity of the origin*" by the condition *"for* $\mathbf{x}$ *everywhere"*, then we have "total asymptotic stability.
(iii) If the condition (iv) of Definition 5 becomes $dV(\mathbf{x})/dt \leq 0$, then we have (simple) L-stability.

Clearly, to establish L-stability of a system, we must find a Lyapunov function. Unfortunately, no general methodology exists for this.

The above definitions and results hold also for discrete-time systems: $\mathbf{x}(k+1) = \mathbf{f}(\mathbf{x}(k))$.

## 7.4 Controllability and Observability

### *7.4.1 Controllability*

The ultimate goal of any automatic control system is to improve (and, if possible, to optimize) the performance of the physical process under control. The question raised here is whether a satisfactory controller can actually be designed that provides this improvement. In many cases, the control input affects the entire system, and so such a proper controller exists. But, in many other cases (especially in MIMO systems), some control inputs affect only part of the dynamic performance. The concept of *controllability* has been developed exactly to study the ability of a

controller to alter the performance of a system in an arbitrary desired way [3, 9, 11, 15].

**Definition** A state $\mathbf{x}_0$ of a system is called *totally controllable* if it can be driven to a final state $\mathbf{x}_f$ as quickly as desired, independently of the initial time $t_0$. A system is said to be *totally controllable* if all of its states are totally controllable.

Intuitively, we can see that, if some state variables do not depend on the control input $\mathbf{u}(t)$, no possibility exists that can drive it to some other desired state. Thus, this state is called a "*noncontrollable state*". If a system has at least one noncontrollable state, it is said to be non-totally controllable or, simply, noncontrollable. The above controllability concept refers to the states of a system, and so it is characterized as *state controllability*. If the controllability is referred to the outputs of a system, then we have the so-called *output controllability*. In general, state controllability and output controllability are not the same.

An $n$-dimensional, linear, time-invariant system with $m$-dimensional input vector

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}, \mathbf{u} = [u_1, u_2, \ldots, u_m]^T$$

is state controllable if and only if the controllability matrix:

$$\mathbf{P}_c = \left[\mathbf{B} \vdots \mathbf{A}\mathbf{B} \vdots \mathbf{A}^2\mathbf{B} \vdots \cdots \vdots \mathbf{A}^{n-1}\mathbf{B}\right]$$

has rank $n$, where $n$ is the dimensionality of the state vector $\mathbf{x}$, i.e., if and only if

$$rank\mathbf{P}_c = n$$

An analogous controllability criterion can be formulated for discrete-time systems, where $\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + \mathbf{B}\mathbf{u}$. In particular, the controllability matrix has exactly the same form:

$$\mathbf{P}_c = \left[\mathbf{B} \vdots \mathbf{A}\mathbf{B} \vdots \mathbf{A}^2\mathbf{B} \vdots \cdots \vdots \mathbf{A}^{n-1}\mathbf{B}\right]$$

and the discrete-time system is totally state controllable if and only if

$$rank\mathbf{P}_c = n$$

Two further controllability criteria can be formulated using the Grammian matrix

$$\mathbf{W}_c(t_0, t_f) = \int_{t_0}^{t_f} e^{\mathbf{A}(t_0-\tau)}\mathbf{B}\mathbf{B}^T e^{\mathbf{A}^T(t_0-\tau)}\mathrm{d}\tau$$

and the Jordan canonical form.

### 7.4.2 Observability

Observability theory deals with the problem of determining whether the state variables of a system can be measured or estimated using only the input/output relation of the system and the measured output history from the initial time to the present. In practice, not all state variables can be directly measured because the location of a particular state variable may not be physically accessible, or there are not available proper measurement instruments for this state variable or this particular state is not a real physical variable but a dummy one. *Observability* is a dual concept to *controllability*. State observability is defined as follows:

(a) The state $\mathbf{x}(t)$ of a system is said to be observable at some time instant, $t$, if the knowledge of the input $\mathbf{u}(t)$ and the output $\mathbf{y}(\tau)$ for some finite interval $t_0 \leq \tau \leq t$ determines completely $\mathbf{x}(t)$.
(b) A system is said to be totally observable if all states $\mathbf{x}(t)$ are observable.
(c) If the observability depends on the initial time $t_0$, then the system is called observable at the time $t_0$.
(d) If the state can be determined for $\tau$ in any arbitrarily small time interval, independently of $t_0$, then it is called a totally observable state.
(e) The system $\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t), \mathbf{y}(t) = \mathbf{C}(t)\mathbf{x}(t)$, for $t \geq t_0$ where $\mathbf{x}$ is $n$-dimensional, $\mathbf{u}$ is $r$-dimensional, and $\mathbf{y}$ is $m$-dimensional, is observable if the initial state $\mathbf{x}(t_0)$ can be determined on the basis of knowledge of $\mathbf{u}(\tau)$ and $\mathbf{y}(\tau), t_0 \leq \tau \leq t_f$ for some finite $t_f$. In analogy to Definition 9, observability can again be total observability at the time $t_0$.
(f) A similar definition of observability holds for the discrete-time system $\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + \mathbf{B}(k)\mathbf{u}(k), \mathbf{y}(k) = \mathbf{C}(k)\mathbf{x}(k), k \geq k_0$ ($\mathbf{x}$ is an n-vector, $\mathbf{u}$ is an $r$ vector, and $\mathbf{y}$ is an m-vector).

In analogy to controllability, we have the following observability criteria:

- Observability criterion via Jordan form
- Criterion via the observability Grammian matrix $\mathbf{W}_0(t_0, t_f)$

$$\mathbf{W}_0(t_0, t_f) = \int_{t_0}^{t_f} \mathbf{\Phi}^T(t, t_0)\mathbf{C}^T(t)\mathbf{C}(t)\mathbf{\Phi}(t, t_0)\mathrm{d}t$$

- Criterion via the observability matrix

$$\mathbf{S}_0 = \begin{bmatrix} \mathbf{C} \\ \mathbf{CA} \\ \vdots \\ \mathbf{CA}^{n-1} \end{bmatrix}$$

The system is totally observable if and only if the observability Grammian matrix is invertible (positive definite) for some $t_f > t_0$. Similarly, the system is totally observable if and only if $rank\mathbf{S}_0 = n$.

Analogously, observability criteria hold for the case of discrete-time systems

$$\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + \mathbf{B}\mathbf{u}(k), \mathbf{y}(k) = \mathbf{C}\mathbf{x}(k).$$

### 7.4.3 *Controllability-Observability, Duality, and Kalman Decomposition*

The duality between controllability and observability was discovered by Kalman who observed that the determinant of the Grammian matrix is analogous to Shannon's entropy and formulated the following theorem (principle) [17, 18, 48].

**Duality Principle**

Let the system $\mathbf{S}_1$:

$$\mathbf{S}_1 \begin{cases} \dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), \mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) & \text{(Continuous time)} \\ \mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + \mathbf{B}\mathbf{u}(k), \mathbf{y}(k) = \mathbf{C}\mathbf{x}(k) & \text{(Discrete time)} \end{cases}$$

where the dimensionality of $\mathbf{x}$, $\mathbf{u}$ and $\mathbf{y}$ is $n$, $r$, and $m$, respectively, and the system $\mathbf{S}_2$:

$$\mathbf{S}_2 \begin{cases} \dot{\mathbf{x}}^*(t) = \mathbf{A}^T\mathbf{x}^*(t) + \mathbf{C}^T\mathbf{v}(t), \mathbf{y}^*(t) = \mathbf{B}^T\mathbf{x}^*(t) & \text{(Continuous time)} \\ \mathbf{x}^*(k+1) = \mathbf{A}^T\mathbf{x}^*(k) + \mathbf{C}^T\mathbf{v}(k), \mathbf{y}^*(k) = \mathbf{B}^T\mathbf{x}^*(k) & \text{(Discrete time)} \end{cases}$$

Then, the duality principle states that $\mathbf{S}_1$ totally states controllable (observable) if and only if the system $\mathbf{S}_2$ totally states observable (controllable).

This follows directly by inspecting the controllability and observability matrices of the two systems, namely:

| **Controllability matrix** | **Observability matrix** |
|---|---|
| $\mathbf{P} = \left[\mathbf{B}\vdots\mathbf{A}\mathbf{B}\vdots\cdots\vdots\mathbf{A}^{n-1}\mathbf{B}\right]$ | $\mathbf{S}_0 = \begin{bmatrix} \mathbf{C} \\ \mathbf{C}\mathbf{A} \\ \vdots \\ \mathbf{C}\mathbf{A}^{n-1} \end{bmatrix}$ |
| $\mathbf{P}^* = \left[\mathbf{C}^T\vdots\mathbf{A}^T\mathbf{C}^T\vdots\cdots\vdots\left(\mathbf{A}^T\right)^{n-1}\mathbf{C}^T\right]$ | $\mathbf{S}_0^* = \begin{bmatrix} \mathbf{B}^T \\ \mathbf{B}^T\mathbf{A}^T \\ \vdots \\ \mathbf{B}^T\left(\mathbf{A}^T\right)^{n-1} \end{bmatrix}$ |

We observe that

$$\left(\mathbf{S}_0^*\right)^T = \mathbf{P} \text{ and } \left(\mathbf{P}^*\right)^T = \mathbf{S}_0$$

Using the duality principle, one can study the controllability of a system by studying the observability of its dual, and vice versa.

**Kalman decomposition** is the decomposition of the state vector $\mathbf{x}$ of a system $\Sigma$ into the following four parts: $\mathbf{x}^T = [\mathbf{x}^{1T}, \mathbf{x}^{2T}, \mathbf{x}^{3T}, \mathbf{x}^{4T}]$:

$\mathbf{x}^1$ Controllable and observable (System $\Sigma_1$)
$\mathbf{x}^2$ Uncontrollable and observable (System $\Sigma_2$)
$\mathbf{x}^3$ Controllable and unobservable (System $\Sigma_3$)
$\mathbf{x}^4$ Uncontrollable and unobservable (System $\Sigma_4$)

This can be done using a proper similarity state transformation $\mathbf{T}$ that leads to a system with state $\hat{\mathbf{x}}$ which is split into the four parts: $\hat{\mathbf{x}}^1, \hat{\mathbf{x}}^2, \hat{\mathbf{x}}^3$ and $\hat{\mathbf{x}}^4$. Kalman decomposition is pictorially illustrated in Fig. 7.3.

Closely related to controllability and observability are the concepts of *reachability* and *constructability* that are defined as follows.

**State reachability** A state $\mathbf{x}_1$ is said to be reachable if there exists an input that drives the system from the initial state $\mathbf{x}_0$ to $\mathbf{x}_1$ in some finite time interval $t_0 \leq \tau \leq t$. More precisely, the state $\mathbf{x}_1$ is reachable at time $t_1$ if, for some given initial time $t_0$, there exists an input $\mathbf{u}(t)$ that drives the state $\mathbf{x}(t)$ from the zero state (origin) at $t_0$ to $\mathbf{x}_1$.

**System reachability** A system is reachable at time $t_1$ if every state $\mathbf{x}_1$ of the system is reachable at time $t_1$.

**System constructability** A system is constructable if its present state can be determined from the present and past inputs. An observable system is also constructable, but the converse does always not hold.

Reachability and constructability can be confirmed with criteria analogous to controllability and observability.

## 7.5 State-Feedback Controllers

### *7.5.1 General Issues*

State-feedback control is more powerful than classical control because the design of a total controller for a **MIMO** system is performed in a unified way for all control loops simultaneously and not serially one loop after the other, which does not guarantee the overall system stability and robustness. The three standard linear state- feedback controllers are the following [1, 8, 9, 17]:

- Eigenvalue placement controller
- Noninteracting (decoupling) controller

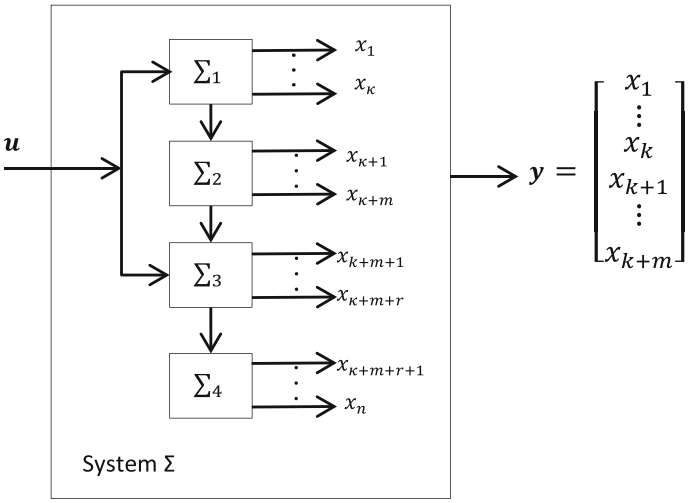



**Fig. 7.3** Graphical representation of Kalman decomposition

- Model matching controller

Let a **SISO** system

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}u(t), y(t) = \mathbf{C}\mathbf{x}(t) + Du(t)$$

where $\mathbf{A}$ is an $n \times n$ constant matrix, $\mathbf{B}$ is an $n \times 1$ constant matrix (column vector), $\mathbf{C}$ is an $1 \times n$ matrix (row vector), $u$ is a scalar input, and $D$ is a scalar constant. In this case, a state-feedback controller has the form

$$u(t) = \mathbf{F}\mathbf{x}(t) + \upsilon(t),$$

where $\upsilon(t)$ is a new control input and $\mathbf{F}$ is an $n$-dimensional constant row vector: $\mathbf{F} = [f_1, f_2, \ldots, f_n]$. Introducing this control law into the system, we get the state equations of the closed-loop (feedback) system:

$$\dot{\mathbf{x}}(t) = (\mathbf{A} + \mathbf{B}\mathbf{F})\mathbf{x}(t) + \mathbf{B}\upsilon(t), y(t) = (\mathbf{C} + \mathbf{D}\mathbf{F})\mathbf{x}(t) + D\upsilon(t)$$

A similar state-feedback controller can also be formulated for discrete-time systems:

$$\begin{aligned}
\mathbf{x}(k+1) &= \mathbf{A}\mathbf{x}(k) + \mathbf{B}u(k), y(k) = \mathbf{C}\mathbf{x}(k) + Du(k) \\
u(k) &= \mathbf{F}\mathbf{x}(k) + \upsilon(k), \mathbf{F} = [f_1, f_2, \ldots, f_n] \\
\mathbf{x}(k+1) &= (\mathbf{A} + \mathbf{B}\mathbf{F})\mathbf{x}(k) + B\upsilon(k)
\end{aligned}$$

### 7.5.2 Eigenvalue Placement Controller

Here, the problem is to select the controller gain matrix $\mathbf{F}$ so that the eigenvalues of the closed-loop matrix $\mathbf{A}+\mathbf{BF}$ are placed at desired positions $\lambda_1, \lambda_2, \ldots, \lambda_n$. It can be shown that this can be done (i.e., the system eigenvalues are controllable by state feedback) if and only if the system $(\mathbf{A}, \mathbf{B})$ is totally controllable.

Three fundamental techniques by which the feedback matrix can be selected are:

- Use of the controllable canonical form
- Equating the characteristic polynomials
- Ackerman technique

Here, the first two techniques will be outlined.

**Technique via the Controllable Canonical Form**

(i) We first write down the characteristic polynomial $\chi_{\mathbf{A}}(s)$ of the matrix $\mathbf{A}$:

$$\chi_{\mathbf{A}}(s) = |s\mathbf{I} - \mathbf{A}| = s^n + a_1 s^{n-1} + \cdots + a_{n-1}s + a_n$$

(ii) Then, we find a similarity transformation $\mathbf{T}$ that converts the given system to its controllable canonical form $\hat{\mathbf{A}} = \mathbf{T}^{-1}\mathbf{A}\mathbf{T}$.

(iii) From the desired eigenvalues of the closed-loop system, we determine the desired characteristic polynomial:

$$\chi_{\text{desired}}(s) = s^n + \tilde{a}_1 s^{n-1} + \cdots + \tilde{a}_{n-1}s + \tilde{a}_n$$

(iv) The feedback-gain matrix $\hat{\mathbf{F}}$ of the controllable canonical model is given by:

$$\hat{\mathbf{F}} = \mathbf{F}\mathbf{T} = \left[\hat{f}_n, \hat{f}_{n-1}, \ldots, \hat{f}_1\right]$$

(v) Equating the last rows of $\mathbf{A}+\mathbf{BF}$ and $\hat{\mathbf{A}}+\mathbf{B}\hat{\hat{F}}$ we find:

$$a_1 - \hat{f}_1 = \tilde{a}_1, a_2 - \hat{f}_2 = \tilde{a}_2, \ldots, a_n - \hat{f}_n = \tilde{a}_n$$

and, so solving for $\mathbf{F} = \hat{\mathbf{F}}\mathbf{T}^{-1}$, gives:

$$\begin{aligned}\mathbf{F} &= [\hat{f}_n, \hat{f}_{n-1}, \ldots, \hat{f}_1]T^{-1} \\ &= [a_n - \tilde{a}_n, a_{n-1} - \tilde{a}_{n-1}, \ldots, a_1 - \tilde{a}_1]\mathbf{T}^{-1}\end{aligned}$$

**Technique via Direct Equation of the Characteristic Polynomials**

We first check if the system is controllable and, in the positive case, we proceed to the following steps:

(i) From the desired eigenvalues, we write the desired characteristic polynomial:

$$\chi_{\text{desired}}(s) = s^n + \tilde{a}_1 s^{n-1} + \ldots + \tilde{a}_{n-1}s + \tilde{a}_n$$

(ii) We find the closed-loop characteristic polynomial $|s\mathbf{I} - \mathbf{A} - \mathbf{BF}|$ by computing the indicated determinant.

(iii) We equate the two polynomials

$$|s\mathbf{I} - \mathbf{A} - \mathbf{BF}| = \chi_{\text{desired}}(s)$$

and find the components $f_i (i = 1, 2, \ldots, n)$ of $\mathbf{F}$.

The difficulty in this methodology is that the resulting equations involve the coefficients $f_i (i = 1, 2, \ldots, n)$ nonlinearly.

### 7.5.3 Discrete-Time Systems

All the above techniques are properly applicable to discrete-time systems. In particular, if the desired characteristic polynomial is chosen as $\chi_{\text{desired}}(z) = z^n$, the resulting controller can drive the system from any initial state $\mathbf{x}(0) \neq \mathbf{0}$ to the zero state $\mathbf{x}(n) = \mathbf{0}$ in $n$ sampling periods (steps). In this case, the controller is called a *dead-beat controller*. Analogous, but more complicated, techniques are available for state-feedback eigenvalue-placement controllers for *multi-input* systems.

### 7.5.4 Decoupling Controller

Consider a MIMO system with $m$ inputs and $m$ outputs. We say that the system is input-output decoupled (or noninteracting) if each output is affected by one only input and each input affects only one output. The transfer matrix $\mathbf{G}(s)$ of an input-output decoupled system is diagonal, i.e.,:

$$\mathbf{G}(s) = \text{diag}[g_{11}(s), \ldots, g_{ii}(s), \ldots, g_{mm}(s)]$$

in which case, the outputs $y_i(s)(i = 1, 2, \ldots, m)$ are given by $y_i(s) = g_{ii}(s)u_i(s)$, where $u_i(s)(i = 1, 2, \ldots, m)$ are the inputs of the system.

Falb and Wolovich [11, 13] have shown that a system, $\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{Bu}, \mathbf{y} = \mathbf{Cx}$ ($\mathbf{u}$ is an m-vector, $\mathbf{y}$ is an m-vector) that has no diagonal transfer matrix $\mathbf{G}(s)$ with $|\mathbf{G}(s)| \neq 0$ can be decoupled by a state-feedback controller:

$$\mathbf{u}(t) = \mathbf{Fx}(t) + \mathbf{G}\upsilon(t)$$

if and only if the matrix (called 'decouplability matrix'):

$$\mathbf{B}^* = \begin{bmatrix} \mathbf{c}_1\mathbf{A}^{d_1}\mathbf{B} \\ \mathbf{c}_2\mathbf{A}^{d_2}\mathbf{B} \\ \vdots \\ \mathbf{c}_m\mathbf{A}^{d_m}\mathbf{B} \end{bmatrix}$$

is nonsingular (invertible), i.e., if and only if $|\mathbf{B}^*| \neq 0$. Here, $\mathbf{c}_i$ is the $i$th row of $\mathbf{C}$ and the indexes $d_i(i = 1, 2, \ldots, m)$ are equal to:

$$d_i = \begin{cases} \min\{j : \mathbf{c}_i\mathbf{A}^j\mathbf{B} \neq \mathbf{0}, j = 0, 1, \ldots, n-1\} \\ \quad n-1 \quad \textit{if } \mathbf{c}_i\mathbf{A}^j\mathbf{B} = \mathbf{0} \quad \text{for all } j \end{cases}$$

A solution for $\mathbf{F}$ and $\mathbf{G}$ that decouples the system is:

$$\mathbf{F} = -(\mathbf{B}^*)^{-1}\mathbf{A}^*, \mathbf{G} = (\mathbf{B}^*)^{-1}$$

where

$$\mathbf{A}^* = \begin{bmatrix} \mathbf{c}_1\mathbf{A}^{d_i+1} \\ \vdots \\ \mathbf{c}_m\mathbf{A}^{d_m+1} \end{bmatrix}$$

*Remarks*

(i) If for a system with $|\mathbf{G}(s)| \neq 0$ we find $|\mathbf{B}^*| = 0$, then we cannot decouple its inputs and outputs by static-state feedback. In this case, we use an integral controller with state equation $\mathrm{d}\bar{\mathbf{x}}/\mathrm{d}t = \mathbf{I}\bar{\mathbf{u}}$ ($\mathbf{I}$ = unit matrix).

(ii) If for a system we find that $|\mathbf{G}(s)| = 0$ and $|\mathbf{B}^*| = 0$, then it cannot be input-output decoupled, because it has *strong internal coupling*. If $|\mathbf{G}(s)| \neq 0$ and $|\mathbf{B}^*| \neq 0$, the system has no internal coupling, and its inputs and output can be decoupled by the above static-state-feedback controller $(\mathbf{F}, \mathbf{G})$.

(iii) The above controller is valid when the system has an equal number of inputs and outputs. If the number of inputs is not the same as the number of outputs,

then, under certain conditions, it is possible to decompose the inputs and outputs into an equal number of groups and then decouple these groups by designing a proper state-feedback controller.

(iv) In some cases, it is possible to decouple the inputs and outputs (or groups of them) by using an output feedback controller $\mathbf{u} = \mathbf{K}\mathbf{y} + \mathbf{G}\upsilon$, where $\mathbf{K}$ is a suitable output feedback-gain matrix.

### 7.5.5 Model Matching Controller

In this case, the problem is to find a state-feedback controller such that:

$$\mathbf{u} = \mathbf{F}\mathbf{x} + \mathbf{G}\upsilon$$

which, when applied to the system $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}, \mathbf{y} = \mathbf{C}\mathbf{x} + \mathbf{D}\mathbf{u}$, leads to a closed-loop system that matches the transfer function of a desired model:

$$\dot{\hat{\mathbf{x}}} = \widehat{\mathbf{A}}\hat{\mathbf{x}} + \widehat{\mathbf{B}}\hat{\upsilon}, \quad \hat{\mathbf{y}} = \widehat{\mathbf{C}}\hat{\mathbf{x}} + \widehat{\mathbf{D}}\hat{\upsilon}$$

This means that it is desired to match the zeros, the poles, and the D.C. gains of the closed-loop system to those of the desired model.

A suitable technique is to use the controllable canonical form of the system under control $(\mathbf{A}, \mathbf{B}; \mathbf{C}, \mathbf{D})$. Under certain conditions, it is possible to have *exact model matching*. If not, then one may obtain *approximate model matching,* depending on the approximation criterion used. In the general case, the derivation of the controller is somewhat complicated.

A SISO system can be put in one of the representations shown in Fig. 7.4 [3, 8, 11].

It should be noted that the state feedback does not change the zeros of a system, i.e., the numerator of its transfer function, and also its degree. Therefore, we must first check if the following conditions hold:

**Condition 1** The degree of the desired model $H_{\text{desired}}(s)$ is the same as the degree of the system under control $G(s)$.

**Condition 2** The zeros of $H_{\text{desired}}(s)$ and $G(s)$ are the same.

In general, if the *condition 1* does not hold, it is impossible to achieve model matching. If *condition 2* does not hold, model matching can be obtained by adding series compensators that add the zeros of the desired model that the system $G(s)$ does not have.

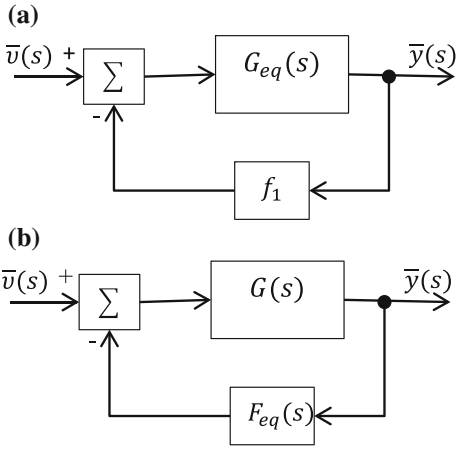



**Fig. 7.4** **a** All internal loops are combined to give the equivalent $G_{\text{eq}}(s)$, except for the outer feedback loop. **b** We keep the direct branch $G(s)$ and combine all feedback paths in an overall equivalent feedback $F_{\text{eq}}(s)$

### 7.5.6 State-Observer Design

All previous deterministic state-feedback controllers are based on the assumption that the whole state of the system is precisely known. In many actual cases, this is not the case, and so we have to reconstruct in the best way the state vector $\mathbf{x}(t)$ using only the available (measured) input and output signals $\mathbf{u}(t)$ and $\mathbf{y}(t)$. This problem is called *state-observer* design and was first solved by Luenberger [14–16].

Consider the system

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), \mathbf{y}(t) = \mathbf{C}\mathbf{x}(t),$$

where all symbols have the usual meaning. An observer of the full state system state $\mathbf{x}(t)$ provides an estimate (reconstructed) $\hat{\mathbf{x}}(t)$ on the basis of $\mathbf{u}(t)$ and $\mathbf{y}(t)$. For the above system, we can select the *observer* to have the following form:

$$\dot{\hat{\mathbf{x}}} = \hat{\mathbf{A}}\hat{x} + \mathbf{K}\mathbf{y} + \mathbf{G}\mathbf{u}$$

where $\hat{\mathbf{x}}$ has the same dimensionality as $\mathbf{x}$, and $\hat{\mathbf{A}}, \mathbf{K}, \mathbf{G}$ have appropriate dimensions. The matrices $\hat{\mathbf{A}}, \mathbf{K}$, and $\mathbf{G}$ must be selected such that $\hat{\mathbf{x}}(t) \to \mathbf{x}(t)$ as $t \to \infty$. It can be shown that the estimation error $\mathbf{e}(t)$ given by $\mathbf{e}(t) = \mathbf{x}(t) - \hat{\mathbf{x}}(t)$ tends to zero if and only if the original system $(\mathbf{A}, \mathbf{C})$ completely states observable and $\mathbf{K}$ is selected such that all eigenvalues of $\hat{\mathbf{A}} = \mathbf{A} - \mathbf{K}\mathbf{C}$ have negative real parts, and $\mathbf{G} = \mathbf{B}$. This follows from the fact that the error dynamics is described by:

$$\dot{\mathbf{e}}(t) = \hat{\mathbf{A}}\mathbf{e}, \mathbf{e}(t_0) = \mathbf{e}(0) \neq \mathbf{0},$$

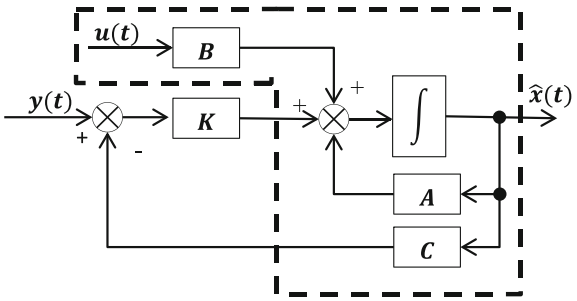



**Fig. 7.5** Block diagram of the full-order state observer

where $\mathbf{e}(0)$ is the initial value of the error (in general nonzero). Clearly, the matrix $\mathbf{K}$ can be determined by using an eigenvalue placement-control method or optimizing (minimizing) a squared error as in the case of the Kalman filter.

Introducing the conditions $\hat{\mathbf{A}} = \mathbf{A} - \mathbf{KC}$ and $\mathbf{G} = \mathbf{B}$ into the assumed observer equation $\dot{\hat{\mathbf{x}}} = \hat{\mathbf{A}}\hat{\mathbf{x}} + \mathbf{Ky} + \mathbf{Gu}$, we find the model: $\dot{\hat{\mathbf{x}}} = \mathbf{A}\hat{\mathbf{x}} + \mathbf{Bu} + \mathbf{K}(\mathbf{y} - \mathbf{C}\hat{\mathbf{x}})$ which has the block diagram shown in Fig. 7.5.

We observe from Fig. 7.5 that the observer consists of a copy of the original system $(\mathbf{A}, \mathbf{B})$ with an extra input $\mathbf{K}(\mathbf{y} - \mathbf{C}\hat{\mathbf{x}}) = \mathbf{K}\tilde{\mathbf{y}}$, where $\tilde{\mathbf{y}} = \mathbf{y} - \mathbf{C}\hat{\mathbf{x}}$ is called the *output residual* (or the *innovation signal*). The above observer reconstructs all $n$ components of the state vector independently of the number $p$ of available measurements (hence the name *full-order observer*). But, we can also design, by the same technique, an observer that reconstructs only $n - p$ states. This observer is known as a *reduced-order observer*. Similar equations can also be derived for discrete-time systems.

## 7.6 Optimal and Stochastic Control

### 7.6.1 General Issues: Principle of Optimality

*Optimal multivariable control theory* in its current form started at the beginning of the 1960s; it is concerned with the design of state-feedback controllers that minimize (optimize) several *performance criteria (cost functions)* that depend on the system variables. Historically, the first optimal control problem is the so-called "*brachystochrone problem*" solved by *Bernoulli* at Groningen (the Netherlands) in 1696. In general, optimal-control theory is an extension of the calculus of variations, mainly originated by *Lev Pontryagin* and *Richard Bellman* (see [5, 7, 12]). The performance criteria are not simple static functions of the state and control variables because the system is dynamic and changes with time. Thus, the performance criteria depend on all the values of the variables over the time interval concerned.

*Stochastic control theory* is an extension of standard (deterministic) optimal control theory applied to stochastic dynamic systems with stochastic performance criteria. Here we will deal with linear systems where the state-feedback controller has to minimize a quadratic (energy-like) cost function. To this end, Bellman *dynamic programming (principle optimality)* will be used.

**Principle of Optimality**

An optimal policy (or optimal control policy) has the property that, for every initial state and initial decision, the remaining decisions constitute an optimal policy with respect to the state that results from the initial decision.

This principle will be applied to the following general discrete-time optimal control problem:

*Given* a discrete-time system

$$\mathbf{x}_{k+1} = \mathbf{F}_k(\mathbf{x}_k, \mathbf{u}_k), k = 1, 2, \ldots, N,$$

*find* the control sequence $\{\mathbf{u}_k\} = \{\mathbf{u}_k : k = 1, 2, \ldots, N\}$ that minimizes the cost function

$$J_N = \sum_{k=1}^{N} L_k(\mathbf{x}_k, \mathbf{u}_k)$$

According to the principle of optimality, $J_N$ can be written as:

$$J_N = L_1(x_1, u_1) + J_{N-1}(x_2) = L_1(x_1, u_1) + J_{N-1}\{F(x_1, u_1)\},$$

where the first term is the initial cost and the second term is the optimal cost resulting from the next $N-1$ decisions. Thus, the optimal cost $J_N^0(x_1)$ is given by:

$$J_N^0(x_1) = \min_{u_1}[L_1(x_1, u_1) + J_{N-1}\{F(x_1, u_1)\}], N \geq 2$$

For $N = 1$, we have $J_1^0(x_1) = \min_{u_1} L_1(x_1, u_1)$. Thus, for $N = N, N-1, N-2, \ldots, 2, 1$ we get:

$$J_N^0(x_1) = \min_{u_1}\left[L_1(x_1, u_1) + J_{N-1}^0(x_2)\right], x_2 = F_1(x_1, u_1)$$
$$J_{N-1}^0(x_2) = \min_{u_2}\left[L_1(x_2, u_2) + J_{N-2}^0(x_3)\right], x_3 = F_2(x_2, u_2)$$
$$\ldots$$
$$\ldots$$
$$J_2^0(x_{N-1}) = \min_{u_{N-1}}\left[L_{N-1}(x_{N-1}, u_{N-1}) + J_1^0(x_N)\right], x_N = F_{N-1}(x_{N-1}, u_{N-1})$$
$$J_1^0(x_N) \min_{u_N}[L_N(x_N, u_N)]$$

Consequently, starting from $J_1^0(x_N)$, we compute $u_N$, then we compute $u_{N-1}$ from $J_2^0(x_{N-1})$, and so on. This means that, applying the principle of optimality, we find an optimal control policy in which the current control action depends only on the current state and the current time and does not depend on previous states or control actions. A policy of this type is known as *Markovian policy*.

### 7.6.2 Application of the Principle of Optimality to Continuous-Time Systems

Consider a general (nonlinear) continuous-time dynamic system

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}, \mathbf{u}, t), \mathbf{x}(t_0) = \mathbf{x}_0$$

It is desired to determine the control input $\mathbf{u}(t), t_0 \leq t \leq t_f$, which minimizes the cost functional:

$$J(\mathbf{u}) = \int_{t_0}^{t_f} L(\mathbf{x}, \mathbf{u}, t)\mathrm{d}t$$

To this end, we define

$$J^0(\mathbf{x}, t) = \min_{\mathbf{u}(\tau):\tau\in[t,t_f]} \int_t^{t_f} L(\mathbf{x}, \mathbf{u}, \tau)\mathrm{d}\tau,$$

where $J^0(\mathbf{x}, t_0) = J^0(\mathrm{x}_0)$. Applying the principle of optimality, the optimization can be performed in two stages, i.e., (i) from $t$ up to $t + \Delta t$ and (ii) from $t + \Delta t$ up to $t_f$. Thus:

$$\begin{aligned} J^0(\mathbf{x}, t) &= \min_{\mathbf{u}:\tau\in[t,t_f]} \left\{ \int_t^{t+\Delta t} L(\tau)\mathrm{d}\tau + \int_{t+\Delta t}^{t_f} L(\tau)\mathrm{d}\tau \right\} \\ &= \min_{\mathbf{u}(t)} \left\{ \int_t^{t+\Delta t} L\mathrm{d}\tau + J^0(\mathbf{x} + \Delta\mathbf{x}, t + \Delta t) \right\} \end{aligned}$$

Taylor series expansion of $J^0(\mathbf{x} + \Delta\mathbf{x}, t + \Delta t)$ gives:

$$J^0(\mathbf{x}+\Delta\mathbf{x}, t+\Delta t) = J^0(\mathbf{x}, t) + \frac{\partial J^{0T}}{\partial \mathbf{x}}\Delta\mathbf{x} + \frac{\partial J^0}{\partial t}\Delta t + 0(\Delta)$$

where $0(\Delta)$ represents the higher order terms. Therefore, the previous equation gives:

$$0 = \min_{\mathbf{u}(t)}\left\{L\Delta t + \left(\partial J^{0T}/\partial \mathbf{x}\right)\Delta\mathbf{x} + \left(\partial J^0/\partial t\right)\Delta t + 0(\Delta)\right\}$$

which divided by $\Delta t$, at the limit $\Delta t \to 0$ gives:

$$-\frac{\partial J^0}{\partial t} = \min_{\mathbf{u}(t)} H(\mathbf{x}, \mathbf{u}, t),$$

where $H(\mathbf{x}, \mathbf{u}, t)$ is the system's *Hamiltonian*:

$$H = L(\mathbf{x}, \mathbf{u}, t) + \frac{\partial J^{0T}(\mathbf{x}, t)}{\partial \mathbf{x}} \cdot \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = L(\mathbf{x}, \mathbf{u}, t) + \frac{\partial J^{0T}(\mathbf{x}, t)}{\partial \mathbf{x}}\mathbf{f}(\mathbf{x}, \mathbf{u}, t)$$

The above equation is known as the *Hamilton-Jacobi-Bellman* **(H-J-B)** equation. Its solution gives both the control $\mathbf{u}(t) : t \in \left[t_0, t_f\right]$ and the optimal cost $J^0(\mathbf{x}, t)$. In general, the solution of the H-J-B equation can be found by computational algorithms. But, if the system is linear and the cost function quadratic, the H-J-B is reduced to a *Riccati matrix-differential equation* as will be shown next [12, 45].

### 7.6.3 Linear Systems with Quadratic Cost

Consider, the case of a linear system $\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x} + \mathbf{B}(t)\mathbf{u}, \mathbf{x}(t_0) = \mathbf{x}_0$, with quadratic cost function:

$$J = \int_{t_0}^{t_f} L\mathrm{d}t, L = \frac{1}{2}\mathbf{x}^T\mathbf{Q}(t)\mathbf{x} + \frac{1}{2}\mathbf{u}^T\mathbf{R}(t)\mathbf{u}$$

Then, the optimal cost is also quadratic:

$$J^0(\mathbf{x}, t) = \frac{1}{2}\mathbf{x}^T(t)\mathbf{P}(t)\mathbf{x}(t)$$

where $\mathbf{P}(t)$ is a symmetric positive-definite matrix. The Hamiltonian is:

$$H = \frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} + \frac{1}{2}\mathbf{u}^T\mathbf{R}\mathbf{u} + \frac{1}{2}\mathbf{x}^T\mathbf{P}(\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}) + \frac{1}{2}(\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u})^T\mathbf{P}\mathbf{x}$$

Now, equating to zero the derivative $\partial H/\partial \mathbf{u}$, i.e., $\partial H/\partial \mathbf{u} = R\mathbf{u} + \mathbf{B}^\mathbf{T}\mathbf{P}\mathbf{x} = \mathbf{0}$, gives:

$$\mathbf{u}^0(t) = -\mathbf{R}^{-1}(t)\mathbf{B}^T(t)\mathbf{P}(t)\mathbf{x}(t)$$

Thus, the optimal Hamiltonian $H^0 = \min_{\mathbf{u}(t)} H$ is found to be:

$$H^0 = \min_{\mathbf{u}(t)} H = \frac{1}{2}\mathbf{x}^T\left(\mathbf{Q} + \mathbf{A}^T\mathbf{P} + \mathbf{P}\mathbf{A} - \mathbf{P}\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^T\mathbf{P}\right)\mathbf{x}$$

and therefore using the derivative:

$\partial J^0/\partial t = \frac{1}{2}\mathbf{x}^T[\mathrm{d}\mathbf{P}(t)/\mathrm{d}t]\mathbf{x}$, in the H-J-B equation, we obtain the Riccati equation for $\mathbf{P}(t)$ [8–12]:

$$-\mathrm{d}\mathbf{P}(t)/\mathrm{d}t = \mathbf{A}^T\mathbf{P}(t) + \mathbf{P}(t)\mathbf{A} + \mathbf{Q} - \mathbf{P}(t)\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^T\mathbf{P}(t)$$

The required boundary condition is found from the relationship

$$J^0(\mathbf{x}, t_f) = (1/2)\mathbf{x}(t_f)\mathbf{P}(t_f)\mathbf{x}(t_f) = 0$$

and is equal to:

$$\mathbf{P}(t_f) = 0$$

Solving the Riccati equation in reverse time, from $t_f$ to $t_0$, we find $\mathbf{P}(t) : t \in [t_0, t_f]$, which is then used to compute $\mathbf{u}^0(t)$ in forward time. The overall optimal closed-loop system has the flowchart shown in Fig. 7.6.

The linear optimal controller of Fig. 7.6 has found numerous applications in aircraft systems, industrial systems, and other modern industrial systems. Actually, it is one of the cornerstones of modern control systems.

### 7.6.4 *Pontryagin Minimum Principle*

This principle (originally formulated by Pontryagin as *maximum principle*) is applicable to the cases in which the control signal is subject to constraints specified by a lower and an upper bound. By definition, the optimal-control signal $\mathbf{u}^0$ corresponds to a local minimum of the cost function if:

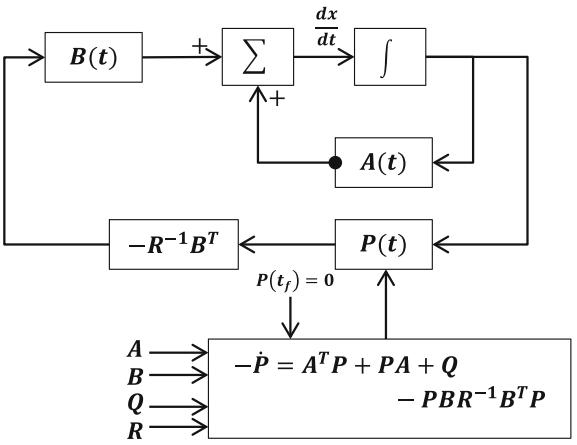



**Fig. 7.6** Closed-loop optimal-control system

$$J(\mathbf{u}) - J\left(\mathbf{u}^0\right) = \Delta J \geq 0$$

for all the allowable signals $\mathbf{u}$ close to $\mathbf{u}^0$.

Let $\mathbf{u} = \mathbf{u}^0 + \delta\mathbf{u}$. Then:

$$\Delta J\left(\mathbf{u}^0, \delta\mathbf{u}\right) = \delta J\left(\mathbf{u}^0, \delta\mathbf{u}\right) + (\text{higher} - \text{order terms})$$

The variation $\delta J$ is a linear function of $\delta\mathbf{u}$, and the higher order terms tend to zero for $\|\delta\mathbf{u}\| \to 0$. If the control signal is free (i.e., if it is not subject to some constraint), then the control variation $\delta\mathbf{u}$ can take any arbitrary value, and the necessary condition for $\mathbf{u}^0$ to minimize $J(\mathbf{u})$ is:

$\delta J(\mathbf{u}^0, \delta\mathbf{u}) = 0$, for $\|\boldsymbol{\delta}\|$ sufficiently small.

But, if the control signal is subject to constraints, the control variation $\delta\mathbf{u}$ is arbitrary only if the total $\mathbf{u}$ lies in the interior of the permissible control region for all $t \in \left[t_0, t_f\right]$. As long as this is valid, the constraints do not have any effect in the solution of the problem. However, if $\mathbf{u}$ lies at the boundary of the allowable region, at least for some time instants $t \in [t_1, t_2], t_0 \leq t_1 \leq t_2 \leq t_f$, then there exist allowable variations $\delta\mathbf{u}$ for which their opposite variations $-\delta\mathbf{u}$ are not allowable. If we consider only these variations, the necessary condition for $\mathbf{u}^0$ to minimize $J$ is:

$$\delta J\left(\mathbf{u}^0, \delta\hat{\mathbf{u}}\right) \geq 0$$

Therefore, the necessary condition for $\mathbf{u}^0$ to minimize $J$ is:

$$\delta J\left(\mathbf{u}^0, \delta\mathbf{u}\right) \geq 0$$

where $\|\delta\mathbf{u}\|$ is sufficiently small such that the sign of $\Delta J$ to be specified by the sign of $\Delta J$, and the signal $\mathbf{u} = \mathbf{u}^0 + \delta\mathbf{u}$ is *allowable* (i.e., it does not go outside the boundary of the allowable control region). This is the *minimum principle of Pontryagin* (see [5, 12]).

## 7.6.5 Stochastic Optimal Control

### 7.6.5.1 The Gauss–Markov Model

The continuous-time Gauss–Markov model has the following state-space equations (see [17, 18, 48]):

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \Gamma(t)\mathbf{w}(t), \mathbf{y}(t) = \mathbf{C}(t)\mathbf{x}(t) + \mathbf{v}(t), \mathbf{x}(t_0) = \mathbf{x}_0, t \geq t_0,$$

where all stochastic signals (processes) $\mathbf{w}(t), \mathbf{v}(t)$ and $\mathbf{x}(t_0)$ have Gaussian distributions with the following properties ($E$ is the expectation/averaging operator):

$$\begin{aligned}
&E[\mathbf{w}(t)] = \bar{\mathbf{w}}(t), t \geq t_0 \\
&E\{[\mathbf{w}(t) - \bar{\mathbf{w}}(t)][\mathbf{w}(\tau) - \bar{\mathbf{w}}(\tau)]^T\} = \mathbf{Q}(t)\delta(t-\tau); \mathbf{Q}(t) \geq 0, t, \tau \geq t_0 \\
&E[\mathbf{v}(t)] = \bar{\mathbf{v}}(t), t \geq t_0 \\
&E\{[\mathbf{v}(t) - \bar{\mathbf{v}}(t)][\mathbf{v}(\tau) - \bar{\mathbf{v}}(\tau)]^T = \mathbf{R}(t)\delta(t-\tau), \mathbf{R}(t) > 0; t, \tau \geq t_0\} \\
&E[\mathbf{x}(t_0)] = \bar{\mathbf{x}}(t_0) \\
&E\{[\mathbf{x}(t_0) - \bar{\mathbf{x}}(t_0)][\mathbf{x}(t_0) - \bar{\mathbf{x}}(t_0)]^T\} = \Sigma(t_0), \Sigma(t_0) \geq 0 \\
&E\{[\mathbf{x}(t_0) - \bar{\mathbf{x}}(t_0)][\mathbf{w}(t) - \bar{\mathbf{w}}(t)]^T\} = \mathbf{0}, t \geq t_0
\end{aligned}$$

inwhere $\mathbf{S}(t) = \mathbf{0}, t \geq t_0$ (and in some cases $\mathbf{S}(t) \neq 0, t \geq t_0$), and $\delta(t)$ is the *Dirac* (*impulse*) *function*. The expressions $\mathbf{Q}(t)\delta(t-\tau)$ and $\mathbf{R}(t)\delta(t-\tau)$ for the covariances of $\mathbf{w}(t)$ and $\mathbf{v}(t)$ indicate that $\mathbf{w}(t)$ and $\mathbf{v}(t)$ are white processes. The above Gauss–Markov model is diagrammed Fig. 7.7, which is analogous to Fig. 7.1.

Analogous equations describe the discrete-time Gauss–Markov model ($k = 0, 1, 2, \ldots$):

$$\mathbf{x}(k+1) = \mathbf{A}(k)\mathbf{x}(k) + \mathbf{\Gamma}(k)\mathbf{w}(k), \mathbf{y}(k) = \mathbf{C}(k)\mathbf{x}(k) + \mathbf{v}(k)$$

with matrices $\mathbf{Q}(k), \mathbf{R}(k), \mathbf{\Sigma}(0), \mathbf{S}(k)$, and $\delta(t)$ being replaced by the Kronecker delta $\delta_{jk} = 1$ for $j = k$ and $\delta_{jk} = 0$ for $j \neq k (j, k = 0, 1, 2, \ldots)$.

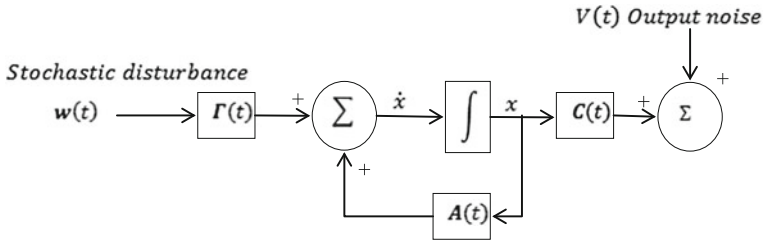



**Fig. 7.7** Pictorial representation of Gauss–Markov model

### 7.6.5.2 The Kalman–Bucy Filter

The Kalman–Bucy filter is the solution to the following state estimation problem [18, 44].

**Given** the matrices $\mathbf{A}(t), \mathbf{\Gamma}(t), \mathbf{C}(t), \mathbf{Q}(t), \mathbf{R}(t)$ and $\mathbf{\Sigma}(t_0)$ of the Gauss–Markov model and the measured output from $t = t_0$ up to the current time $t = t$, i.e., $\{y(\tau), t_0 \leq \tau \leq t\}$.

**Find** an optimal estimate $\mathbf{x}(t)$ of the current state $\mathbf{x}(t)$ using a suitable optimization criterion.

The available estimation criteria are the following:

(i) Minimization of the mean-square error
(ii) Maximization of the a *posteriori* probability-density function $p(\mathbf{x}(t)/\mathbf{Y}_t)$
(iii) Maximization of the likelihood function:

$L(\mathbf{Y}_t, \mathbf{x}) = \ln[p(\mathbf{Y}_t|\mathbf{x})]$ of the output measurement from $t_0$ to $t$.

If the stochastic system is governed by the above Gauss–Markov model, or use is made of statistics up to second-order, all the above criteria yield the optimal estimate $\hat{\mathbf{x}}(t)$ as the output of the so-called *Kalman–Bucy filter*, which is described by the following equations [18, 44]:

$$\begin{aligned}
\dot{\hat{\mathbf{x}}}(t|t) &= \mathbf{A}(t)\hat{\mathbf{x}}(t|t) + \mathbf{K}(t)[\mathbf{y}(t) - \mathbf{C}(t)\hat{\mathbf{x}}(t|t)], \hat{\mathbf{x}}(t_0|t_0) = \bar{\mathbf{x}}(t_0) \\
\mathbf{K}(t) &= \mathbf{\Sigma}(t|t)\mathbf{C}^T(t)\mathbf{R}^{-1}(t) \\
\dot{\mathbf{\Sigma}}(t|t) &= \mathbf{A}(t)\mathbf{\Sigma}(t|t) + \mathbf{\Sigma}(t|t)\mathbf{A}^T(t) - \mathbf{\Sigma}(t|t)\mathbf{C}^T(t)\mathbf{R}^{-1}(t)\mathbf{C}(t)\mathbf{\Sigma}(t|t) \\
&\quad + \mathbf{\Gamma}(t)\mathbf{Q}(t)\mathbf{\Gamma}^T(t), \mathbf{\Sigma}(t_0|t_0) = \mathbf{\Sigma}_0
\end{aligned}$$

where $\mathbf{\Sigma}(t|t)$, the covariance matrix of the error $\tilde{\mathbf{x}}(t|t) = \mathbf{x}(t) - \hat{\mathbf{x}}(t|t)$, is described by a Riccati equation. We observe that the Kalman–Bucy filter is composed of a copy of the system model with a correction (feedback) term equal to

$\mathbf{K}(t)[\mathbf{y}(t) - \mathbf{C}(t)\hat{\mathbf{x}}(t|t)] = \mathbf{K}(t)\tilde{\mathbf{y}}(t|t)$. Thus, the filter has diagram in Fig. 7.8.

An analogous form has the discrete-time-optimal filter:

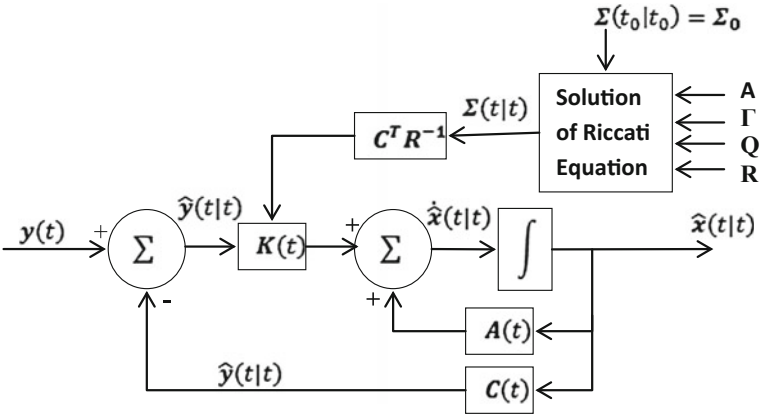



**Fig. 7.8** The Kalman–Bucy filter receives the measurement $\mathbf{y}(t)$ and provides the estimate $\hat{\mathbf{x}}(t|t)$

$$\hat{\mathbf{x}}(k+1|k+1) = \mathbf{A}(k)\hat{\mathbf{x}}(k|k) + \mathbf{K}(k+1)[\mathbf{y}(k+1) - \mathbf{C}(k+1)\mathbf{A}(k)\hat{\mathbf{x}}(k|k)], \hat{\mathbf{x}}(0|0) = \bar{\mathbf{x}}(0)$$
$$\mathbf{K}(k+1) = \sum (k+1|k)\mathbf{C}^T(k+1)\left[\mathbf{C}(k+1)\sum (k+1|k)\mathbf{C}^T(k+1) + \mathbf{R}(k+1)\right]^{-1}$$
$$\sum (k+1|k+1) = \sum (k+1|k) - \mathbf{K}(k+1)\mathbf{C}(k+1)\sum (k+1|k)$$
$$\sum (k+1|k) = \mathbf{A}(k)\sum (k|k)\mathbf{A}^T(k) + \mathbf{\Gamma}(k)\mathbf{Q}(k)\mathbf{\Gamma}^T(k), \sum (0|0) = \mathbf{\Sigma}_0$$

where

$$\sum (k|k) = E\left[\tilde{\mathbf{x}}(k|k)\tilde{\mathbf{x}}^T(k|k)\right]$$
$$\sum (k+1|k) - E\left[\tilde{\mathbf{x}}(k+1|k)\tilde{\mathbf{x}}^T(k+1|k)\right], \tilde{\mathbf{x}}(k+1|k) = \mathbf{x}(k+1) - \hat{\mathbf{x}}(k+1|k)$$

#### 7.6.5.3 Optimal Linear-Quadratic Gaussian Control

Now, the optimal control problem will be considered for a Gauss–Markov system where a control term **Bu** is also added in the state equation, i.e.,:

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t) + \mathbf{\Gamma}(t)\mathbf{w}(t), \mathbf{x}(t_0) = \mathbf{x}_0$$

The cost function to be minimized is

$$J = E\left[\int_{t_0}^{t_f} L\mathrm{d}t + J_f\right]$$

where $E[\cdot]$ is the expectation operator, and

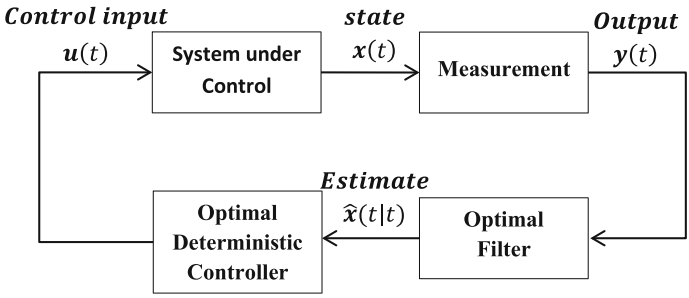



**Fig. 7.9** Block diagram illustration of the (linear) separation principle between estimation and control

$$L = \frac{1}{2}\mathbf{x}^T(t)\mathbf{Q}(t)\mathbf{x}(t) + \frac{1}{2}\mathbf{u}^T(t)\mathbf{R}(t)\mathbf{u}(t)$$
$$J_f = \frac{1}{2}\mathbf{x}^T(t_f)\mathbf{Q}_f\mathbf{x}(t_f)$$

Again, $\mathbf{Q}(t)$ is a semi-positive definite-square matrix, and $\mathbf{R}(t)$ is positive definite. The optimal control law here has the form:

$$\hat{\mathbf{u}}^0(t) = -\mathbf{R}^{-1}(t)\mathbf{B}^T(t)\mathbf{P}(t)\hat{\mathbf{x}}(t|t), t \geq t_0,$$

where $\hat{\mathbf{x}}(t|t)$ is the optimal state estimate of $\mathbf{x}(t)$ provided by the Kalman–Bucy filter, and control is the solution of the Riccati equation. The fact that the replacement of $\mathbf{x}(t)$ by $\hat{\mathbf{x}}(t|t)$ in this law leads to overall optimality is known as the **separation principle** between linear estimation and control. That is, the *optimal stochastic controller* is found by combining the *deterministic controller* (i.e., $\mathbf{P}(t)$ provided by the *Riccati control equation*) and the *stochastic state estimator* (i.e., the optimal Kalman–Bucy state estimate $\hat{\mathbf{x}}(t|t)$ as shown in Fig. 7.9. It is noted this separation principle does not in general hold when the system is nonlinear and/or the cost functional is not quadratic [18].

## 7.7 Adaptive and Predictive Control

### 7.7.1 General Issues

*Adaptive control* involves always an *on-line parameter estimator* and a *standard controller*. Depending on the method of estimation and control design, adaptive control is divided into one these types:

- Model-reference adaptive control **(MRAC)**
- Self-tuning control (**STC**)

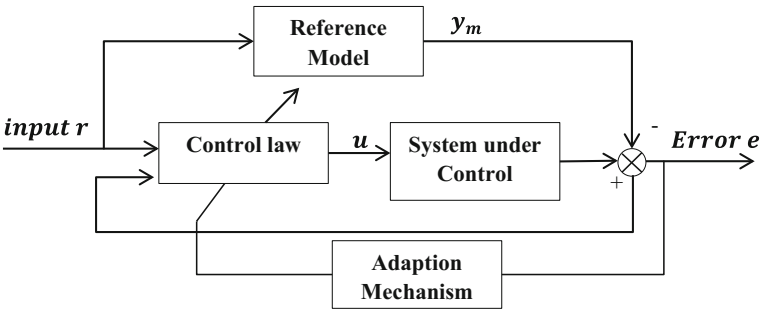
- Gain-scheduling control (**GSC**)
- Multiple-model adaptive control (**MMAC**)

Essentially, adaptive control involves the modification of the standard control law used to face the fact that the systems under control have uncertain or slowly-varying parameters (e.g., the mass of an aircraft is slowly decreasing due to the fuel consumption, etc.). Adaptive control does not require the *a-priori* knowledge (or information) about the bounds on these uncertain or varying parameters, in contrast to robust control, where such bounds must be available [19–22].

*Model-Predictive control* (also called *model-based predictive control*) uses an internal model for the simulation of the future behavior of the system and a reference trajectory for a smooth transition of the system from its present output to the desired output within a to ensure an optimal transition of the system's output much nearer to the desired reference trajectory. Model-based predictive control (**MBPC**) has been a popular type of control in chemical and physical process industries since the 1980s. The models are used to balance the impact of nonlinearities of variables and the jumps caused by noncoherent process devolution. If the use of linear models is not successful because of strong process nonlinearities, some kind of nonlinear MBPC must be tried [32].

### 7.7.2 Model-Reference Adaptive Control

*Model-reference adaptive control* (**MRAC**) was formulated by Landau. In MRAC, the parameter-estimation (or parameter adaptation) mechanism searches to find the parameter values so that the system response under MRAC follows (matches) the response of a given reference model. This means that the error between the



closed-loop system states and the reference model states (or responses) must tend asymptotically to zero, i.e., the error dynamics must be asymptotically stable, a fact that is assured by using Lyapunov's criterion. The general structure of **MRAC** has the form shown in Fig. 7.10 (see [21, 22]).

Suppose that we have the system:

$$\dot{\mathbf{y}} = \mathbf{A}(\mathbf{e}, t)\mathbf{y} + \mathbf{B}(\mathbf{e}, t)\mathbf{u}, \mathbf{y} \in R^n, \mathbf{u} \in R^p$$

and the reference model is:

$$\dot{\mathbf{x}} = \mathbf{A}_m\mathbf{x} + \mathbf{B}_m\mathbf{u}, \mathbf{x} \in R^n, \mathbf{u} \in R^p$$

here $R^n$ and $R^p$ denote the n-dimensional and p-dimensional Euclidean spaces, respectively. Note that the state vectors **y** and **x** must have the same dimensionality, and the control vector is applied to both the system and the reference model. To derive the adaptation mechanism, we start with the generalized state vector arbitrarily governed by the following dynamic (state-space) equation:

$$\dot{\mathbf{e}} = \mathbf{A}_m\mathbf{e} + [\mathbf{A}_m - \mathbf{A}(\mathbf{e}, t)]\mathbf{y} + [\mathbf{B}_m - \mathbf{B}(\mathbf{e}, t)]\mathbf{u}$$

Then, we define the following candidate Lyapunov function so:

$$V = \mathbf{e}^T\mathbf{P}\mathbf{e} + \text{trace}\left\{[\mathbf{A}_m - \mathbf{A}(\mathbf{e}, t)]^T\mathbf{F}_\mathbf{A}^{-1}[\mathbf{A}_m - \mathbf{A}(\mathbf{e}, t)]\right\} + \text{trace}\left\{[\mathbf{B}_m - \mathbf{B}(\mathbf{e}, t)]^T\mathbf{F}_\mathbf{B}^{-1}[\mathbf{B}_m - \mathbf{B}(\mathbf{e}, t)]\right\}$$

where $\mathbf{P}, \mathbf{F}_\mathbf{A}^{-1}$ and $\mathbf{F}_\mathbf{B}^{-1}$ are positive definite matrixes to be defined below. Computing the total derivative $\dot{V}$ of the Lyapunov function, it can be easily shown that the parameter adaptation laws

$$\dot{\mathbf{A}}(\mathbf{e}, t) = \mathbf{F}_\mathbf{A}(\mathbf{P}\mathbf{e})\mathbf{y}^T, \dot{\mathbf{B}}(\mathbf{e}, t) = \mathbf{F}_B(\mathbf{P}\mathbf{e})\mathbf{u}^T$$

assure that $\lim_{e\to\infty} \mathbf{e}(t) = \mathbf{0}$ when

$$\mathbf{A}_m^T\mathbf{P} + \mathbf{P}\mathbf{A}_m = -\mathbf{Q},$$

where **Q** is an arbitrary positive definite matrix that allows the computation of a suitable matrix **P**. These adaptation laws (mechanisms) are used as shown in Fig. 7.12 in combination with a suitable deterministic controller. Extensions and variations of these laws have been applied in various industrial systems with great success.

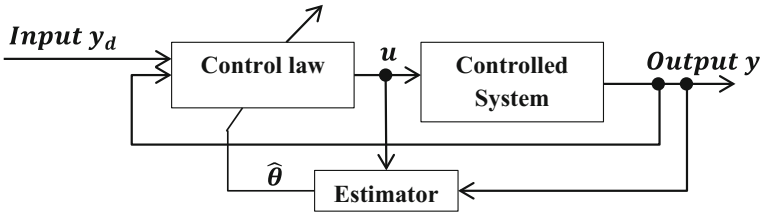



**Fig. 7.11** Architecture of STC where $\hat{\theta}$ denotes the estimated parameters

### 7.7.3 Self-tuning Control

The *self-tuning controller* (**STC**) is usually designed using a system model of the self/recursive form [19, 23]

$$A\left(z^{-1}\right)y(k) = B\left(z^{-1}\right)u(k-1) + h + e(k)$$

where $u(k)$ is the input, $y(k)$ the output, $e(k)$ a disturbance, and $h$ an unknown constant. Here $z^{-1}$ denotes the unit delay operator, $z^{-1}y(k) = y(k-1)$ and $A(z^{-1}), B(z^{-1})$ are polynomials of $z^{-1}$:

$$A\left(z^{-1}\right) = 1 + a_1 z^{-1} + \cdots + a_n z^{-n},\ B\left(z^{-1}\right) = b_1 z^{-1} + b_2 z^{-2} + \cdots + b_n z^{-n}$$

The general structure of a self-tuning control has the form shown in Fig. 7.11.

Typically, some variant of least-squares estimator is used combined with a minimum variance-type or PID controller [19, 23, 25]. The estimator is derived writing the system equation as:

$$y(k) = \mathbf{M}^T(k-1)\boldsymbol{\theta} + e(k)$$

where $\boldsymbol{\theta}$ is the vector of unknown parameters:

$$\boldsymbol{\theta} = [a_1, \ldots, a_n; b_1, \ldots, b_n, h]^T = [\theta_1, \theta_2, \ldots, \theta_{2n+1}]^T$$

and $\mathbf{M}(k-1)$ the vector of measurements:

$$\mathbf{M}(k-1) = [-y(k-1), \ldots, -y(k-n); u(k-1), \ldots, u(k-n); 1]^T,$$

the parameters must be selected that the estimated outputs $\hat{y}(k) = \mathbf{M}^T(k-1)\hat{\boldsymbol{\theta}}$ are as close to the measurements $y(k)$ as possible, in the least/squares sense. Omitting the derivation details, the estimate $\hat{\boldsymbol{\theta}}(k)$ can be computed recursively as:

$$\begin{aligned}
\boldsymbol{\theta}(k) &= \hat{\boldsymbol{\theta}}(k-1) + \mathbf{K}(k)\left[y(k) - \mathbf{M}^T(k-1)\hat{\theta}(k-1)\right] \\
\mathbf{K}(k) &= \sum(k)\mathbf{M}(k-1) \\
&= \sum(k-1)\mathbf{M}(k-1)/\left[1 + \mathbf{M}^T(k-1)\sum(k-1)\mathbf{M}(k-1)\right] \\
\sum(k) &= \sum(k-1) - \frac{\sum(k-1)\mathbf{M}(k-1)\mathbf{M}^T(k-1)\sum(k-1)\boldsymbol{\Sigma}(k-1)}{1 + \mathbf{M}^T(k-1)\sum(k-1)\mathbf{M}(k-1)} \\
&= \left[\mathbf{I} - \mathbf{K}(k)\mathbf{M}^T(k-1)\right]\sum(k-1).
\end{aligned}$$

This estimator is very good if the parameters $a_i, b_i, (i = 1, 2, \ldots, n)$ are constant,

$$y(k) = \mathbf{M}^T(k-1)\theta + e(k)$$

where $\boldsymbol{\theta}$ is the vector of unknown parameters

$$\boldsymbol{\theta} = [a_1, \ldots, a_n; b_1, \ldots, b_n, h]^T = [\theta_1, \theta_2, \ldots, \theta_{2n+1}]^T,$$

and $\mathbf{M}(k-1)$ the vector of measurements:

$$\mathbf{M}(k-1) = [-y(k-1), \ldots, -y(k-n); u(k-1), \ldots, u(k-n); 1]^T,$$

the parameters must be selected so that the estimated outputs $\hat{y}(k) = \mathbf{M}^T(k-1)\hat{\boldsymbol{\theta}}$ are as close to the measurements $y(k)$ as possible, in the least/squares sense. Omitting the derivation details, the estimate $\hat{\theta}(k)$ can be computed recursively as:

$$\begin{aligned}
\hat{\theta}(k) &= \hat{\theta}(k-1) + \mathbf{K}(k)\left[y(k) - \mathbf{M}^T(k-1)\hat{\theta}(k-1)\right] \\
\mathbf{K}(k) &= \sum(k)\mathbf{M}(k-1) \\
&= \sum(k-1)\mathbf{M}(k-1)/\left[1 + \mathbf{M}^T(k-1)\sum(k-1)\mathbf{M}(k-1)\right] \\
\sum(k) &= \sum(k-1) - \frac{\sum(k-1)\mathbf{M}(k-1)\mathbf{M}^T(k-1)\sum(k-1)}{1 + \mathbf{M}^T(k-1)\sum(k-1)\mathbf{M}(k-1)} \\
&= \left[\mathbf{I} - \mathbf{K}(k)\mathbf{M}^T(k-1)\right]\sum(k-1)
\end{aligned}$$

This estimator is good if the parameters $a_i, b_i (i = 1, 2, \ldots, n)$ are constant. If they slowly vary, then we use a data-forgetting factor $\mu (0 < \mu \leq 1)$, in which case the relations for $\mathbf{K}(k)$ and $\sum(k)$ must be replaced by:

$$\mathbf{K}(k) = \frac{\sum (k-1)\mathbf{M}(k-1)}{\mu + \mathbf{M}^T(k-1)\sum (k-1)\mathbf{M}(k-1)}$$

$$\sum (k) = \frac{1}{\mu}\left[\mathbf{I} - \mathbf{K}(k)\mathbf{M}^T(k-1)\right]\sum (k-1), \sum (0) = a\mathbf{I}$$

### 7.7.4 Gain-Scheduling Control

*Gain-scheduling control* is suitable for nonlinear plants and is applied in a variety of industrial systems. Typically, the dynamics of a physical or chemical process change when the operating conditions of the process change. This may be due to several reasons, e.g., to known nonlinearities. In these cases, it is possible to vary the controller parameters according to the monitored operating conditions of the system. This has motivated the development of *gain-scheduling control*. The structure of gain-scheduling control is as shown in Fig. 7.12.

A gain-scheduling controller is actually a special kind of nonlinear controller consisting of a linear controller whose parameters are changed (scheduled) as a function of the operating conditions in a programmed manner by the gain scheduler. General unified rules for designing gain-scheduling controllers are difficult to give, but typically the design of a **GSC** proceeds in the following steps:

Step 1: Several operating conditions are selected that cover the whole range of system's operation.
Step 2: For each of these operating conditions, a linear time-invariant approximation of the system is constructed and a linear controller (compensator) is designed for this linearized plant.
Step 3: In between operating points, the gains or parameters of the controllers are interpolated (scheduled) so that a global compensator is obtained that is applicable to the entire operation range.
Step 4: The loop-sampling period is specified, as well as any other design parameters that will remain constant despite the system nonlinearities or nonlinear behavior.

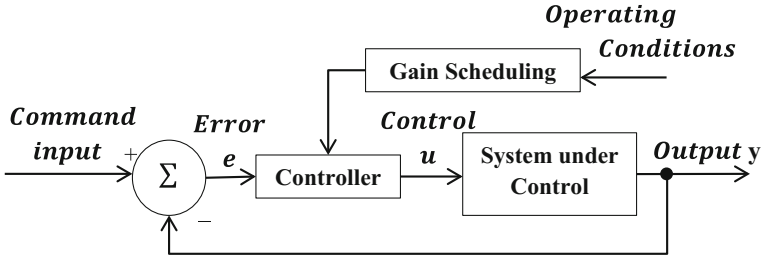



**Fig. 7.12** Structure of gain-scheduling control

The gain values are programmed as a look-up table (schedule) where the measured process variable indicates the current operating condition, and, as such, "directs" to the suitable controller-tuning values in the table at any time instant or sampling time. The controller (compensator) may be accordingly proportional (only), PI, PD, PID, or a pole-placement controller. Since the local designs are based on the linear time-invariant approximations, global stability cannot be secured beforehand. This can only be evaluated using computer simulations. Actually, one can design a look-up scheduling table along a reference trajectory on the basis of the system's measured variable(s).

### 7.7.5 *Model-Predictive Control*

Actually, there have been proposed several variants of **MPC** (or **MBPC**: Model-Based Predictive Control), but all of them generate the control inputs by optimizing online (in real-time) a certain criterion function of the process measurements and the outputs of an assumed process prediction model. Among the early **MPCs,** we mention the *model-predictive-heuristic control* **(MPHC)** technique of Richalet [24] and the *dynamic-matrix control* (**DMC**) technique of Ramaker [25]. MPC is suitable for industrial systems by taking into account the actuator limitations, enabling operation closer to the constraints (thus increasing the economic profit) and providing easy-to-tune schemes. The basic structure of MPC is as shown in Fig. 7.13.

The formulation of the fundamental MPC problem is as follows:

*Given* a finite-step response **(FSR)** model

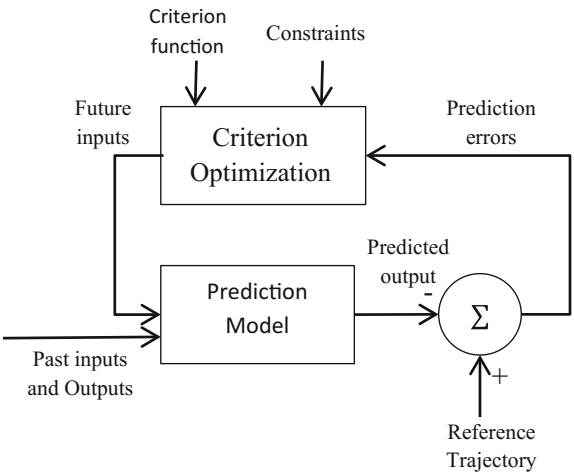



**Fig. 7.13** The basic structure of MPC

$$y(k) = \sum_{j=0}^{n-1} h_j u(k-j-1),$$

where $h_j$ is the impulse-response (or step-response) coefficient, $u(k-j-1)$ is the system input $j+1$ steps in the past, $y(k)$ is the system output at the current time $k$ and $n$ is the truncated order of the system, *minimize* the cost function:

$$J = E\left[\sum_{j=N_1}^{N_2} [\hat{y}(k+j|k) - r(k+j)]^2 + \sum_{j=1}^{N_u} [\gamma \Delta u(k+j-1)]^2 | k\right],$$

where $\hat{y}(k+j|k)$ is the predicted output $j$ steps ahead on the basis of data up to time $k$, $r(k+j)$ is the reference signal $j$ steps ahead, $\Delta u(k) = u(k) - u(k-1), \gamma$ is the control weighting coefficient, $N_1, N_2$ indicate the time horizon of $N_2 - N_1$ steps, and $N_u$ is the control horizon.

The above formulation refers to the **DMC** technique. Other techniques include:

- The generalized predictive control (**GPC**) [27]
- The extended horizon adaptive control (**EHAC**) [28, 29]
- The adaptive Predictive Control (**APC**) [30]

Details on the derivation of these controllers and their variants can be found in the MPC literature. If the problem involves operational or state or control constraints, the most common optimization technique employed is *quadratic programming* (**QP**). A good reference book composed of a set of papers on advanced model-based predictive control is [31]. A good book on nonlinear MPC is [32].

## 7.8 Robust Control

### *7.8.1 General Issues*

Among the modern control-design techniques, a dominant position is possessed by the techniques of *robust control*, which combine the stabilization of a system with the optimization of a "*norm*" of the transfer function. The term "*robust controller*" was coined by *Davison* in the 1970s; he formulated state-space controllers with asymptotic stability and/or of weakening a class of deterministic signals. In general, "*robustness*" of a system means "the ability of the system to maintain a desired performance in spite of the presence of uncertainties in the system". The simplest measure of robustness is the "*distance*" of the system from "*instability*". From a mathematical point of view, frequency-response methodology has provided the most effective set of analytical and computational tools for the study of "sensitivity," "robust stability,", and "robust performance". Actually, the techniques, which are based on the *spectrum norms* (singular values), generalize the concepts of **SISO**

stability margins to **MIMO** systems. These techniques include the techniques based on the norms $H_2, H_\infty$ and $l_1$, and the $\mu$-methodology.

**Robust stability** is concerned with the determination and evaluation of stability margins and the parameterization of stabilizing controllers for systems involving uncertainties.

**Robust optimal control** is concerned with the design of optimal controllers that maintain the internal stability of the system and minimize a certain weighted measure with weights the sensitivity matrices for the system.

**Robust performance** is concerned with the study and evaluation of the performance robustness under the influence of the uncertainties. The $H_2$ controllers are identical with the LQG controllers, the $H_\infty$ controllers were coined by *Zames* [33–36], the $l_1$-methodology is due to *Vidyasagar* [37, 38], and the $\mu$-methodology was developed by Doyle [41] (see also [39, 40]).

### 7.8.2 Non-stochastic Uncertainty Modeling

By the term *non-stochastic* (i.e., deterministic) *uncertainty*, we mean the uncertainty that is not modeled and studied using probabilities and random processes as was demonstrated in Sect. 7.6.5.1. In general, uncertainty in control systems appears in two forms:

*signal uncertainty* (i.e., uncertainty in the signals) and *system uncertainty* (i.e., uncertainty in the relation that transforms the input signals to output signals). The uncertainties that are due to the system's environment (external uncertainties) can be either *measured* (known) or *nonmeasured* (unknown). The *signal uncertainty* is modeled using the $L_2$—norm or the $L_\infty$—norm, i.e.:

$$W_2(c) = \left\{ \mathbf{w}(t) | \|\mathbf{w}(t)\|_2 = \left( \int_{-\infty}^{\infty} \mathbf{w}^T(t)\mathbf{w}(t)\mathrm{d}t \right)^{1/2} \leq c \right\}$$

and

$$W_\infty(c) = \left\{ \mathbf{w}(t) | \|\mathbf{w}(t)\|_\infty = \sup_t \max_i |w_i(t)| \leq c \right\}$$

where $W_2(c)$ denotes the permissible set of n-dimensional vector signals $\mathbf{w}(t)$ of which the energy (Euclidean norm) $\|\mathbf{w}(t)\|_2$ is less than a certain positive constant $c$, and $W_\infty(c)$ denotes the permissible set of signals with maximum amplitude $\|\mathbf{w}(t)\|_\infty$ less than $c$ with $w_i(t)$ being the ith component of $\mathbf{w}(t)$.

The *system uncertainty* may be one of the following:

- Nonstructured uncertainty
- Structured uncertainty

Examples of nonstructured uncertainty are: the unmodeled dynamic features of a system (e.g., high frequency modes), the varying time delays, the nonlinear effects, disturbances, etc.

### 7.8.3 Formulation of Robust Control Design

The general robust control problem is "to design the controller **K** in the system of Fig. 7.14 that maintains the norm (size) of the performance vector **z** small despite the exogenous signals **w**".

The classical *disturbance-rejection* problem falls within the framework this problem. The effect of **w** upon **z** is expressed by the transfer function $\mathbf{T}_{zw}(s)$ from **w** to **z**, which must have a small "norm". The same happens also for the problem of tracking a command signal, in which case **z** contains the tracking error. Therefore, the robust-control-design problem reduces to the problem of selecting the control signal to minimize the size (i.e., the norm) of the MIMO closed-loop-transfer function $\mathbf{T}_{zw}(s)$. The norm of a transfer-matrix function $\mathbf{T}_{zw}(s)$ is defined in the same way as the norm of a signal, i.e., by $H_2$ and $H_\infty$:

**Norm** $H_2$

$$\|T_{zw}\|_2 = \left\{ \frac{1}{2\pi} \int_{-\infty}^{\infty} trace\left[\mathbf{T}_{zw}(j\omega)\mathbf{T}_{zw}^*(j\omega)\right] \mathrm{d}\omega \right\}^{1/2}$$

$$= \left\{ \frac{1}{2\pi} \int_{-\infty}^{\infty} \sum_{i=1}^{r} \sigma_i^2(\mathbf{T}_{zw}(j\omega)) \mathrm{d}\omega \right\}^{1/2}$$

where $\sigma_i$ is the $i$th singular value of $\mathbf{T}_{zw}(j\omega)$, $\mathbf{T}_{zw}^*(j\omega)$, is the conjugate transpose of $\mathbf{T}_{zw}(j\omega)$ (the conjugate of $\mathbf{T}_{zw}^T(j\omega)$, and $r$ is the rank of the matrix $T(j\omega)$.

**Norm** $H_\infty$

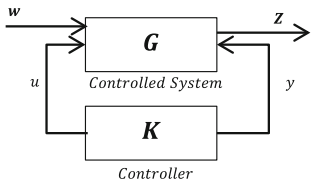



**Fig. 7.14** Typical configuration for robust controller design

$$\|\mathbf{T}_{zw}(s)\|_{\infty} = \sup_{w} \sigma_{\max}(\mathbf{T}_{zw}(j\omega))$$

where "sup" is the minimum upper bound of the function $\sigma_{\max}(\cdot)$. In practice, we can use the maximum ("max") operator in place of "sup". Here, $\sigma_{\max}(T(j\omega))$ is the *maximum singular value* of $\mathbf{T}(j\omega)$, i.e., the square root of the maximum eigenvalue of $\mathbf{T}^{*}(j\omega)\mathbf{T}(j\omega)$. Now, consider a system in state-space form: $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}, \mathbf{y} = \mathbf{C}\mathbf{x} + \mathbf{D}\mathbf{u}$, symbolized by $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$ or $\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}$

The transfer function matrix of this system is given by:

$$\mathbf{G}(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D}, \text{ Symbolically } \mathbf{G}(s) : \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}$$

The norm $H_{\infty}$ is suitable for the design of controllers that minimize $\|\mathbf{T}_{zw}\|_{\infty}$ to satisfy robustness criteria based on norm bounds, whereas $H_2$ is suitable for the minimization of $\|\mathbf{T}_{zw}\|_2$ when the disturbances $\mathbf{w}$ are stochastic as we have seen in the LQG control problem (Sect.7.6.5.3). Actually, the $H_2$ controllers are nothing more than the LQG controllers. Exactly similar $H_2$ and $H_{\infty}$ control designs are also valid in the case of discrete-time systems. The solution of the above robust control design can be found in the literature (e.g., [39–43]). **MATLAB** and other software packages provide suitable tools for implementing the solution(s) and checking robustness.

## 7.9 Nonlinear Control

The typical classes of nonlinear control are:

- Stabilizing and trajectory-tracking control via state-feedback linearization [49–51]
- Optimal-nonlinear control [45–47]
- Robust sliding-mode control [49–51].

### *7.9.1 State-Feedback Linearizing Control*

This methodology converts the nonlinear dynamics of a system into equivalent linear dynamics to make possible the application of linear control laws. This methodology is useful in practical nonlinear systems, such as fixed-wing aircraft, helicopters, industrial robots, and biomedical systems [49].

For simplicity, consider a special nonlinear system of the form:

$$\mathrm{d}x^n/\mathrm{d}t^n = f(\mathbf{x}) + b(\mathbf{x})u$$

where $u$ is a scalar input and $x$ a scalar output. The state vector of the system is $\mathbf{x} = \left[x_1, \dot{x}_2, \ldots, x^{(n-1)}\right]$ where $x^{(n-1)} = \mathrm{d}x^{n-1}/\mathrm{d}t^{n-1}$, and so the state-vector equation is:

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix} = \begin{bmatrix} x_2 \\ x_3 \\ \vdots \\ x_n \\ f(\mathbf{x}) + b(\mathbf{x})u \end{bmatrix}$$

which has the well-known controllable canonical form. Thus, if $\upsilon$ is the new control variable, choosing the control law as:

$$u = \frac{1}{b(\mathbf{x})}[\upsilon - f(\mathbf{x})], b(\mathbf{x}) \neq 0$$

the system reduces to:

$$x^{(n)} = \upsilon, x^{(n)} = \mathrm{d}x^n/\mathrm{d}t^n$$

and so the linear-control law now becomes:

$$\upsilon = -f_0 x - f_1 \dot{x} - \cdots - f_{n-1} x^{(n-1)},$$

where the feedback gains can be selected so all the roots of the characteristic polynomial (eigenvalues, poles) are placed on the left-hand s-semiplane and the system is *stable*. If the output $x(t)$ to *track a desired trajectory* $x_d(t)$, is desired, then the state-feedback control law must be selected as:

$$\upsilon = x_d^{(n)} - f_0 e - f_1 \dot{e} - \cdots - f_{n-1} e^{(n-1)}, e(t) = x(t) - x_d(t),$$

in which case the error tends to zero as $t$ goes to infinity. If the output is vectorial, then $\mathbf{b}(\mathbf{x})$ must be a matrix that is invertible in the state-space region of concern.

The above *input-state linearization* technique can be extended to the general nonlinear system:

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, u)$$

where $u$ is a scalar input and $\mathbf{x}$ an n-vector.

### 7.9.2 Optimal Nonlinear Control

The optimal control problem of nonlinear systems has been solved by the *calculus of variations,* the *minimum principle* of Pontryagin, and the *principle of optimality* of Bellman. The resulting equations are nonlinear and can be solved only by approximate computational techniques which are currently available in scientific or commercial software packages [45–47].

**Dynamic Programming Approach**

As described in Sect. 7.6.2, the solution of the optimal control problem of a nonlinear system: $\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}, \mathbf{u}, t), \mathbf{x}(t_0) = \mathbf{x}_0$ with cost functional: $J(\mathbf{u}) = \int_{t_0}^{t_f} L(\mathbf{x}, \mathbf{u}, t)\mathrm{d}t + I(\mathbf{x}_f)$ is given by the solution of the following *Hamilton-Jacobi-Bellman* **(H-J-B)** equation:

$$-\partial J^0/dt = \min_{\mathbf{u}(t)} H(\mathbf{x}, \mathbf{u}, t)$$

where

$$H(\mathbf{x}, \mathbf{u}, t) = L(\mathbf{x}, \mathbf{u}, t) + \frac{\partial J^{0T}(\mathbf{x}, t)}{\partial \mathbf{x}} f(\mathbf{x}, \mathbf{u})$$

is the *Hamiltonian* of the problem.

If the system is linear, i.e., $f(\mathbf{x}, \mathbf{u}, t) = \mathbf{Ax} + \mathbf{Bu}$ and the cost quadratic, i.e., $L = (\mathbf{x}^T\mathbf{Qx} + \mathbf{u}^T\mathbf{Ru})/2$, the solution of the **H-J-B** equation can be found in closed analytic form, i.e.:

$$\mathbf{u}^0(t) = -\mathbf{R}^{-1}(t)\mathbf{B}^T(t)\mathbf{P}(t)\mathbf{x}(t)$$

where the matrix $\mathbf{P}(t)$ is provided by the solution of the *Riccati equation*. In the general nonlinear case, the solution can be given by numerical (computational) algorithms of first or second order, by assuming the availability of a initial approximation $\bar{\mathbf{u}}(t), t_0 \le t \le t_f$ of the optimal control signal $\mathbf{u}^0(t), t_0 \le t \le t_f$.

### 7.9.3 Robust Nonlinear Sliding-Mode Control

Basic sliding-mode control was developed for the trajectory-tracking problem of a nonlinear system in the canonical form provided in Sect. 7.9.1 [49]:

$$d^n x/dt^n = f(\mathbf{x}) + b(\mathbf{x})u(t) + \delta(t), \mathbf{x}(0) = \mathbf{x}_0$$

where $\delta(t)$ is a disturbance input, $u(t)$ is the control input (e.g., the torque applied to a robotic join), $x(t)$ is the scalar output of concern, and $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T = \left[x, dx/dt, \ldots, d^{n-1}x/dt^{n-1}\right]^T$, is the state vector. The nonlinear function $f(\mathbf{x})$ is not known exactly but with some error (or precision) $|\Delta f|$ that is bounded from above by a known continuous function of $\mathbf{x}$. Similarly, the control-input gain $b(\mathbf{x})$ is not exactly known. We know its sign and an upper-bounding function. The problem under consideration is the following: *It is desired to find that* $u(t)$ *which drives the state on a desired trajectory* $\mathbf{x}_d = \left[x_d, dx_d/dt, \ldots, d^{n-1}x_d/dt^{n-1}\right]$ *in spite of the presence of the disturbance* $\delta(t)$ *and the fact that* $f(\mathbf{x})$ *and* $b(\mathbf{x})$ *are known with uncertainty.*

This problem will be first treated under the assumption that $\mathbf{x}_d(t=0) = \mathbf{x}(0) = \mathbf{x}_0$. The tracking error $\tilde{\mathbf{x}}(t)$ is $\tilde{\mathbf{x}}(t) = \mathbf{x}(t) - \mathbf{x}_d(t) = \left[\tilde{x}, d\tilde{x}/dt, \ldots, d^{n-1}\tilde{x}/dt^{n-1}\right]$. Defining a time-varying sliding surface $S(t)$ within the state-space $R^n$ as:

$$s(\mathbf{x}, t) = 0, s(\mathbf{x}, t) = (d/dt + \Omega)^{n-1}\tilde{\mathbf{x}}(t)$$

where $\Omega$ is a positive constant that represents the control-signal bandwidth, it can be shown that, to assure the trajectory tracking $\mathbf{x}(t) \to \mathbf{x}_d(t)$, we must maintain $s(\mathbf{x}, t) = 0$, which can be done if $u(t)$ is selected such that outside the surface $S(t)$ the following sliding condition holds:

$$\frac{1}{2}\frac{d}{dt}s^2(\mathbf{x}, t) \leq -\gamma|s|,$$

where $\gamma$ is a positive constant. This condition forces all trajectories to slide toward the surface $S(t)$, and, for this reason, the technique was named the "*sliding-mode*" technique (Fig. 7.15).

Essentially, this condition says that the function $s^2$ is, and continues to be, a Lyapunov function of the closed-loop system, despite the presence of the model uncertainty and the disturbance. This condition also assures that, if $\tilde{\mathbf{x}}(0) \neq \mathbf{0}$, again the trajectory $x(t)$ will arrive at $S(t)$ after a lapse of time less than or equal to

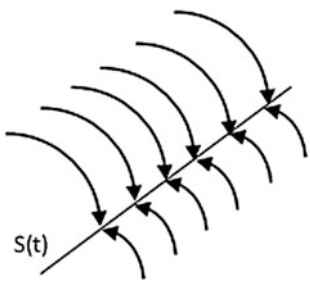



**Fig. 7.15** The sliding condition forces all trajectories to fall on the surface $S(t)$

$|s(0)|/\gamma$. Moreover, since $s(x,t) = 0$ is an $n$th order differential equation, once the trajectory arrives at $S(t)$ the tracking error will tend to zero asymptotically with the time constant $(n-1)/\Omega$. On the basis of this, the design of the sliding-mode robust controller involves two steps:

Step 1 We select a control that satisfies the sliding condition. This controller turns out to be discontinuous (a switching type) involving the sign function:

$$sign(s) = \begin{cases} +1, & s > 0 \\ -1, & s < 0 \end{cases}$$

Such a controller is not desired in practice because it might excite the non-modeled high- frequency dynamics of the system.

Step 2 In this step the switching-type controller is properly smoothened such to get a compromise between trajectory-tracking accuracy and control-signal bandwidth. This can be typically achieved by approximating the sharply varying "*sign*" function by a "*saturation*" function within a region $B(t)$:

$$B(t) = \{\mathbf{x} : |s(\mathbf{x};t)| \leq \phi\}, \phi > 0$$

of $\mathbf{x}$ where $\phi$ is a positive constant.

Step 3 Outside the boundary layer $B(t)$, the control law is defined as before, i.e., to satisfy the standard sliding-surface condition.

## 7.10 Intelligent Control

*Intelligent control* (**IC**) is the branch of automatic control in which the designed controllers mimic in one or the other way *biological control* [53, 54]. Practically, intelligent control is an enhancement of classical and traditional modern control with the following capabilities:

- Sensing
- Learning/recognition
- Reasoning under uncertainty
- Optimization
- Adaptability and flexibility
- Robustness/failure restoration
- Planning and decision making

Intelligent control is structured in a hierarchical (multilevel) way as shown in Fig. 7.16, so that the operations and tasks executed are given top-down-wise, i.e., from more abstract forms (valid for large horizons) to more exact data-rich forms (valid for smaller horizons). This means that intelligent control obeys the so-called "*principle of increasing precision with reducing intelligence*" coined by Saridis [52], which is applicable to engineering/technological, as well as social/managerial, systems.

Actually, intelligent control is based on the interaction of **AI** (Artificial Intelligence), **OR** (Operational Research), and **Control**, as shown in Fig. 7.17 [52–54].

The global goal of IC is to develop, implement, and use fully autonomous systems. According to *Albus*, intelligence is the integration of feedback into a sensory-interactive, goal-directed, autonomous, control system. Other architectures that have been proposed for the analysis and design of intelligent control systems include the following:

- *Meystel's* nested/multiresolutional architecture [55]
- *Arkin's* behavior-based architecture [56, 57]
- *Albus* reference model architecture [58]

*Biologically inspired intelligent control* involving multi-agents and possessing short-term memory for learning the environment, and long-term memory for behavior learning and task execution, is sometimes called *cognitive control*.

In general, the available methodologies for the design of intelligent controllers are:

- Model-based methods [19, 20, 25]
- Knowledge-based methods [59]

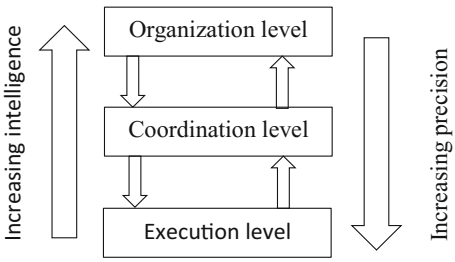



**Fig. 7.16** General hierarchical structure of intelligent control

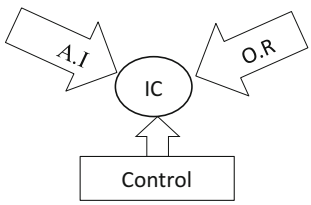



**Fig. 7.17** The three synergetic constituents of IC

- Fuzzy-logic methods [60–63]
- Neural-network methods [61–70]
- Hybrid neuro-fuzzy methods [71–81]
- Behavior-based methods [56–58]

In addition, a recently innovated class of optimization algorithms, which can be used in all the above methods, is the class of *genetic-evolutionary algorithms* discussed in Chap. 8 (Sect. 8.11).

**Model-based Methods** These methods assume the availability of a mathematical model of the system to be controlled; they produce intelligent controllers based on some kind of numerical learning and estimation combined with adaptation mechanisms. All adaptive controllers discussed in Sect. 7.7 belong to this class of intelligent controllers.

**Knowledge-based Methods** These methods use *symbolic* (*linguistic, nonnumerical*) *knowledge* and *models* of the systems under control, and they produce intelligent controllers structured in the same way as the knowledge-based (expert) systems discussed in Sect. 5.3.1.6 (Fig. 5.3). That is, a knowledge-based intelligent controller embeds control procedures in the inference engine and uses the relevant knowledge base. The controller is interacting with the user (control operator, control scientists) via the proper interface and explanation unit. The knowledge-based (or expert) controller is one of the following two kinds [59, 74]:

- Direct expert controllers
- Indirect expert controllers.

In direct expert control, the expert system plays the role of controller and controls the system directly as shown in Fig. 7.18.

In indirect expert control the expert system is used for the supervision (parameter tuning, etc.) of a conventional controller (classical or modern) as shown in Fig. 7.19.

**Fuzzy Logic Methods** The *fuzzy logic controllers* (**FLCs**) are based on the *fuzzy logic* and *fuzzy reasoning* and (like the neural controllers) do not need a mathematical model of the system under control. Thus, they are not described by mathematical (differential/algebraic) equations but by fuzzy (linguistic) rules. Fuzzy-logic controllers are categorized (like the expert controllers) as *direct* or *indirect fuzzy controllers*, depending on whether they control directly the system or they supervise and coordinate some traditional controller (e.g., a PID controller).

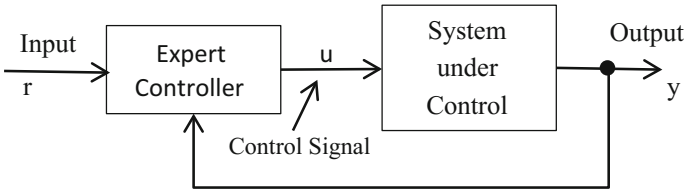



**Fig. 7.18** Structure of direct expert controller

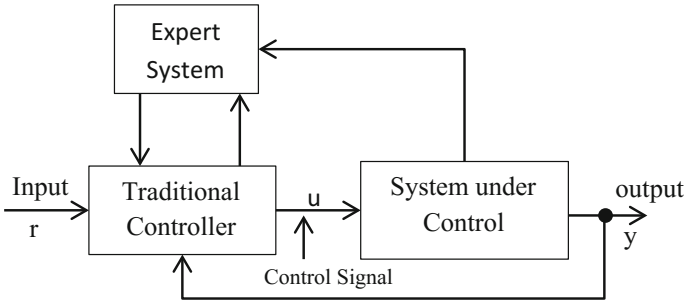



**Fig. 7.19** Structure of indirect expert controller

FLCs are formulated with the aid of linguistic expressions, and so they can mimic human operators [60–63]. The general structure of a *fuzzy system* (or *fuzzy-decision algorithm*) involves the following four units (Fig. 7.20a).

- A fuzzy rule base, i.e., a base of IF-THEN rules (FRB).
- A fuzzy-inference mechanism (FIM).
- An input-fuzzification unit (IFU).
- An output-defuzzification unit (ODU).

The *fuzzy-rule base* usually contains, besides the fuzzy or linguistic rules, a standard arithmetic database section. The fuzzy rules are provided by human experts or are derived through simulation. The *input-fuzzification unit* (fuzzifier) receives the non-fuzzy input values and converts them in fuzzy or linguistic form. The

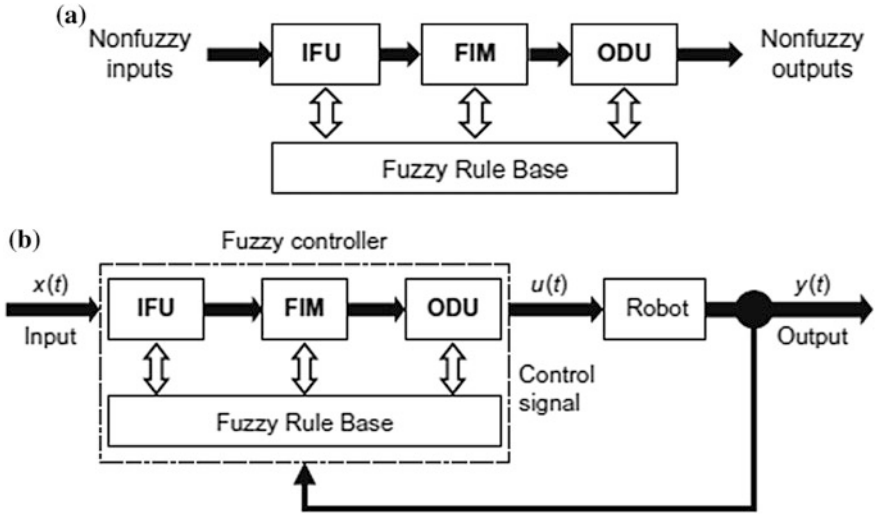



**Fig. 7.20** **a** General structure of a fuzzy system. **b** Basic fuzzy system control loop

*fuzzy-inference mechanism* is the core of the system and involves the fuzzy-inference logic (e.g., the max–min rule of Zadeh, etc.). Finally, the *output-defuzzification unit* converts the fuzzy results provided by FIM to non-fuzzy form using a defuzzification method. The basic fuzzy-control system loop is built using the fuzzy system of Fig. 7.20a, and has the form shown in Fig. 7.20b, where the fuzzy-rule base stores all the relevant knowledge, i.e., how to control the system (robot, etc.), and eliminates the need to have available an analytical mathematical model of the system.

The concept of *fuzzy set*, coined by *Zadeh* in 1965 [156], has broken the dichotomy of classical sets according to which an element $x$ belongs to a set $X$ or does not belong to $X$, symbolically $x \in X$ or $x \notin X$. Let $X = \{x_1, x_2, x_3, x_4, x_5\}$ be a classical set. The set $X$ is called the *reference superset*. Now, let $A = \{x_1, x_3, x_5\}$ be a classical subset of $X$. An equivalent representation of $A$ is:

$$A = \{(x_1, 1), (x_2, 0), (x_3, 1), (x_4, 0), (x_5, 1)\}$$

which is an ordered set of pairs $(x, \mu_A(x))$, where $x$ is the element of $x \in X$ of concern, and $\mu_A(x)$ is the membership of $x$ in the subset $A$, where:

$$\mu_A(x) = \left\{ \begin{array}{ll} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{array} \right\}$$

That is, here, we have $\mu_A : A \rightarrow \{0, 1\}$, where the set $\{0, 1\}$ has two elements, namely 0 and 1. If we allow the membership function $\mu_A(x)$ to be:

$$\mu_A : A \rightarrow [0, 1]$$

where [0, 1] is the fully closed interval between 0 and 1 (i.e., $0 \leq \mu_A(x) \leq 1$), then we have the fuzzy subset $A$ of $X$, defined as:

$$A = \{(x, \mu_A(x)) | x \in X, \mu_A(x) : X \rightarrow [0, 1]\}$$

**Neural-Network Methods** Intelligent controllers which are based on neural networks (briefly called *neuro-controllers*) have the following properties [64–70]:

- They learn by experience (not via modeling or programming)
- They can generalize (i.e., they can act successfully even if they don't know the situation at hand)
- They can find and provide arbitrary, continuous, nonlinear relations or functions (i.e., they are *universal approximators*)
- They have endogenously parallel and distributed structures.

Neuro-controllers can be sorted into three types:

- Direct neuro-controllers (the neural network controls directly the system)
- Indirect neurocontrollers (the neural network supervises a certain traditional controller)
- Neuro-controllers that use a neural model of the system.

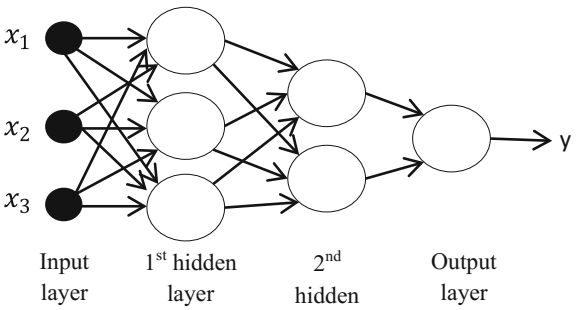



**Fig. 7.21** The multilayer perceptron

Actually, a neuro-controller realizes some form of adaptive control, where the controller is a neural network of a certain type, and the adaptation mechanism is based on the updating the weights in the neural network.

Two very popular types of neural networks (**NNs**) used for designing neuro controllers are the *multilayer perceptron* (**MLP**) and the *radial-basis functions* (**RBF**) networks.

**The Multilayer Perceptron** This NN has the structure shown in Fig. 7.21.

All neurons of the network contain a *sigmoid nonlinearity* of the *logistic* type:

$$y_k = f(v_k) = \frac{1}{1 + e^{-v_k}},$$

where $v_k = \sum_{j=1}^{p} w_{kj} y_j - \theta_k$ is the net internal activity of the neuron (node) $k$, $\theta_k$ is the threshold of the neuron $k$, and $y_k$ is the output of this neuron. The weights are adjusted using the *back-propagation* (*BP*) *algorithm* of learning, which is based on the weight updating formula:

$$w_{ij}(t+1) = w_{ij}(t) + \eta \delta_j(t) x_i(t),$$

where $w_{ij}$ is the weight connecting node $i$ with node $j$ of the next layer at time $t$. If node $j$ is a node of the output layer, then:

$$\delta_j = y_j(1 - y_j)(d_j - y_j)$$

otherwise:

$$\delta_j = x'_j\left(1 - x'_j\right) \sum_k \delta_k w_{jk},$$

where $k$ extends over all nodes of the previous layers.

**The Radial-Basis Function (RBF) Network** An RBF network approximates an input–output mapping by employing a linear combination of radially symmetric functions (see Fig. 7.22). The $k$th output $y_k$ is given by:

$$y_k(\mathbf{x}) = \sum_{i=1}^{m} w_{ki}\phi_i(\mathbf{x})$$

where:

$$\phi_i(\mathbf{x}) = \phi(\|\mathbf{x} - \mathbf{c}_i\|) = \phi(r_i) = \exp\left(-\frac{r_i^2}{2\sigma_i^2}\right), r_i \geq 0, \sigma_i \geq 0$$

is a Gaussian-like function.

The RBF networks have always one hidden layer of computational nodes with non- monotonic transfer functions $\phi(\cdot)$. Theoretical studies have shown that the choice of $\phi(\cdot)$ is not very crucial for the effectiveness of the network. In most cases, the Gaussian RBF given above is used, where $\mathbf{c}_i$ and $\sigma_i (i = 1, 2, \ldots, m)$ are selected centers and widths, respectively.

The *training procedure* of the RBF network involves the following steps:

Step 1: Group the training patterns in $\mathbf{M}$ subsets using some clustering algorithm (e.g., the $k$-means clustering algorithm) and select their centers $\mathbf{c}_i$.
Step 2: Select the widths, $\sigma_i (i = 1, 2, \ldots, m)$, using some heuristic method (e.g., the $p$ nearest-neighbor algorithm).
Step 3: Compute the RBF activation functions, $\phi_i(\mathbf{x})$, for the training inputs.
Step 4: Compute the weights by least squares. To this end, write the $k$th output relation as $\mathbf{b}_k = \mathbf{A}\mathbf{w}_k (k = 1, 2, \ldots, p)$ and solve for $\mathbf{w}_k$, i.e.,:

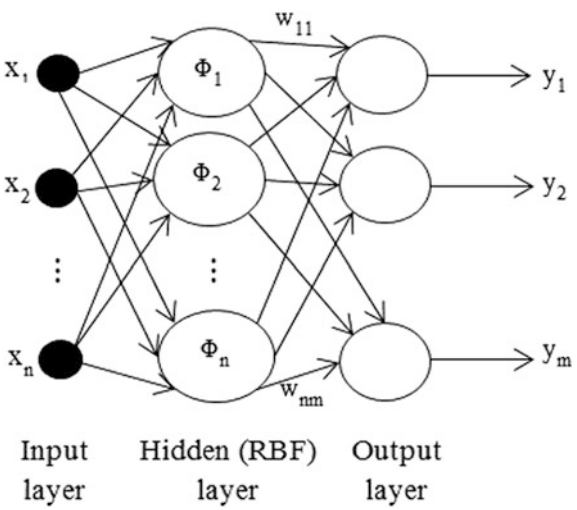



**Fig. 7.22** The radial-basis-function network

$$\mathbf{w}_k = \mathbf{A}^{\dagger}\mathbf{b}_k, \mathbf{w}_k = [w_{k1}, \ldots, w_{km}]^T,$$

where $\mathbf{A}^{\dagger}$ is the generalized inverse of $\mathbf{A}$ given by:

$$\mathbf{A}^{\dagger} = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T$$

and $\mathbf{b}_k$ is the vector of the training values for the output $k$.

The general classification of neuro-controllers is threefold, i.e., the same as that of NNs:

- Neuro-control with supervisory learning
- Neuro-control with non-supervisory learning
- Neuro-control with reinforcement learning

The most popular types are the neuro-controllers with supervisory and reinforcement learning. The basic structure of supervised learning controller is shown in Fig. 7.23.

The teacher trains the neuro-controller by presenting to it examples of successful control signals. The teacher may be an experienced human operator or any man-made traditional controller. Clearly, this type of neuro-controller cannot exceed the efficiency of the teacher.

**Hybrid Neuro-Fuzzy Methods** These methods are based on the general idea to design controllers that combine **FL** methods and **NN** methods in several ways. The FL component enables us to embed experiential knowledge, and the NN enables the system to learn. In many cases, a genetic algorithm is also involved, which makes it possible for the system to be self-optimized [71–81].

The pure fuzzy-logic systems have two main drawbacks:

- They do not have available a specific method for the determination of the membership functions.
- They do not have available a learning or adaptation component.

The above two drawbacks are overcome if we use neural nets for driving the fuzzy reasoning as shown in Fig. 7.24.

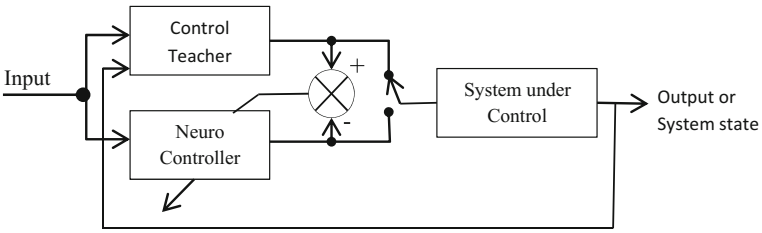



**Fig. 7.23** Basic structure of NN controller with supervisory learning

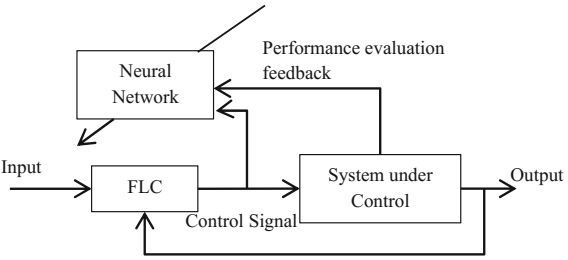



**Fig. 7.24** General form of hybrid neuro-fuzzy controller

Indeed the NNs can be trained so as to be able to select the membership functions (i.e., the fuzzy sets) in an autonomous way and for selecting the form and/or the number of the fuzzy rules. To this end, various techniques have been developed with relative variations in their generality, simplicity, and applicability. Four such systems platforms are the following:

- *ANFIS*: Adaptive-Neurofuzzy Inference System [75].
- *GARI*: Generalized Approximate Reasoning-based Intelligent-Control System [76].
- *ART-ARTMAC*: Adaptive-Resonance Technique [77–80].
- *FALCON*: Fuzzy Adaptive-Learning Control Network [71, 81].

**Behavior-based Methods** These methods are based on the concept of "agent", and couple tightly the sensing and action functions of intelligent control. They avoid the symbolic representation of knowledge, decomposing it into contextually meaningful units (behavior or situation-action units). Using the concept of agent, it is possible to describe and explain both the hierarchical and nested intelligent-control architectures.

## 7.11 Control of Further System Types

So far we have reviewed modern control techniques for the fundamental types of physical/technological systems, which are described by ordinary differential equations and traditional algebraic equations. These systems are characterized as *lumped-parameter systems* (**LPS**). In this section, a quick tour will be taken of the control of some further types of systems, namely:

- Large-scale lumped-parameter systems
- Distributed-parameter systems
- Systems with time delays
- Finite-state automata
- Discrete-event systems.

### 7.11.1 Control of Large-Scale Systems

The mathematical models of many physical and engineering systems are frequently of high dimension or possess interacting dynamic phenomena. The information processing and requirements for experimenting with these models for control purposes are usually excessive. It is therefore natural to seek techniques that reduce the computational effort. Approaches to and methodologies of large-scale systems provide such techniques through the manipulation of system structure in some way.

The essential differences between *large-scale systems* (**LSS**) in contrast to small-scale (conventional) systems can be summarized as follows:

- More than one controller or decision maker shares the planning, allocation, and management responsibility which results in decentralized computations.
- The information patterns and measurement records that are frequently made available to controllers have different but correlated forms.
- Despite the fact that each controller may have a specific task to fulfill, a certain degree of coordination among all controllers, or groups of them, seems to be the rule rather than the exception.

The requirements of distributed-computation effort, the large number of controllers involved, and the need for coordination among their operation have led quite naturally to *hierarchical* and *multilevel* structure.

The solution of nonlinear optimal-control problems leads ultimately to the solution of **TPBV** problems, which require successive approximation techniques for their solution [82]. To facilitate the implementation of such techniques in **LSSs**, a number of techniques have been developed based on the idea of *decomposition–coordination* [83, 84]. It has been argued that such techniques have computational advantages both from the point of view of achievable accuracy and the issue of demanding less memory and computation time.

The hierarchical optimization is based on the general idea to write the cost function under minimization in the form of a separable part and a non-separable part and to write the nonlinear dynamic equations in a form of a linear part that is separable by block and another term that contains the nonlinearities and the interaction terms.

The role of the higher level of the hierarchy is to fix the nonseparable part in the criterion function and the nonlinear part in the dynamic equation. This leaves a set of low-order dynamic-optimization problems to be solved at the lowest level in the hierarchy. The higher level could successively approximate the specified variables by their optimal values. A *multilevel control structure* can also be used for stabilizing LSSs. If the system is linear, then more concrete results can be obtained. Here, two approaches can be followed: the first one treats the interactions as perturbations, while the second considers the interactions as part of the original system, and tries to stabilize the system under structural perturbations. Filtering, stochastic, adaptive, and robust controllers of LSSs all can be designed following the multilevel approach [85–89].

Another class of methods developed for LSSs is the class of *decentralized-stabilization, control,* and *estimation methods* [90–95]. These methods require the system to be stabilizable, controllable, and observable in decentralized form. For example, consider an interconnected LSS of the form:

$$S : \dot{\mathbf{x}}_i = \mathbf{A}_i\mathbf{x}_i + \mathbf{B}_i\left(\mathbf{u}_i + \sum_{j=1}^{N}\mathbf{D}_{ij}\mathbf{x}_j\right), \mathbf{y}_i = \mathbf{x}_i$$

for $i = 1, 2, \ldots, N$, where $\mathbf{x}_i$ and $\mathbf{y}_i$ are the state vectors and output vectors, respectively, of the $i$th interconnected system, and $\mathbf{A}_i, \mathbf{B}_i$ and $\mathbf{D}_{ij}$ are matrices of appropriate dimensions.

If the decoupled systems $(\mathbf{A}_i, \mathbf{B}_i) : \dot{\mathbf{x}}_i = \mathbf{A}_i\mathbf{x}_i + \mathbf{B}_i\mathbf{u}_i$ are controllable, then we can easily establish that the system $S$ has no *frozen* (non-stabilizable) decentralized components. Thus, the system $S$ can be stabilized with dynamic state-feedback controllers of the type:

$$\dot{\mathbf{z}} = \mathbf{K}_D\mathbf{z} + \mathbf{G}_D\mathbf{y}, \mathbf{u} = -\mathbf{H}_D\mathbf{z} - \mathbf{F}_D\mathbf{y},$$

where $\mathbf{z} = \left[\mathbf{z}_1^T, \mathbf{z}_2^T, \ldots, \mathbf{z}_N^T\right]^T, \mathbf{y} = \left[\mathbf{y}_1^T, \mathbf{y}_2^T, \ldots, \mathbf{y}_N^T\right]^T$ and $\mathbf{u} = \left[\mathbf{u}_1^T, \mathbf{u}_2^T, \ldots, \mathbf{u}_N^T\right]^T$, with $\mathbf{u}$ being the control signal vector provided at the output of the controller. Since here $\mathbf{y}_i = \mathbf{x}_i$ (i.e., the subsystem outputs are their entire states), it is not necessary to use dynamic controllers. In the framework of LSS stability, the so-called *vector Lyapunov functions* have been developed.

Given an interconnected system of the type:

$$\dot{\mathbf{x}} = \mathbf{A}_i\mathbf{x}_i + \sum_{j=1}^{N}\mathbf{A}_{ij}\mathbf{x}_j + \mathbf{B}_i\mathbf{u}_i, \mathbf{A}_{ij} = \mathbf{B}_i\mathbf{D}_{ij}$$

we can derive a *decentralized-optimal controller*, i.e., a controller where the control vector of each subsystem uses only the local state vector, i.e.:

$$\mathbf{u}_i^0(t) = -\mathbf{F}_i(t)\mathbf{x}_i$$

where $\mathbf{F}_i(t)$ is determined by the local problem matrices, i.e.: $\mathbf{F}_i(t) = \mathbf{R}_i^{-1}\mathbf{B}_i^{-1}\mathbf{P}_i(t)$, with $\mathbf{P}_i(t)$ being the solution of the local Riccati equation. Analogous results are also obtained in the case of discrete-time systems.

### 7.11.2 Control of Distributed-Parameter Systems

Dynamic systems are divided into two main categories: *lumped-parameter* (*LPS*) and *distributed-parameter* (*DPS*) *systems*. Lumped-parameter systems are modeled

by ordinary differential or integrodifferential equations, whereas distributed-parameter systems are described by partial equations. Many physical systems are composed of a combination of a pure distributed and a pure lumped-parameter system. Such systems are known as *mixed-distributed and lumped-parameter systems* and are described by mixtures of ordinary and partial equations. Strictly speaking, all physical systems are distributed in nature (i.e., they occupy a certain spatial domain), but, if the wavelength is much larger than the size of the system, or equivalently if the energy spatial distribution is sufficiently concentrated, we regard them as lumped. In those situations where the above condition does not hold, for an adequate description, analysis, and design of the system, one has to resort to partial differential or integral or integrodifferential equation models.

Partial differential equations (or the distributed-parameter systems that are described by them) are classified in many ways, the most important of which are the following [96]:

- On the basis of how the response quantities of the system propagate in space and time, namely, "*diffusion-like*" and "*wave-like*" systems.
- On the basis of the dimensionality of the spatial domain occupied, that is, 1-D or M-D partial differential equations.
- Single-variable and multivariable systems.
- Fixed-domain systems—systems with time-invariant boundary surfaces—and variable-domain systems—systems with moving boundary surfaces.
- Finite-domain, semi-infinite-domain and infinite-domain systems; finite-domain systems are characterized by certain boundary conditions, whereas infinite- or semi-infinite-domain systems must satisfy certain limit conditions (e.g., the well-known radiation condition).
- Free or unforced (no external distributed or boundary inputs exist) and forced or controlled (from the interior or from the boundary of the spatial domain) systems.
- Distributed systems with homogenous media (characteristic parameters of the system are the same over the whole occupied spatial domain) or inhomogeneous media (the parameters are different in various parts of the domain).
- Time-invariant/varying or space invariant/varying distributed-parameter systems.

The equations modeling a given process can be derived by using basic conservation principles (energy, mass, momentum, monetary, for instance), along with some restrictive relations connecting their structural parameters. Much of classical and modern science and engineering has been concerned with the basic problem of determining the system parameters, such as heat transfer coefficients, chemical-rate constants, specific heats, elastic moduli, strain properties, electromagnetic properties, and so on. The dynamic models of DPS are analogous to LPS.

Consider a fixed spatial domain $D$ which is a simply connected, open subset of the $n$-dimensional Euclidean space $E^n$ with boundary surface $\partial D \in E^n$. Then, a

general linear state-space distributed-parameter model covering the majority of linear systems is:

$$\frac{\partial X(x,t)}{\partial t} = AX(x,t) + BU(x,t), (x,t) \in \Omega = D \times \left[t_0, t_f\right], X(x,t_0) = X_0(x), x \in D$$
$$Y(x,t) = CX(x,t), (x,t) \in \Omega = D \times \left[t_0, t_f\right]$$
$$A_b X(x,t) = B_b U_b(x,t), (x,t) \in \partial\Omega = \partial D \times \left[t_0, t_f\right]$$

where $x \in D$ is the spatial variable; $t \in \left[t_0, t_f\right]$ is the time variable; $X(x,t) \in \chi = L_m^2(\Omega)$ is an m-dimensional state vector; $Y(x,t) \in \mathsf{Y} = L_q^2(\Omega)$ is a q-dimensional output vector; $U(x,t) \in \mathsf{U} = L_p^2(\Omega)$ is a p-dimensional volume input vector; $U_b(x,t) \in \mathsf{U}_b = L_{pb}^2(\partial\Omega)$ is a $p_b$-dimensional boundary input vector; $A, A_b$ are linear matrix partial differential (or integral or integrodifferential) operators with respect to the spatial coordinates $x_1, x_2, \ldots, x_n$ of $x$; and $B, B_b, C$ are matrix functions or integral operators.

The order of the operator $A_b$ must be less than that of $A$. This guarantees the well-posedness of the model. A special case of this model, which has received much attention is:

$$A(x,t)\frac{\partial X(x,t)}{\partial t} = \sum_{j=1}^{n} A_j(x,t)\frac{\partial X(x,t)}{\partial x_j} + A_0(x,t)X(x,t) + B(x,t)U(x,t)$$

with the properties: (a) the matrix $A$ is nonsingular; and (b) for arbitrary values of the real parameters $\mu_1, \mu_2, \ldots, \mu_n$ the roots $\lambda$ of the characteristic polynomial $\left|\lambda A - \sum_{j=1}^{n} \mu_j A_j\right| = 0$ are all real-valued functions of $x \in D$ and have associated with them a full set of $m$ linearly independent and nontrivial characteristic vectors. Of particular importance is the special case where $D$ is a region of the real line, say [0, 1].

Concerning the nonlinear distributed-parameter systems, the most general dynamic model is:

$$X_t(x,t) = F[x,t;X,X_x,X_{xx},U(x,t)], x \in (0,1)$$
$$Y(x,t) = M[x,t,X(x,t)]; X(x,0) = X_0(x), x \in (0,1)$$
$$G_0[x,t,X_x,U_{b0}(t)] = 0, x = 0$$
$$G_1[x,t,X_x,U_{b1}(t)] = 0, x = 1$$

for $t \in \left[t_0, t_f\right]$, where $N$, $N_b$ and $C$ are nonlinear spatial operators and $F$, $G_0$, $G_1$ and $M$ are conventional functions of their arguments, sufficiently smooth; $U(x,t)$, $U_{b0}(t)$ and $U_{b1}(t)$ are known inputs.

The preceding partial-differential-equation models constitute the core of distributed-parameter system theory (simulation, stability, stabilization, state estimation, optimization, and control) but do not cover all cases. It is useful to mention

here that all physical systems are subject to random internal and external disturbances and noise. Hence, in many cases we are using not deterministic models but stochastic ones, involving additive or nonadditive stochastic terms, which are used to model the random fluctuations. A small representative subset of publications in the DPS control area can be found in [96–110].

Examples of DPSs are:

- *Mechanical systems* (e.g., a transverse vibrating slender beam)
- *Electrical systems* (e.g., a transmission line)
- *Magnetohydrodynamic systems* (plasma systems)
- *Chemical systems* (gas separation, distillation columns, etc.)
- *Geophysical systems* (e.g., petroleum reservoirs and subsurface aquifers)
- *Environmental systems* (e.g., water quality/estuary and stream systems)
- *Population systems* (e.g., population growth and diffusion, interacting population motion)

Distributed-parameter-systems (DPS) theory has been developed along the classical avenues of modeling, simulation, stability, controllability/observability, parameter identification, state estimation, optimal control, stochastic control, adaptive control, sensitivity, and computational methods. In addition to these problems, new areas strictly specific to DPS such as spatial sensor/actuator allocation, boundary condition identification, and the study of control/observation spillover, have been pursued. DPS estimation and control theory is at a very advanced level, with many results being applied in modern technological and nontechnological applications.

Some representatives of the most recent results include: boundary-condition estimation, DPS control from the boundary, simultaneous parameter and state estimation, optimal sensor and actuator allocation, adaptive control, stabilizing control, model-predictive control, robust control, nonlinear control, and intelligent control. All these types of control use the same core principles and methods developed for LPS, integrating and embedding in each case peculiarities due to the space dependence of the system. Details on them are provided in the literature on DPS which is still growing [96–110].

### *7.11.3 Control of Time-Delay Systems*

Real systems have a multitude of time delays typically due to material flow and information transmission. Physical and chemical processes, tele-operation systems, communication networks, and so on, are examples of systems with time delays. Time-delay systems are infinite dimensional like distributed-parameter systems, and so their efficient control needs special care and enhanced complicated techniques. A time-delay system can be one of the following:

- Linear or nonlinear
- Continuous-time or discrete time
- Deterministic or stochastic
- With constant or time-varying delays
- With input/output or internal (state) delays
- With single or multiple delays
- Lumped-parameter or distributed-parameter
- Hierarchical or decentralized

Controllability-observability, stability, optimality, adaptability, and robustness are all governed by the same principles as the non-delay systems, but the development and design of controllers is much more complex and needs more advanced mathematical methods.

The first powerful method for treating time-delay systems is the well-known *Smith predictor* developed in the 1950s, which was enhanced to be applicable to unstable time-delay systems in the late 1970s. Today, many adaptive- and robust-control schemes are available for treating time-delay systems in all the above categories.

Examples of existing time-delay system modern controllers include: decoupling controllers, stabilizing controllers, optimal controllers, adaptive controllers, stochastic controllers, and $H_\infty$ robust controller [111–121]. Typical examples of time-delay systems control are process control, tele-operator control, communication systems control, and control of technological or societal systems via the Internet.

In practice, several approximations of the transfer function

$$\bar{y}(s)/\bar{u}(s) = e^{-sT_d}$$

of a pure time-delay block $y(t) = u(t - T_d)$ at time $t$, where $t - T_d$ is the input time, are used. The simplest one is the linear approximation:

$$e^{-sTd} = 1 - T_d s$$

which is good only at low frequencies for a very small time delay $T_d$.

When the value of $T_d$ is relatively large, then the design of an efficient compensator of the time-delay system is very difficult. The two classical approaches to treat this problem are:

- Use of rational approximations of $e^{sT_d}$ (e.g., Padé approximation)
- Use of a suitable predictor, e.g., Smith predictor, that eliminates (or reduces) the effect of the time delay

The simplest Padé approximation is:

$$e^{-sT_d} \simeq (2 - T_d s)/(2 + T_d s)$$

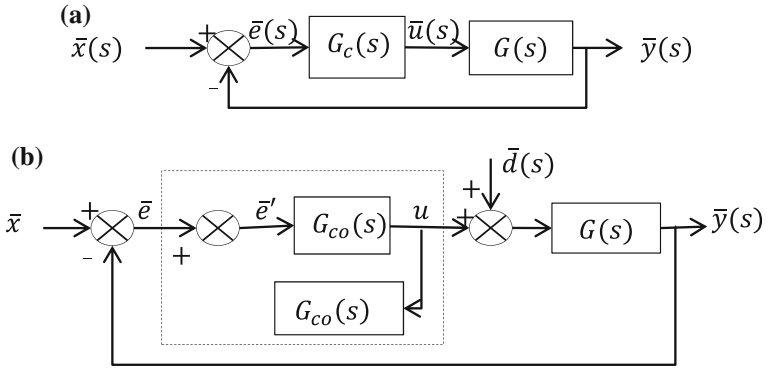



**Fig. 7.25** **a** Feedback loop with $G(s)$ involving time delay $T_d$. **b** Basic structure of controller using a Smith predictor

Another rational approximation of $e^{sT_d}$ which is used in process control is:

$$e^{-sT_d} = 1/(1 + sT_d/m)^m, \quad \text{for } m \geq 5.$$

The second approach was first studied by Smith through the now-called *Smith predictor* in a standard series-compensated system feedback loop [120, 121] (Fig. 7.25).

We observe that the total controller $G_c(s)$ in the system with the Smith predictor contains an internal (secondary) loop made-up by a primary series controller $G_{c0}(s)$ and a feedback controller $\hat{G}_0(s) - \hat{G}(s)$, where $\hat{G}_0(s)$ and $\hat{G}(s)$ are nominal models of $G_0(s)$ and $G(s)$, respectively, and

$$G(s) = G_0(s)e^{-sT_d}$$

The primary controller $G_{c0}(s)$ is usually a PI or PID controller. From Fig. 7.25b, we find that the total controller $G_c(s)$ has the transfer function:

$$G_c(s) = \frac{G_{c0}(s)}{1 + G_{c0}(s)\left[\hat{G}_0(s) - \hat{G}(s)\right]}$$

One can see that the signal $\upsilon(t)$ contains a prediction of the output signal $y(t)$ at a time period equal to $T_d$. This is exactly the reason why the secondary loop around the primary controller is called a predictor. In the ideal case, where $\hat{G}(s) = G(s)$, the overall transfer function of the system is:

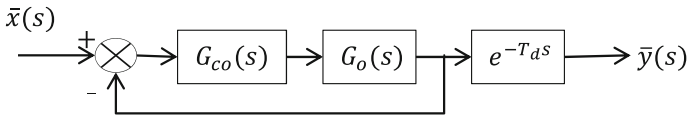



**Fig. 7.26** Equivalent closed-loop system when $\hat{G}(s) = G(s)$ (exact modeling)

$$
\begin{aligned}
H(s) &= \frac{\bar{y}(s)}{\bar{x}(s)} = \frac{G_c(s)G(s)}{1 + G_c(s)G(s)} \\
&= \frac{G_{c0}(s)G(s)/\{1 + G_{c0}(s)[\hat{G}_0(s) - G(s)]\}}{1 + G_{c0}(s)G(s)/\{1 + G_{c0}(s)[\hat{G}_0(s) - G(s)]\}} \\
&= \frac{G_{c0}(s)G(s)}{1 + G_{c0}(s)[\hat{G}_0(s) - G(s)] + G_{c0}(s)G(s)} \\
&= G_{c0}(s)G(s)/[1 + G_{c0}(s)\hat{G}_0(s)]
\end{aligned}
$$

i.e., the time delay was removed from the denominator. This means that the characteristic polynomial of the overall system does not depend on the delay. and the design of $G_{c0}(s)$ could be made on the basis of $G_0(s)$ only as shown in Fig. 7.26.

In practice, however, we have, in general, $\hat{G}(s) \neq G(s)$ and the delay $\exp(-sTd)$ is not removed from the denominator. In this case, the Smith predictor of Fig. 7.26 can only be used if the system under control is stable.

Standard state-space models used for the design of controllers in time-delay systems include the following:

$$
\begin{aligned}
&\dot{\mathbf{x}}(t) = \mathbf{A}x(t) + \mathbf{A}_1\mathbf{x}(t-\tau) + \mathbf{B}\mathbf{u}(t) \\
&\mathbf{x}((k+1)T) = \mathbf{G}\mathbf{x}(kT) + \mathbf{F}\mathbf{x}(t-\tau) + \mathbf{H}\mathbf{u}(kT) \\
&\dot{\mathbf{x}}(t) = (\mathbf{A} + \Delta\mathbf{A}(t))\mathbf{x}(t) + (\mathbf{A}_1 + \Delta\mathbf{A}_1(t))\mathbf{x}(t-\tau) + \mathbf{B}\mathbf{f}(x,t) \\
&\dot{\mathbf{x}}(t) - \mathbf{F}\dot{\mathbf{x}}(t-\tau_1) = \mathbf{A}_0\mathbf{x}(t) + \mathbf{A}_1\mathbf{x}(t-\tau_0(t)), \mathbf{x}(\xi) = \varphi(\xi) \\
&\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}_t, \boldsymbol{\theta}) + \mathbf{g}(\mathbf{x}, \boldsymbol{\theta})\mathbf{u}, \mathbf{x}(\xi) = \varphi(\xi), \xi \in [-\tau, 0]
\end{aligned}
$$

where $\Delta A_1(t)$ is a time-varying (random) parameter disturbance, and $\mathbf{x}_t \in C_n = C([-\tau, 0], R^n)$ is the state, with $C_n = C([a,b], R^n)$ denoting a *Banach space* of continuous functions mapping the interval $[a,b]$ into $R^n$ under the topology by uniform convergence. In the discrete-time model, we have two cases: (i) $\tau = mT$ and (ii) $\tau = mT + \tau'$ where $m$ is an integer $m \geq 1$ and $\tau' < T$. The output equation may or may not involve a delay. More general models with multiple delays, random delays, and periodically varying parameters were also studied over the years to handle particular physical or societal applications. A useful toolbox for the analysis and control of time-delay systems is available in MATLAB (Time Delays: LTI Models) [122, 123].

### 7.11.4 Control of Finite-State Automata

A *finite-state automaton* (or a *finite-state machine*/**FSM**) is an abstract system used to design logic-based systems and computer programs. It has a finite number of states in which it can go through transitions from state to state. These transitions as well the machine's actions are finite. Finite-state automata are applicable to a large number of problems, such as logic design automation, parsing, communication protocol design, flexible manufacturing systems, and, in general, discrete-event systems [124–129].

Actually there are two classes of finite automata (**f.a.**):

- Sequence detectors
- Transducers

A **sequence detector** (or **acceptor** or **recognizer**) provides a binary output *yes* or *no* to indicate whether the input is accepted by the automaton or not. The **f.a.** states are either *accept* (or accepting) states or *non-accept* (not accepting) states. An accept state occurs when the machine has successfully completed its process. In the state-transition diagram an "accept (final)" state is symbolized by a *double circle*. When a **f.a.** is used to define a language, the language is accepted by the **f.a.** if every word of the language is accepted and no word is rejected. The languages accepted by a **f.a.** are called *regular languages* which means that a language is regular if it is accepted by some **f.a.** The *starting state* is symbolized by an *arrow* pointing at some state of the **f.a.**

**Acceptor–Receptor**

Mathematically, a deterministic *f.a. recognizer* (acceptor) $\mathsf{M}$ over an alphabet (finite nonempty set of symbols) $\Sigma$ is a system:

$$\mathsf{M} = (Z, \Sigma, f, z_0, F)$$

where $Z$ is a finite non empty set of states $z_i$,$f$ is a *mapping* of $Z \times \Sigma$ into $Z$, $z_0$ in $Z$ is the initial state, and $F \subseteq Z$ is the set of final states. The mapping:

$$f : Z \times \Sigma \rightarrow Z$$

is called the *state-transition function* of the f.a. $\mathsf{M}$.

The model shown in Fig. 7.27 represents a finite control that reads symbols from a linear input tape. Here, the interpretation of $f(q, a) = p$, for "$q$" and "$p$" in $Z$ and "a" in $\Sigma$, is that $\mathsf{M}$ in state "$q$" and scanning the input symbol "$a$", moves its input head one cell to the right, and goes (transits) to the state "$p$". The transition function $f$ is from $Z \times \Sigma$ to $Z$.

An f.a. that accepts the set of strings with an even number of 0's and an even number of 1's is defined as follows:

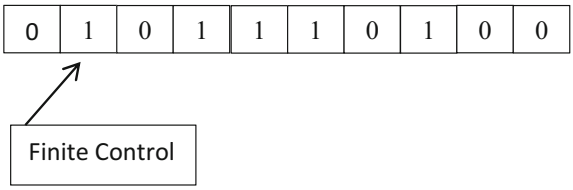



**Fig. 7.27** A finite automaton (control)

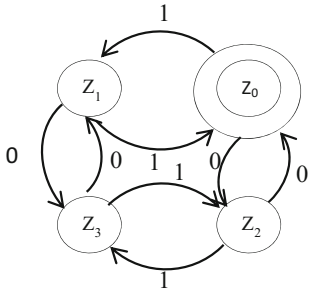



**Fig. 7.28** State-transition diagram of an f.a. that accepts sets of strings with an even number of 0's and an even number of 1's

$$\mathsf{M} = (\Sigma, \Sigma, f, z_0, F), \Sigma = \{0, 1\}$$
$$Z = \{z_0, z_1, z_2, z_3\}, F = \{z_0\}$$
$$f(z_0, 0) = z_2, f(z_1, 0) = z_3, f(z_2, 0) = z_0, f(z_3, 0) = z_1$$
$$f(z_0, 1) = z_1, f(z_1, 1) = z_0, f(z_2, 1) = z_3, f(z_3, 1) = z_2$$

Its state-transition diagram is shown in Fig. 7.28.

Indeed, suppose that the input to $\mathsf{M}$ is 110101. From $f(z_0, 1) = z_1$ and $f(z_1, 1) = z_0$, it follows that $f(z_0, 11) = z_0$, i.e., 11 is in $S(\mathsf{M}) = \{s | f(z, s)\ is\ in\ F\}$. Continuing, we obtain $f(z_0, 0) = z_2$, i.e., $f(z_0, 110) = z_2, f(z_2, 1) = z_3$ i.e., $f(z_0, 1101) = z_3$. Finally, $f(z_3, 0) = z_1$ and $f(z_1, 1) = z_0$. Thus, in overall, the input string 110101 produces the output:

$$f(z_0, 110101) = z_0$$

which means that $s = 110101$ is in $S(\mathsf{M})$. One can show that here $S(\mathsf{M})$ is the set of all sentences $s$ in $\{0, 1\}$ that contain both an even number of 0's and an even number of 1's.

**Transducer**

A *transducer f.a.* produces an output as a response to input and/or state actions. Transducers find application in control and computational linguistic applications. They are classified into two types:

- *Moore machine* in which the output depends only on the state.
- *Mealy machine* in which the output depends on both the input and the state.

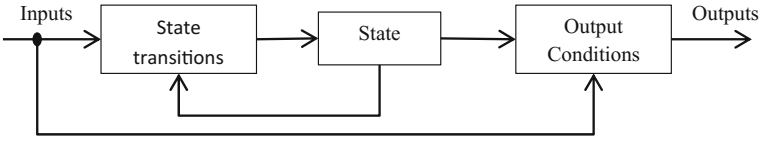



**Fig. 7.29** Structure of Mealy machine

Mathematically, a *transducer f.a.* is a system M:

$$\mathsf{M} = (Z, \Sigma, Y, f, z_0, g)$$

where $Z$ is a finite, nonempty set of states, $\Sigma$ is the (finite) input alphabet, $Y$ is the (finite) output alphabet, $f$ is the state transition $f : Z \times \Sigma \rightarrow Z$, $z_0$ is the initial state, and $g$ is the output function. When the output function depends on both the input alphabet $\Sigma$ and a state in $Z$, i.e., $g : Z \times \Sigma \rightarrow Y$, then we have the *Mealy* machine, otherwise if $g : Z \rightarrow Y$, we have the *Moore* machine.

A Mealy machine has the block diagram shown in Fig. 7.29.

From Fig. 7.29, it can be seen that the Mealy machine has the structure of a control system in state space, having a state-transition relation and an output relation.

A very important class of finite automata is the class of *hybrid automata* that are generalized finite-state machines in which the discrete transitions transfer the system between a finite number of modes $P$. Various types of controllers for hybrid automata are available in the literature, including time-optimal controllers and predictive controllers [130, 131]. The input–output modeling, the controllability and observability, and the canonical forms of finite-state machines are continuously studied since the 1970s [132–135]. The same is true for multivariable control and optimal control of linear sequential machines and several estimation problems of linear and nonlinear finite-state machines [136–141].

### *7.11.5 Control of Discrete-Event Systems*

*Discrete-Event Systems* (**DES**s) are dynamic systems the "*state*" of which changes at discrete instants of time. The term "*event*" characterizes the occurrence of discontinuous changes at known or unknown intervals. Therefore, **DES**s are appropriate for dealing with ordering and sequencing discrete operation [142–150]. Actually, there is no available unified model that covers all the features that a full theory of DES would be desired to include. The case, in which we are interested only in the order in which the events occur, needs a kind of *logical model*. In this case, the time is implicit and the times at which the events take place are ignored. If we want to study the time evolution (trajectory) of the DES, we need some kind of *timed model*. The trajectories are checked if they meet the performance

requirements and satisfy the constraints involved. Examples of man-made systems for which control of discrete-event systems is needed include embedded systems, manufacturing systems, communication protocols, failure detection in telephone switches, feature interaction in telephone networks, verification and validation of software systems, organization systems, vehicular traffic, etc.

The six main methodologies for modeling discrete-event systems are:

- Petri-Nets (**PN**s) theory
- Finite automata (**f.a.**) theory
- Process (min-max) algebra theory
- Logical calculus (temporal logic)
- Markov chains
- Queuing theory

From among them, Petri nets [142–145] and finite automata are the most popular models because they are very easy to use, and they are more tightly connected to engineering models. The control theory for PNs has been developed by many researchers including Y.Ch. Ho, S. Lafortune, F. Capkovic, C. G. Cassandras. CM. Zhou, K. Venkatesh, and others [143–146]. The control theory for finite-automata-based DES was initiated by Wonham and Ramadge and enhanced by B. Kumar, V. Gong, S.I. Marcus, and others [149–153].

**Control of DES Using PN Models** A *Petri Net* is a directed bipartite graph that consists of:

- Places
- Transitions
- Directed arcs

*Places* usually contain a certain number of tokens. *Marking* is called a distribution of tokens over the places of a PN. A *transition* of a PN may *fire* if there exists a token at the start of all input arcs. When it fires, it consumes these tokens and places the tokens at the end of all output arcs. *Arcs* run from a place to a transition or conversely, but never between places or between transitions. A firing is *atomic*, i.e., a single non-interruptible step.

Formally, a PN $N$ is defined as:

$$N = (P, T, U, Y)$$

where $P$ is a set of places, $T$ is a set of transitions, $U$ is the input function, and $Y$ is the output function. Transitions represent the actions that occur in the system, which are controlled by the state of the system described as a set of conditions. Preconditions of a transition are the conditions that must be satisfied in order for the transition to occur.

In the graph of a PN, the places are symbolized by *circles*, and the transitions by *bars*. The input and output functions are represented by *directed arcs* from places to transitions and from transitions to places. The tokens are symbolized by *solid dots* inside the places (*circles*) of the PN graph. If each of the input places of a transition

contains at least one token, then this transition is said to *be enabled*. Thus, a transition can fire only if it is enabled. The number of tokens residing at the places $p_1, p_2, \ldots, p_n$, are given by the components of the marking vector $\mathbf{m} = [m_1, m_2, \ldots, m_n]$, respectively.

An example of a simple PN is [151]:

$$N = (P, T, U, Y)$$

where:

$$P = \{p_1, p_2, p_3, p_4, p_5\}, T = \{t_1, t_2, t_3, t_4\}$$
$$U(t_1) = \{p_1\}, U(t_2) = \{p_2, p_3, p_5\}, U(t_3) = \{p_3\}, U(t_4) = \{p_4\}$$
$$Y(t_1) = \{p_2, p_3, p_4\}, Y(t_2) = \{p_5\}, Y(t_3) = \{p_4\}, Y(t_4) = \{p_2, p_3\}$$

This PN with the marking $\mathbf{m} = \{1, 0, 1, 0, 2\}$ is denoted as $N = \{P, T, U, Y, \mathbf{m}\}$ and has the marked PN graph of Fig. 7.30a.

As it is evident from this diagram, the transition $t_1$ is enabled (one token resides in place $p_1$), but $t_2$ and $t_4$ are not enabled (because no token resides in $p_2$ and $p_4$). Thus, $t_1$ can fire, and, after its firing, the PN diagram is obtained. Note that one token has moved from $p_1$ to all outputs of $t_1$ (i.e., to the places $p_2$, $p_3$, and $p_5$), and thus no token now exists in $p_1$. Hence, $t_1$ is now nonenabled, and $t_2$ becomes enabled. The set of all markings of a PN is called "*the state of the PN*" and is changed after each transition firing.

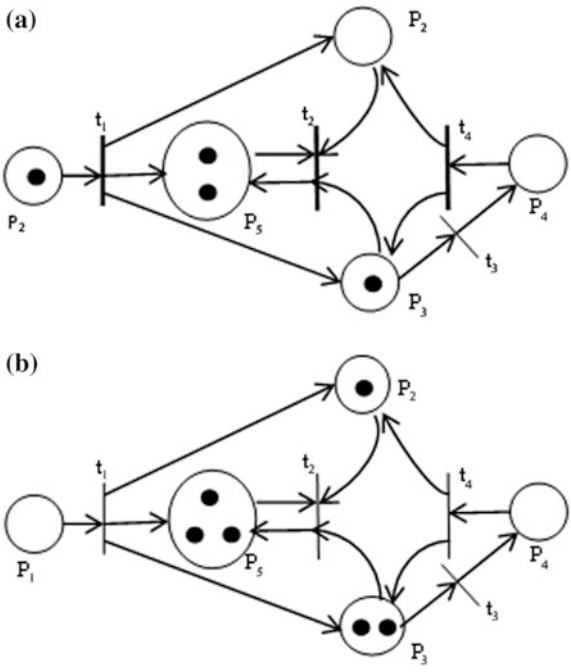



**Fig. 7.30** **a** A simple Petri net. **b** The Petri net after firing of transition 1

One way to take into account the duration of activities in a PN is the following. Consider an activity that needs $\tau$ time units to complete. If the starting time of the activity is $s$ time units, then the end of the activity occurs at $s+\tau$. Thus, the activity can be modeled by two transitions: the first transition occurring at time $s$ (start activity) and the second transition firing at time $s+\tau$ (end activity). Hence, in order for the transition to fire, two conditions must be valid: its input place must have at least one token, and the transition should occur at time $s+\tau$ (current clock time). When all transitions scheduled for the current clock time have been fired, the current clock time is updated to the next scheduled firing of a transition.

The control problem of a dynamic DES is to find the most suitable sequence of the controllable discrete events such that to transfer the system in question from an initial state into a desired final state or into a desired terminal set.

**Control of DES Using Finite-Automata Models**

The concept of finite-state automata (**f.a.**) or finite-state machines (**FSM**) was reviewed in Sect. 7.11.4. To control a DES through finite-state machine modeling, the machine $\mathsf{M}$ should be described as a state-space model:

$$\mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k, \mathbf{u}_k), \mathbf{y}_{k+1} = \mathbf{g}(\mathbf{x}_{k+1}, \mathbf{u}_k),$$

where $\mathbf{x}_{k+1} \in Z$ is the state of $\mathsf{M}$ after the $k$ event, $\mathbf{u}_k$ is the $k$ event, and $\mathbf{y}_{k+1}$ is the vector of the resulting output symbols.

An automaton $\mathsf{M}$ is said to be *deterministic* if for all $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3 \in Z$ and $\mathbf{u} \in U$ (set of control events) the state transitions $\mathbf{x}_1 = f(\mathbf{x}_3, \mathbf{u})$ and $\mathbf{x}_2 = f(\mathbf{x}_3, \mathbf{u})$ imply that $\mathbf{x}_1 = \mathbf{x}_2$. A sequence of states $\mathbf{x}_k\mathbf{x}_{k-1}\ldots\mathbf{x}_1\mathbf{x}_0$ is said to be a *path* (*trajectory*) of $\mathsf{M}$ if there exists an input sequence $\mathbf{u}_k\mathbf{u}_{k-1}, \ldots, \mathbf{u}_1\mathbf{u}_0$ for which $\mathbf{x}_k = \mathbf{f}(\mathbf{x}_{k-1}, \mathbf{u}_{k-1})$ for all $k = 1, 2, \ldots, n$. A path of $\mathsf{M}$ is called a *cycle*, if $\mathbf{x}_k = \mathbf{x}_0$. If $\mathsf{M}$ has no cycles, it is called *acyclic*. The control of a DES, modeled by an f.a., can be *static* or *dynamic*.

The control-design problem is to determine a control sequence that assures the satisfaction of the performance specifications and constraints. The controller that provides this switching control pattern is called a *supervisor*. The supervisor is actually a pair $C\mathsf{M} = (\mathsf{M}, \upsilon)$ where $\mathsf{M}_c$ is a f.a. (deterministic):

$$\mathsf{M}_c = \left(C, U, Y, f, c_0, c_f\right)$$

where $C$ is the state space of $\mathsf{M}_c$, $U$ is the event alphabet, $c_0$ and $c_f$ are the initial and final states in $C$, and $\upsilon$ is an overall function that maps supervisor states to control patterns:

$$\upsilon : C \rightarrow u$$

Obviously, the f.a. under control $\mathsf{M}$ and the supervisor can be coupled in a feedback loop by permitting the state transitions in $\mathsf{M}_c$ to be forced by $\mathsf{M}$ and the states of $\mathsf{M}$ to be driven by the control sequence generated by the supervisor $\mathsf{M}_c$. Wonham and Ramadge [147–150] have worked on the static DES control problem

using the above f.a. formulation and designed the supervisor controller in terms of a formal language $L$ under the condition that $L$ is closed and controllable. The control specifications were stated in terms of state trajectories using predicate calculus [152, 153].

The above results have been extended to the case of dynamic state-feedback controllers that are not static but dynamic depending on the history of the f.a. M. For this dynamic control case, the use of static (memoryless) algorithms (via, e.g., predicate calculus) is not sufficient but use must be made of auxiliary devices that take into account the machine's history information. Such devices are called "*memories*". Details on the derivations and the results can be found in [140, 147–153].

## 7.12 Conclusions

*Modern control* has strongly influenced human technological and nontechnological activity from the operation of simple devices to the control of very complex systems (such as robots, airplanes, spaceships, nuclear reactors, environmental control systems, economic systems, etc.). "Modern control" has very quickly developed to include a wide repertory of techniques and implementations. The birth of control and the history of calculus of variations are discussed in [154, 155]. In this chapter, the fundamental concepts and principles have been were covered at a conceptual, introductory level keeping the mathematics to a minimum. The material of the chapter is supported by a sufficient set of references where the reader can find the details of particular topics in which he/she is interested. The actual space available for the chapter was naturally limited, and so only a few simple examples were included that illustrate many of the concepts and principles discussed in the chapter.

These examples were concerned with the following *how-to-do* issues at a "*paper-and-pencil*" level:

- Controllability and observability testing.
- Lyapunov-based stability testing.
- State-feedback pole shifting control.
- Input–output decoupling control.
- State-observer design.
- Optimal deterministic and stochastic (LQG) control.
- Model-reference adaptive control.

More examples of these issues with simulation and/or real physical results, as well as of issues of practical robust, intelligent, and nonlinear control applications, can be found in the bibliography provided in the previous and the present chapter. The impact of modern control system design no human life and activities will be addressed in Chap. 12.

# References


1. R.C. Dorf, R.H. Bishop, *Modern Control Systems*, 7th edn. (Addison-Wesley, Reading, MA, 1995)
2. J.J. DiStefano III, A.R. Stubberud, I.J. Williams, *Theory and Problems of Feedback Control Systems Design* (Mc Graw-Hill, New York, 1990)
3. L.A. Zadeh, C.A. Desoer, *Linear Systems Theory: The State Space Approach* (Mc Graw-Hill, New York, 1963)
4. R.E. Kalman, J.E. Bertram, Control system analysis and design via the "Second Method" of Lyapunov, (I) continuous-time systems. Trans. ASME J. Basic Eng., 371–393 (1960)
5. L.S. Pontryagin, V.G. Boltyansky, R.V. Gamkrelidze, E.F. Mishchenko, *The Mathematical Theory of Optimal Processes* (Wiley, New York, 1962)
6. J.P. Lasalle, Some extensions of Lyapunov's second method. IRE Trans. Circuit Theor. **7**, 520 (1960)
7. R. Belmann, *Dynamic Programming* (Princeton University Press, New Jersey, 1957)
8. A.P. Sage, C.C. White III, *Optimum Systems Control* (Prentice-Hall, Englewood Cliffs, NJ, 1977)
9. M. Jamshidi, M. Malek-Zavarei, *Linear Control Systems: A Computer-Aided Approach* (Pergamon Press, Oxford, 1986)
10. M. Wonham, *Linear Multivariable Control: A Geometric Approach*, 3rd edn. (Springer, Berlin, 1985)
11. W.A. Wolowitz, *Linear Multivariable Systems* (Springer, Berlin/New York, 1974)
12. M. Athans, P. Falb, *Optimal Control* (Mc Graw-Hill, New York, 1966)
13. P.L. Falb, W.A. Wolowitz, Decoupling in the design of multivariable feedback systems. IEEE Trans. Autom. Control **12**, 651–659 (1967)
14. D.G. Luenberger, An introduction to observers. IEEE Trans. Autom. Control **16**, 233–239 (1971)
15. D.G. Luenberger, Canonical forms for multivariable feedback systems. IEEE Trans. Autom. Control **12**, 290–293 (1967)
16. D.G. Luenberger, Observers for multivariable systems. IEEE Trans. Autom. Control **11**(2), 190–197 (1966)
17. H. Kwakernaak, R. Sivan, *Linear Optimal Control Systems* (Wiley, New York, 1972)
18. J.S. Meditch, *Stochastic Optimal Linear Estimation and Control* (McGraw-Hill, New York, 1969)
19. K.J. Aström, B. Wittenmark, *Adaptive Control* (Addison-Wesley, Reading, MA, 1989)
20. K.S. Narendra, A. Annaswamy, *Stable Adaptive Systems* (Prentice-Hall, New Jersey, 1989)
21. I.D. Landau, *Adaptive Control: The Model Reference Approach* (Marcel-Dekker, New York, 1979)
22. I.D. Landau, R. Lozano, M. M'Saad, *Adaptive Control* (Springer, New York, 1998)
23. K.J. Aström, B. Wittenmark, Problems of identification and control. J. Math Anal. Appl. **34**, 90–113 (1971)
24. J. Richalet, A. Rault, L. Testaud, J. Papon, Model predictive heuristic control: applications to industrial processes. Automatica **14**, 413 (1978)
25. G. Tao, *Adaptive Control Design and Analysis* (Wiley, Hoboken, NJ, 2003)
26. C.R. Cutler, B.L. Ramaker, Dynamic Matrix Control: A Computer Control Algorithm, in *Proceedings of the JACC* (San Francisco, CA, WP-5, 1980)
27. D.W. Clarke, C. Mohtadi, P.S. Tuffs, Generalized predictive control, Part I: the basic algorithm. Automatica **23**(2), 137–148 (1987); Part II: Extensions and interpretations, ibid, p. 149 (1987)
28. R.M.G. De Keyser, A.R. Van Cauwenberghe, Extended Prediction Self-Adaptive Control, in *Proceedings of the 7th IFAC Symposium Identification Systems Parameter Estimation* (York, U.K, 1985), pp. 1255–1260

29. R.M.C. De Keyser, P.H.G.A. Van de Velde, F.A.G. Dumortier, A comparative study of self-adaptive long range predictive control methods. Automatica **24**(2), 149 (1988)
30. J., Martin-Sànchez, A New Solution to Adaptive Control, in *Proceedings of the IEEE,* vol. 64, no. 8 (1976), pp. 1209–1218
31. D.W. Clarke (ed.), *Advances in Model-Based Predictive Control* (Oxford University Press, Oxford, 1994)
32. F. Allgöwer, A. Zheng, *Nonlinear Model Predictive Control* (Birkhauser Basel, Switzerland, 2000)
33. G. Zames, On the input-output stability of time-varying non-linear feedback systems, Part I: conditions derived using concepts of loop gain, conicity, and positivity. IEEE Trans. Autom. Control **11**(2), 228–238 (1966). Part II: conditions involving circles in the frequency plane and sector nonlinearities, *ibid,* no. 3, pp. 465–476 (1966)
34. G. Zames, Feedback and optimal sensitivity: model reference transformations. G. Zames, On the input-output stability of time-varying non-linear feedback systems, Part I: conditions derived using concepts of loop gain, conicity, and positivity. IEEE Trans. Autom. Control **11** (2), 228–238 (1966). Part II: conditions involving circles in the frequency plane and sector nonlinearities, *ibid,* no. 3, pp. 465–476 (1966)
35. G. Zames, Feedback and optimal sensitivity: model reference transformations multiplicative semi-norms and approximate inverses. IEEE Trans. Autom. Control **26**, 301–320 (1981)
36. G. Zames, N.A. Shneydor, Structural stabilization and quenching by dither in nonlinear systems. IEEE Trans. Autom. Control **22**(3), 352–361 (1977)
37. M. Vidyasagar, Normalized coprime factorizations for non-strictly proper systems. IEEE Trans. Auto. Control **33**, 300–301 (1988)
38. M. Vidyasagar, *Control System Synthesis: A Factorization Approach* (MIT Press, Cambridge, MA, 1985)
39. J.F. Doyle, A. Tannenbaum, *Feedback Control Theory* (Macmillan, New York, 1992)
40. T. Chen, B. Francis, *Optimal Sampled-Data Control Systems* (Springer, London/Berlin, 1995)
41. J.C. Doyle, Analysis of Feedback Systems with Structured Uncertainties, in *IEEE Proceedings,* vol. 129, Part D (6) (1982), pp. 242–250
42. G.J. Balas, J.C. Doyle, R. Glover, A. Packard, R. Smith, *The μ-Analysis and Synthesis Toolbox* (Mathworks, Natick, MA, 1991)
43. I. Kurylowicz, J. Jaworska, S.G. Tzafestas, *Robust Stabilizing Control: An Overview, Applied Digital Control* (Marcel Dekker, New York, pp. 289–324, 1993). 41 J.C. Doyle, Analysis of Feedback Systems with Structured Uncertainties, *IEEE Proc.,* Vol.129, Part D (6), pp.242–250, 1982
44. R.E. Kalman, A new approach to linear filtering and prediction problems. ASME J. Basic Eng. **82**, 34–45 (1960)
45. D.E. Kirk, *Optimal Control Theory* (Prentice Hall, Englewood Cliffs, NJ, 1970)
46. B.D.O. Anderson, J.B. Moore, *Optimal Control* (Prentice Hall, Englewood Cliffs, NJ, 1990)
47. A.E. Bryson Jr., Y.-C. Ho, *Applied Optimal Control* (Hemisphere, New York, 1975)
48. T. Kailath, *Linear Systems* (Prentice Hall, Englewood Cliffs, 1980)
49. J.-J.E. Slotine, W. Li, *Applied Nonlinear Control* (Prentice-Hall, Englewood Cliffs, NJ, 1991)
50. A. Isidori, T.J. Tarn (eds.), *Systems Models and Feedback* (Birkhauser, Berlin, 1992), pp. 1–402
51. A. Isidori, *Nonlinear Control Systems* (Springer, Berlin, 1995)
52. G.N. Saridis, Analytic formulation of the principle of increasing precision with decreasing intelligence for intelligent machines. Automatica **25**(3), 461–467 (1989)
53. K.S. Fu, Learning control systems and intelligent control systems: an intersection of artificial intelligence and automatic control. IEEE Trans Autom. Control **16**(2), 70–72 (1971)
54. K.S. Fu, *Syntactic Pattern Recognition and Applications* (Prentice-Hall, Englewood Cliffs, 1982)

55. A. Meystel, Cognitive Controller for Autonomous Systems, in *Proceedings of the IEEE Workshop on Intelligent Control* (RPI, Troy, New York, 1985)
56. R.C. Arkin, Motor schema-based mobile robot navigation. Intl. Robot. Res. **8**(4), 92–112 (1989)
57. R.C, Arkin, *Behavior-Based Robotics* (MIT Press, Cambridge, MA, 1998)
58. J. Albus, R. Quintero, *Towards a Reference Model Architecture for Real-Time Intelligent Control Systems (ARTICS), in Robotics and Manufacturing (ASME)*, vol. 3 (ASME Press, New York, 1990)
59. S.G. Tzafestas (ed.), *Knowledge- Based Systems: Advanced Concepts, Techniques and Applications* (World Scientific, Singapore/London, 1997)
60. S.G. Tzafestas, A. Venetsanopoulos (eds.), *Fuzzy Reasoning in Information, Decision and Control Systems* (Kluwer, Boston/Dordrecht, 1994)
61. S.G. Tzafestas (ed.), *Soft Computing in Systems and Control Technology* (World Scientific, Singapore/London, 1999)
62. S.G. Tzafestas (ed.), *Computational Intelligence in Systems and Control Design and Applications* (Kluwer, Boston/Dordrecht, 1999)
63. Y.-H. Song, A. Johns, R. Aggarwal, *Computational Intelligence Applications to Power Systems* (Kluwer, Boston/Dordrecht, 1996)
64. S. Haykin, *Neural Networks: A Comprehensive Foundation* (Prentice Hall, Upper Saddle River, NJ, 1999)
65. A.G. Barto, R.S. Sutton, W. Aderson, Neuronlike adaptive elements that can solve difficult learning control problems. IEEE Trans. Syst. Man Cybern **13**(5), 834–847 (1983)
66. F.C. Chen, Back Propagation Neural Network for Nonlinear Self-Tuning Adaptive Control, in *Proceedings of the IEEE Intelligent Machines*, pp. 274–279 (1989)
67. K. Tzafestas, Watanabe, Learning Algorithms for Neural Networks with the Kalman Filters. J. Intell. and Robotic Syst. **3**(4), 305–319 (1990)
68. K. Watanabe, T. Fukuda, S.G. Tzafestas, Adaptive control for CARMA systems using linear neural networks. Int. J. Control **56**, 483–497 (1992)
69. S. Omatu, M. Khalil, R. Yusof, *Neuro-Control and its Applications* (Springer, Berlin, 1996)
70. D. Nguyen, B. Widrow, Neural-networks for self-learning control systems. IEEE Control Syst. Mag. **10**(3), 18–23 (1990)
71. C.T. Lin, *Neural Fuzzy Control Systems with Structure and Parameter Learning* (World Scientific, Singapore/London, 1994)
72. F.L. Lewis, J. Campos, R. Selmic, *Neuro-Fuzzy Control of Industrial Systems with Actuator Nonlinearities* (SIAM Publications, Philadelphia, PA, 2002)
73. S.G. Tzafestas, G.G, Rigatos, Neural and neurofuzzy FELA adaptive robot control using feedforward and counterpropagration networks. J. Intell. Robot. Syst. **23**(2–4), 291–330 (1998)
74. C.I. Harris, C.G. Moore, M. Brown, *Intelligent Control: Aspects of Fuzzy Logic and Neural Networks* (World Scientific, Singapore/London, 1993)
75. J.-S.R. Jang, ANFIS: Adaptive-network-based fuzzy inference systems. IEEE Trans. Syst. Man Cybern. **23**, 665–685 (1993)
76. H.R. Berenji, P. Khedkar, Learning and tuning fuzzy logic controllers through reinforcements. IEEE Trans. Neural Netw. **3**, 724–740 (1992)
77. G.A. Carpenter, M.N. Gjaja, S. Gopal, C.E. Woodcock, ART neural networks for remote sensing: vegetation classification from landsat TM and terrain data. IEEE Trans. Geosci. Remote Sens. **35**, 308–325 (1997)
78. G.A. Carpenter, S. Grossberg, N. Markuson, J.H. Reynolds, D.B. Rosen, Fuzzy ARTMAP: a neural network architecture for incremental supervised learning of analog multidimensional maps. IEEE Trans. Neural Netw. **3**, 698–712 (1992)
79. S.G. Tzafestas, K.C. Zikidis, An Online Learning Neuro-fuzzy Architecture Based on Functional Reasoning and Fuzzy ARTMAP, in *Proceedings of the ICSC International Symposium on Fuzzy Logic Applications* (Zurich, Switzerland, 1997)

80. S.G. Tzafestas, K.C. Zikidis, NeuroFAST: on-line neuro-fuzzy ART-based structure and parameter learning TSK model. IEEE Trans. Syst. Man Cybern. Part B: Cybern. **31**(5), 797–802 (2001)
81. T.C. Lin, C.S. Lee, Neural network based fuzzy logic control and decision system. IEEE Trans. Comput. **40**(12), 1320–1336 (1991)
82. R.E. Bellman, R.E. Kalaba, Quasilinearization and Nonlinear Boundary-Value Problems, in *Modern Analytic and Computational Methods in Science and Mathematics*, vol 3, ed. R.E. Bellman, R.E. Kalaba (American Elsevier, New York, 1965)
83. B. Pradin, A. Titli, Methods of Decomposition—Coordination for the Optimization of Interconnected DPS, in *Proceedings of the 6th IFAC Triennial Congress*, Paper 15-3 (Boston, 1975)
84. M.G. Singh, A. Titli, *Systems: Decomposition, Optimization and Control* (Pergamon, Oxford, 1978)
85. D.D. Siljak, M.K. Sundareshan, A Multilevel Optimization of Large Scale Dynamic Systems. IEEE Trans. Auto. Control **21**(3), 363–366 (1976)
86. D.D. Siljak, Multilevel stabilization of large scale systems: a spinning flexible spacecraft. Automatica **12**, 309–320 (1976)
87. M.G. Singh, M.F. Hassan, A. Titli, Multilevel Feedback Control of Interconnected Dynamical Principle Using the Prediction Principle*, IEEE Trans. Syst. Man. Cybern.*, Vol. SMC-6, No.4, pp. 233–239, 1976
88. M.S. Mahmoud, M.F. Hassan, M.G. Darwish, *Large Scale Control Systems Theories and Techniques* (Marcel Dekker, New York, 1985)
89. S.G. Tzafestas, M.F. Hassan, Complex large scale systems methodologies in conjunction with modern computer technology. Control-Theor. Adv. Technol. **2**(2), 105–120 (1986)
90. E.J. Davison, The robust decentralized control of a general servomechanism problem. IEEE Trans. Auto Control **21**(1), 14–24 (1976)
91. D.D. Siljak, M.B. Vukcevic, Decentralization, stabilization and estimation of large scale linear systems. IEEE Trans. Auto. Control **21**(3), 363–366 (1976)
92. M.G., Singh, *Decentralized Control* (North-Holland, Amsterdam, 1981)
93. C.W. Sanders, E.C. Tacker, T.D. Linton, Decentralized Filtering Algorithms for Interconnected Systems, in *Proceedings of the IFAC Congress,* Paper 26-2 (Boston, 1975)
94. C.W. Sanders, E.C. Tacker,T.D. Linton, A new class of decentralized filters for interconnected systems. IEEE Trans. Auto. Control **19**(3), 259–262 (1974)
95. S.H. Wang, E.J. Davison, on the stabilization of decentralized control systems. IEEE Trans. Auto. Control **18**(5), 473–478 (1973)
96. S.G. Tzafestas (ed.), *Distributed Parameter Control Systems: Theory and Application* (Pergamon, Oxford, 1982)
97. S.G. Tzafestas, Distributed Parameter Systems Estimation and Control: An Update, Systems and Control Encyclopedia, vol. 2, ed. M.G. Singh (Pergamon, Oxford, 1992), pp. 710–723
98. P.K.C. Wang, Asymptotic stability of distributed parameter systems with feedback controls. IEEE Trans. Auto. Control **11**(1), 46–54 (1966)
99. A.G. Butkovskii, The maximum principle for optimum systems with distributed parameters. Autom. Remote Control **22**, 1429–1438 (1962)
100. Y. Sakawa, Optimal control of a certain type of linear distributed- parameter systems. IEEE Trans. Autom. Control **11**(1), 42–45 (1966)
101. P.K.C. Wang, Mathematical Modeling of Systems with Distributed- Parameter Systems (ASME Publications, 1959), pp. 49–63
102. P.K.C. Wang, Modeling and control of nonlinear micro-distributed systems, nonlinear analysis. Theor. Methods Appl. **30**(6), 3215–3226 (1997)
103. Y. Sawaragi, T. Soeda, S. Omatu, *Modeling Estimation and Their Applications for Distributed Parameter Systems* (Springer, Berlin/New York, 1978). (Springer, LNCTS-11)
104. S.G. Tzafestas, J.M. Nightingule, Optimal filtering, smoothing and prediction in linear distributed—parameter systems. Proc. IEE **115**(8), 1207–1212 (1968)

105. S.G. Tzafestas, J.M. Nightingale, Optimal control of a class of linear stochastic distributed—parameter systems. Proc. IEE **115**(8), 1213–1220 (1968)
106. S.G. Tzafestas, J.M. Nightingale, Differential dynamic programming approach to optimal nonlinear distributed-parameter control systems. Proc. IEE **116**(6), 1085–1093 (1969)
107. S.G. Tzafestas, On optimum distributed—parameter filtering and fixed-interval smoothing for colored noise. IEEE Trans. Auto. Control **17**(4), 448–458 (1972)
108. Y. Sawaragi, T. Soeda, S. Omatu, *Modeling Estimation and Their Applications for Distributed Parameter Systems* (Springer, Berlin/New York, 1978). (Springer, LNCTS-11)
109. Y. Sunahara, A. Ohsumi, M. Imamura, A method of parameter identification for linear distributed parameter systems. Automatica **12**, 245–256 (1976)
110. G.K. Lausterer, W.H. Ray, H.R. Martenes, Real time distributed-parameter state estimation applied to a two-dimensional heated ingot. Automatica **14**(4), 335–344 (1978)
111. S.G. Tzafestas, Decoupling of nonlinear time-delay systems. Int. J. Syst. Sci. **5**(4), 301–307 (1974)
112. S.G. Tzafestas, T. Pimenides, Partial input-output decoupling in time-delay systems. Electron. Lett. **11**(15), 353–354 (1975)
113. S.G. Tzafestas, Model-matching in time-delay control systems. IEEE Trans. Autom. Control **21**, 426–428 (1976)
114. J.-P. Richard, Time-delay systems: an overview of some recent advances and open problems. Automatica **39**, 1667–1694 (2003)
115. W. Michiels, S.-I. Niculescu, On delay sensitivity of Smith predictors. Int. J. Syst. Sci. **34**(8–9), 543–551 (2003)
116. Z. Wang, K.J. Burnham, Robust filtering for a class of stochastic uncertain nonlinear time-delay systems via exponential state estimation. IEEE Trans. Signal Process. **49**(4) (2001)
117. V. Suplin, E. Fridman, U. Shaked, H∞ control of linear uncertain time-delay systems: a projection approach. IEEE Trans. Auto. Control **51**(4), 680–684 (2006)
118. P.A. Prokopiou, S.G. Tzafestas, W.S. Harwin, A novel scheme for human-friendly and time-delays robust neuropredictive teleoperation. J. Intell. Robot. Syst. **25**(4), 311–340 (1999)
119. Q.-C. Zhong, *Robust Control of Time-Delay Systems* (Springer, London, 2006)
120. O.J.M. Smith, A controller to overcome dead time. ISA J. **6**(2), 28–33 (1959)
121. Z.J. Palmor, Time-delay compensation-Smith predictor and its modifications, in *Control Handbook*, ed. W.S. Levine (CRC and IEEE Press, New York, 1996), pp. 224–237
122. A.V. Kim, W.H. Kwon, V.G. Pimenov, Time Delay System Toolbox (for Use with MATLAB), Users Guide 2000. http://home.imm.wran.ru/fde/toolbox.html, http://fde.imm.uran.ru (for on line simulation)
123. F. Haugen, *Tutorial for Control System Toolbox for MATLAB*, Oct 11, 2003. http://techteach.no/publications/control_system_toolbox/
124. A. Gill, *Introduction to the Theory of Finite-state Machines* (McGraw-Hill, New York, 1962)
125. S. Ginsburg, *An Introduction to Mathematical Machine Theory* (Addison-Wesley, Reading, MA, 1962)
126. T.A. Booth, *Sequential Machines and Automata Theory* (Wiley, New York, 1967)
127. M.A. Arbib, *Theories of Abstract Automata* (Prentice Hall, Englewood Cliffs, NJ, 1969)
128. T. Kam, *Synthesis of Finite State Machines: Functional Optimization* (Kluwer, Boston, MA, 1997)
129. P. Linz, *Formal Languages and Automata* (Jones and Barlett, Sudbury, MA, 2006)
130. Y. Pang, M.P. Spathopoulos, *On Weighted Time-Optimal Control for Linear Hybrid Automata Using Quantifier Elimination*. citeseerx.ist.psu.edu/viewdoc/summary.doi=10.11.1.3.4570
131. Y. Pang, M.P. Spathopoulos, J. Raisch, On Suboptimal Control Design for Hybrid Automata Using Predictive Control Techniques, in *Proceedings of the16th IFAC World Congress* (Prague, 2005)

132. S.G. Tzafestas, Input- output modeling and identification of linear automata. Inform-Process. Lett. **1**(3), 273, 295 (1972)
133. S.G. Tzafestas, Concerning controllability and observability of linear sequential machines. Int. J. Syst. Sci. **4**(6), 833–858 (1973)
134. A. Nerode, W. Kohn, Models for Hybrid Systems: Automata, Topologies, Controllability, Observability, in *Hybrid Systems, LNCS-736* (Springer, Heidelberg, 1993)
135. G.F. Beckhoff, *Controllability-Observability Type Duality Relations of Finite-State Machines, LNCS-1* (Springer, Heidelberg, 1973)
136. S.G. Tzafestas, Multivariable control theory in linear sequential machines. Int. J. Syst. Sci. **4** (3), 363–396 (1973)
137. S.G. Tzafestas, State estimation algorithms for nonlinear stochastic sequential machines. Comput. J. **6**(3), 295–253 (1973)
138. E. Athanasopoulou, C.N. Hadjicostis, Maximum Likelihood Diagnosis in Partially Observable Finite State Machines, in *Proceedings of the MED-2005, IEEE International Symposium on Intelligent Control and Automation* (Limassol, 2005), pp. 896–901
139. Y. Pang, M.P. Spathopoulos, J. Raisch, On Suboptimal Control Design for Hybrid Automata Using Predictive Control Techniques, in *Proceedings of the 16th IFAC World Congress* (Prague, 2005). http://www.nt.ntnu.no/users/skoge/prost/procedings/ifac2005/Fullpapers/03461.pdf
140. A. Ray, J. Fu, C. Lagoa, Optimal supervisory control of finite state automata. Int. J. Control **77**(12), 1083–1100 (2004)
141. E. Tronci, Optimal Finite State Supervisory Control, in *Proceedings 35th IEEE Conference on Decision and Control* (Kobe, Japan, 1996)
142. YCh. Ho, Introduction to dynamics of discrete event systems. Proc. IEEE **77**(1), 3–6 (1989)
143. S.L. Chung, S. Lafortune, F. Lin, Limited lookahead policies in supervisory control of discrete event systems. IEEE Trans. Auto. Control **37**(12), 1921–1935 (1992)
144. F. Capkovic, Knowledge-Based Control Synthesis of Discrete Event Dynamic Systems, in *Advances in Manufacturing: Decision, Control and Information Technology*, ed. S.G. Tzafestas (Springer, London, 1999), pp. 195–180
145. C.G. Cassandras, S. Lafortune, *Introduction to Discrete Event Systems* (Kluwer, Norwell, MA, 1999)
146. M. Zhou, K. Vencatech, *Modeling Simulation and Control of Flexible Manufacturing Systems: A Petri Net Approach* (World Scientific, Singapore, 1998)
147. W.M. Wonham, P.J. Ramadge, On the supremal controllable sublanguage of a given language. SIAM J. Control Optimiz. **25**, 637–659 (1987)
148. P.J. Ramadge, W.M. Wonham, Supervisory control of a class of discrete event processes. SIAM J. Control Optimiz. **25**(1), 206–230 (1987)
149. W.M. Wonham, P.J. Ramadge, Modular supervisory control of discrete event systems. Mathem. Control Signals Syst. **1**(1), 13–30 (1988)
150. F. Lin, W.M. Wonham, Decentralized supervisory control of discrete—event systems. Inform. Sci. **44**, 199–224 (1988)
151. J. Duggan, J. Browne, ESPNET: Expert-Systems-Based Simulator of Petri Nets, Proc. IEEE Part D **135**(4), 239–247 (1988)
152. R. Kumar, V.K. Carg, *Modeling and Control of Logical Discrete Event Systems* (Kluwer, Norwell, MA, 1995)
153. R. Kumar, V. Gong, S.I. Markus, Predicates and predicate transformers for supervisory control of discrete event systems. IEEE Trans. Autom. Control **38**(2), 232–247 (1993)
154. J.C. Willems, The Birth of Optimal Control, in *Proceedings of the 35th CDC* (Kobe, Japan. 1996), pp. 1586–1587
155. H.H. Goldstine, *A History of Calculus of Variations from the 17th through the 19th Century* (Springer, Berlin, 1980)
156. L.A. Zadeh, Fuzzy sets. Inf. Control **8**, 338 (1965)

# Chapter 8
# Adaptation, Complexity, and Complex Adaptive Systems

*Adaptation is the heart and soul of evolution.*
Niles Eldredge
*Expansion means complexity and complexity decay.*
C. Northcote Parkinson

**Abstract** Adaptation is inherent in all biological organisms and societal systems, and provides the means for assuring the fitness and survival of any biological species or society in a given environment. It was of primary concern by biologists and scientists over time and produced strong debates about its nature and impact on life evolution. Complexity is also an inherent property of life, human society, and technology. It is due to the interrelationship, interdependence, and connectivity of elements and entities in the interior and the environment of an organism or system. Complex Adaptive Systems (CAS) have the general properties of complex systems, but they also exhibit several higher level features. In this chapter, an overview of this field is provided including biological, hard science, soft science, and computer science issues. This chapter starts by introducing the concept of adaptation, its manifestations, and its basic properties and mechanisms. The adaptation measurement aspect is also examined. Then, the concept of "emergence", which again is one of the most difficult philosophical concepts strongly connected with delicate questions of life existence and evolution on Earth, is examined. This chapter includes a short historical note highlighting the results and opinions of workers that have initiated and expanded the adaptation, and emergence scientific field.

**Keywords** Adaptation · Adaptation measures · Adaptation mechanisms
Structure adaptation · Function adaptation · Physiological adaptation
Evolutionary adaptation · Complexity · Complex system · Complex adaptive system (CAS) · Nonlinear system · Chaos · Space chaos/fractals
Self-similarity · Time chaos/chaotic system · Strange attractor · Cellular automata chaos · Emergence · Epistemological/ontological emergence · CAS fitness landscape · Genetic algorithm (GA) · Fitness function · Evolutionary cycle



S. G. Tzafestas, *Energy, Information, Feedback, Adaptation, and Self-organization*,
Intelligent Systems, Control and Automation: Science and Engineering 90,
https://doi.org/10.1007/978-3-319-66999-1_8

## 8.1 Introduction

This chapter is devoted to *adaptation* in biology, nature, and society, and to the related discipline of *complexity and complex adaptive systems*. The processes and products of adaptation are of central concern in biology, anthropology, archaeology, and human society's evolution. The dominant thesis of the 1940s was that "*culture is a unique human invention,*" which absolutely separates humankind from other species. Darwinists did not accept such an extreme separatism (or biologism). However, this complete separation dominated the social sciences in the twentieth century, because many human society adaptations appeared to be free of any direct conscious involvement of biologically oriented goals. Today, many workers in this field appear to agree that the societal type of adaptation is not unique to humankind, is not independent of biological adaptation, and is not biologically imperative [1]. This thesis must be appropriately embedded into the three biological types of adaptation (adaptation to inorganic environment, adaptation to organic environment, and adaptation of the living being to its own interior) suggested by Huxley [2]. In all case, *life* is a "*survival enterprise*" which, no matter what kind of perceptions, beliefs, or illusions we have, is applicable to humankind as well. Survival is the principal life problem for any human society and a prerequisite for any other more specialized goal. Adaptation is the means for assuring survival and reproduction of any species or society in a given environment.

Actually, an organism, species, or human society is a complex system of adaptations. Complexity is due to the interrelationship, interaction, interdependence, and connectivity of entities or elements within a system, and between the interior and the environment of the system. The term "complexity" has its origin in the Greek word "πλοκή" (ploki) that has the meaning of *interweaving* or *entwining* or *braiding* (Latin: *plexus*). Complexity theory, in general, was initiated in Europe within the established "natural sciences" framework of systems, cybernetics, chaos, and irreversible processes. But it was predominantly developed fully in the USA, leading to the field of *Complex Adaptive Systems* (**CAS**) at Santa Fe Institute (New Mexico). Complex adaptive systems possess the general properties of complex systems, but they exhibit several higher level properties as will be explained in this chapter.

The purpose of the chapter is as follows:

- To present the concept of adaptation and its manifestations.
- To provide a list of short historical notes highlighting the results derived by the founders of the field.
- To examine the issue of adaptation measurement.
- To define and present an outline of "complex adaptive systems" (CAS).
- To review the theory of chaos and chaotic nonlinear systems, including both spatial chaos (fractals) and time chaos (strange attractors).
- To study the concept of complexity, giving the properties of complex systems described by several authors.
- To discuss the concept of "emergence", one of the most difficult philosophical concepts profoundly connected with delicate questions of the existence and evolution of life on Earth.

– To discuss some further issues of CAS, including the topic of 'genetic algorithms' (GA), that were conceptualized by John Holland at the Santa Fe Institute.

## 8.2 What Is Adaptation?

*Life exhibits adaptation*. This observed fact of life has been recognized by most biologists, anthropologists, archaeologists, philosophers, ecologists, and other scientists who are concerned with "living beings". *Julian Huxley* notes: "The significance of adaptation can only be understood in relation to the total biology of the species" [2]. The object of adaptation varies according to the nature of the system or function to which the reference is made, but it is in general accepted that adaptation means the following (see, e.g., answers.com/topic/adaptation):

- The act or process of adapting
- The state of being adapted
- Something (e.g., a mechanism or device) that is changed or changes to fit a new situation.

All of these meanings and interpretations of the term adaptation are in common use today in science and society. According to the Britannica Encyclopedia (Brittanica.com): "Adaptation in biology is the process by which an animal or plant becomes fitted to its environment. It is the result of natural selection acting on inherited variation." According to the Sci-Tech Encyclopedia: "Adaptation is a characteristic of an organism which makes it fit for its environment or for its particular way of life."

The research on *adaptation* has revealed several forms of adaptation. These include, but are not restricted to, the following:

- Structure adaptation
- Physiological adaptation
- Function adaptation
- Evolutionary adaptation
- Genetic adaptation.

**Structure adaptation** includes, among others, the anatomical adaptation of animals. For example, the shape of the body of the fish is adapted to life in water, the body of the bird is adapted for flight, horse legs are adapted for running on the grass, and the long ears of the rabbits living in desert-like environments enable them to radiate heat more efficiently and so survive under harsh conditions.

**Physiological adaptation** includes the responsive adaptation of sense and action organs. For example, the eye has the ability to adjust to changing conditions of light intensity, and heart function can be adapted to higher workloads, etc. Clearly, physiological adaptations span an organism's lifetime.

**Function adaptation** includes the historical adjustments of the manner in which the organs of an organism work to fit the changing long-term conditions of the organism's habitat. This is again achieved through natural selection that forms and maintains the function concerned. For example, the function of the heart is pumping blood. The resulting sound is not a function but a side effect of pumping. Every organ of an organism has a functional history, which has undergone *selection* for its *survival* in its environment.

**Evolutionary adaptation** refers to adjustment of living matter to environmental conditions and to other living things over very long periods (e.g., over many, many generations of a given population). The adaptation property is a property of life which is not possessed by inanimate matter.

**Genetic adaptation** takes place in a population via *natural selection* which affects its genetic variability (i.e., the population undergoes adaptive genetic adjustments to cope with the circumstances). Genetic adaptation (also called genetic improvisation) occurs at the **DNA** (**d**eoxyribo**n**ucleic **a**cid) level and is brought about by *mutation* and *selection*. Genetic adaptation includes adaptations that may result in visible structures or changes to physiological activity so as to fit the environment's changes.

According to Darwin [3], life processes are categorized as one of the following:

- **Internal processes**, i.e., the processes that generate the organism.
- **External processes**, i.e., the processes that take place in the environment where the organization must live.

Before Darwin, there was no clear distinction between internal and external processes. Adaptations of internal processes have to do with the totally coordinated adjustments in the organism's (or system's) body, e.g., temperature control, blood pressure control, suitable changes in the immune system, etc. The adaptation that comes from the interaction of the organism or population with nature (the habitat) is performed through a suitable feedback process. This means that nature "*selects*" the "*design*" that best solve the adaptation problem. Prior to Darwin, the only systematic explanation of design in nature was the existence of a designer and a purpose for life's diversity. That is, adaptation was thought to be a sign of design and purpose. With his theory of evolution by natural selection, Darwin changed this line of thought, underwriting a fully naturalistic explanation of function in the biological world.

However, the point of view that adaptation provides evidence that the world is governed by design is still held and promoted by many thinkers. Adaptations observed in "nature", which reveal the sensitive fit between live beings and their habitats, are considered by them as a proof of *"Creator's Design"*. There is an ongoing debate regarding adaptations as the act, state, or mechanism of change to fit new circumstances and adaptation as an indicator of design and purpose (i.e., between *evolutionists and creationists*) [4].

## 8.3 Historical Note

Over the years, biologists and other scientists have tackled the issue of adaptation and developed many different theories. Here, we will present only a few representative ones.

In ancient times, *Aristotle* regarded life adaptation as a process toward a purpose, while *Empedocles* had exactly the opposite opinion, i.e., adaptation did not require a purpose, but came about naturally, since such creatures survived [5].

***Descartes*** viewed the world as consisting of two separate parts, viz., an active, striving purposeful psychological part (thinking, perception), and a "dead" physical part (matter, body) defined fully by its extension in space and time without any inherent *arrow of time*. According to Descartes, the active part of the world is limited to human minds [6].

***Kant*** presented another fundamental duality in the world, namely the dualism between the active striving of living beings and their dead environments [6].

***Boltzmann*** *(1886),* as we extensively described in Sect. 3.10.2, promoted the view that the physical world is not actually dead or passive, but persistently moving toward increasing disorder (contrary to Newtonian physics).

***Darwin*** *(1872)* studied the imperfections and limitations of animal and plant life that have existed over the time and concluded that, as the *habitat* changed, so did the *biota* and that their habitats were subject to changes in their biota (e.g., by invasions of species from other regions) [3]. As mentioned before, Darwin divided life processes into *internal* and *external* and developed his theory of *evolution by natural selection*. Building on the Cartesian–Kantian dualistic point of view, most ascendants of Darwin *(neo-Darwinists)* promoted the theory of complete separation between living beings and environments or, equivalently, the full independence of biology from physics [6, 7].

***Donald Fisher*** *(1930)*, a neo-Darwinist, examined deeply the apparent incommensurability between biology (evolution) and physics (thermodynamics), whereby evolution leads toward an increase of order (organization) and entropy change leads toward a decrease of order (disorganization). He put forward the idea that the two opposite directions of evolution and thermodynamics could be unified under a more general principle [6].

***Lotka*** *(1922)* concluded that natural selection tends to maximize energy flux, so far as compatible with the constraints to which the system is subject (see Sect. 3.8.1 ), and so the law of natural selection becomes also the law of evolution (maximum energy flux principle).

***Odum*** *(1963)* and his colleagues have developed further Lotka's "maximum energy flux principle" providing a corollary to it called "maximum empower principle" and pointing out that "in surviving designs a matching of high-quality

energy with larger amounts of low-quality energy is likely to occur" (see Sect. 3.8.1).

***Johannes von Uexküll*** did not accept the Darwinian separation of organism and habitat, promoting the thesis that: "Every animal carries its environment about with it, like an impenetrable shell, all the days of its life. Every animal is bound to its environment by the circle of functions". Uexküll developed the *"pure natural science"* approach of *living beings* and investigated the *"structural plan" (blueprint)* of them, their origin, and their outcomes. According to this theory, the stimuli of the environment, which an animal is able to receive via its blueprint, constitute the only reality present to it. Due to this physical limit, the organism closes itself to all other spheres of existence [7].

***Ehrlich and Raven*** *(1964)* introduced the *coevolution* concept to describe the two-way dynamic interaction between organisms and their environments. Living beings adapt to their environment (living and nonliving) and, in this adaptation process, environments are many times modified, probably in such a way that influences the living beings [8].

***John Maynard Smith*** *(1975)* explained that the *a priori* reasoning of *evolutionists* (according to which there must be an adaptive/functional explanation for any trait, and conversely that natural selection provides an explanation for every biological phenomenon) is not necessarily wrong and may be the best way to proceed. Most traits of living creatures probably evolved toward survival and earning a living, despite the fact that they might not, at the time, be optimal or in any way adaptive [1, 9].

***Theodosius Dobzhansky*** *(1956–1968)* defined the terms adaptation, adaptedness, and adaptive trait as follows [5]: (i) *Adaptation* is the evolutionary process whereby an organism becomes better able to live in its habitat or habitats [10], (ii) *Adaptedness* is the state of being adapted, the degree to which an organism is able to live and reproduce in a given set of habitats [11], and (iii) *An adaptive trait* is an aspect of development of pattern of the organism which enables or enhances the probability of that organism surviving and reproducing [12].

***Donald Hardesty*** *(1977)*, an ecological anthropologist, regarded the adaptation process "as any beneficial response to environment" meaning "biologically beneficial" [1, 13].

***Richard Lewontin*** *(1978, 1979)* adopted the constructionism view, according to which the world is changing because the living creatures are changing. According to evolutionary biologists, the relation between adaptation and ongoing changes in the environment is analogous to what happened in the queen in *Through the Looking Glass,* who realized that she had to continue running just to keep in place. On this basis, they formulated the so-called "*Red Queen's Running Hypothesis*" which makes the constructionist view even stronger [7]. Lewontin reexamined the relationship between outside (environment) and inside (organism) and argued that

the evolutionary process can be better explained by the construction metaphor [14, 15].

***Bronislaw Malinowski*** *(1944),* an anthropologist, adopted the *basic-needs* approach to adaptation in society. For him, a society is actually a self-organized system of cooperatively pursued activities. It has a purposeful behavior, whereby its purpose is primarily the "*satisfaction of basic needs*", i.e., the conditions in the human organism within the cultural and natural environment that are necessary and sufficient for the survival of the organism and the society or group.

***Peter Corning*** [1] has developed further this "basic-needs approach" to societal adaptation on the grounds of the biological need of survival and reproduction, i.e., in connection with biological fundamentals.

***Rod Swenson*** *(1988)* worked out the entropy interpretation "as opposite to potential" and develop the *Law of Maximum Entropy Production* **(MEP)** according to which: "A system will select the path of assemblage of paths that maximizes the entropy at the fastest rate given the constraints." This is the opposite of Boltzmann's interpretation of the entropy law as the "law of disorder" (see Sect. 3.10.4). Going a step further, Swenson and Turvey concluded that the world is in the *order production business* because ordered flow produces entropy faster than disordered flow. This showed that the MEP law has important implications for evolutionary theory, social adaptation, and human development. In other words, it seems that the MEP law is a principle that unifies evolution (biology) and thermodynamic (physics) [6]. Swenson convincingly explained that, from prokaryotes up to our present day, cultural systems' "evolution on Earth can be regarded as an epistemic process where the global system as a whole learns to degrade the cosmic gradient at the fastest possible rate given the constraints" [16].

## 8.4 Adaptation Mechanisms

The basic mechanisms of *adaptation* (or *adaptability*) are the following:

- Feedback mechanism
- Feed-forward mechanism,

both of which have a behavioral structure. This means that, in the interaction of an organism or population with nature, adaptation is performed through feedback, i.e., nature selects the design that best solves the adaptation problem applying *positive feedback* to the change that enhances the capability to cause the selected design's own reproduction. If a change leads to fewer (or no) offspring, a *negative feedback* is applied, and the design will become extinct. Thus, evolution by *natural selection* always tends to enhance *fitness* to the environment, making organisms *better adapt* to their habitat and way of life. Of course, this adaptation, which proceeds in small steps, may not be ultimately perfect in all cases.

The feed-forward and feedback mechanisms are embedded in the so-called *program of a living being* [17], i.e., in the plan that specifies the organisms' body ingredients and the interaction as the organism persists overtime. The feed-forward and feedback processes are responses at the molecular level, the organ and body levels of the organism, and at the population level, which secure "*survival*" in quickly varying habitats.

When the environment (habitat) changes, the resident population (e.g., flying insects, oceanic organisms, etc.) goes to another more suitable place. This feedback response is the so-called "*habitat tracking*", which however does not result in adaptation.

At the *program level,* adaptation is specifically called *improvisation,* which is implemented by the DNA and brought about by *mutation* and *selection.* Program improvisation is actually an optimization of the program to face newly appearing environmental conditions. According to Koshland [17], adaptation and improvisation are handled on Earth by different mechanisms, and so they must be considered as different concepts and treated by different mechanisms in any newly devised or newly discovered system. The relative types of species in a given environment are always subject to change, so, in nature, "*change is the rule*".

According to *Darwin,* the following three conditions are necessary for evolution by natural selection [3, 18]:

- **Variation:** This refers to the variation in phenotype characteristics among population members.
- **Inheritance:** This refers to the fact that these characteristics are heritable to some extent (i.e., offspring have the characteristics of their parents more often than the average characteristics of the population).
- **Differential reproductive success:** This refers to the fact that various different variants produce different numbers of offspring in succeeding generations.

As *Robert Brandon* points out [18], these conditions are not sufficient. He explains "insufficiency" by the following example: "If two physically identical coins (as much as can be) are tossed 100 times, it is highly likely that one of them will yield more heads than the other. Similarly, in the absence of selection, *drift* will occur. Therefore, change (in gene frequencies or in phenotype distribution) is by no means indicative of selection. What is needed more than change to invoke selection is *differential adaptedness* to a *common selective environment*."

The Darwinian explanation of "*differential reproductive success*" is the *Principle of Natural Selection* (**PNS**). Brandon formulated the PNS principle as follows: "If A is better adapted than B in environment E, then (probably) A will have greater reproductive success than B in E." He has argued that the propensity interpretation of adaptedness (or propensity interpretation of fitness) renders the PNS explanatory [19].

In [18], *Brandon* discussed the following three issues of adaptation:

- Adaptation and environment,
- Adaptation and function,

- Hierarchical selection and adaptation.

**Adaptation and Environment** Regarding this issue, the question is why PNS compares entities A and B in a common environment E, and what does this mean? The answer that Brandon gives is that "to meaningfully compare genotypes, this must be done on a common environment." This is the foundation of the long-standing practice in biology of trying to carry out the comparison experiments of various organisms within a common environment (the so-called "*common-garden experiments*"). Moreover, Brandon indicates that "*cumulative adaptive evolution*" requires the existence, indeed the persistence, of common selective environments. The concept of environment is taken to be the twin of the concept of adaptation. The concept of selective environment has two principal virtues: The first is *operational* and the second is *factual*, i.e., this notion is more relevant to population genetic theory than the simple-minded notion of environmental heterogeneity. Some researchers consider that the concept of selective environment is too narrow to be relevant to evolutionary biology. Brandon's answer to them is that "the existence of stable adaptations (whether at the phenotypic level or genetic level) very strongly implies the long-term persistence of selectively homogeneous environments, or at least homogeneous with respect to the trait variants in question." Otherwise, there is no explanation for the long-term persistence of the trait.

**Adaptation and Function** Brandon pointed out that *change* (increase or decrease) in population is not sufficient to indicate that selection has taken place. This indication is secured through the concept of *drift* for which he developed the "*principle of drift*" that he calls "*biology's first law*" [20]. He calls the products of an adaptation, performed through evolution by natural selection, "*adaptations*" and argued with the help of two examples that adaptations have "*functions*". These functions represent their effects that make them adaptively superior to the trait variants with which they compete. However, knowing that some natural selection products are adaptations does not automatically help us to see what types of functions they represent. This can be done if we have available the *ecological explanation* of the adaptive superiority of the trait variant under study, relative to its competition and relative to the corresponding selective environment. This ecological explanation cannot be drawn (directly or indirectly) from the statistical techniques used to see if selection occurs, and so (particularly at the molecular level) we may know that something is an adaptation and hence it has a function, without being able to identify this function. The difficulties in constructing the ecological explanations that are required to identify adaptation functions are discussed in [19].

**Hierarchical Selection and Adaptation** The selection and adaptation process take place on several levels, i.e., it has a *hierarchical* nature. For example, a purely *genic* account may not possibly represent adaptation in nature. Brandon [18] supports this argument by referring to the works on *genic selection* of Oral and Criek [21] and Doolittle and Sapienza [22]. When they first discovered the phenomena of "*repetitive genomic sequences*", they wondered how the organism could benefit from these repetitive sequences. Because no convincing answer was found, they

concluded that the benefit to the organism was not the issue and abandoned the phenotype paradigm. This is because the process that produced the repetitive sequences was not a selection process among organisms, but rather an intra-cellular process of a DNA bit copying itself and inserting this copy somewhere else in the genome. In this way, this bit out-competes bits of DNA not doing that or doing it at a lower speed. Actually, this is selection taking place at a lower level and has to be understood in terms of the benefits that are added to the entities competing on that level. The benefit at the organismic level is not in the first instance the issue, since selection occurs, and can occur, at multiple levels. A good sequence of DNA may, or may not, be good for the organism to which it belongs. Therefore, any successful theory of adaptation must appreciate this issue and must be hierarchical [18].

John Stewart (1977) in [23] showed how evolution tunes the content and frequency of genetic variation to enhance its evolvability. This means that natural selection acting on a genetic system is not only finding adaptations but also discovering more effective ways to find adaptations. He states that genetic evolution is not random or fully blind. Genetic systems work in ways similar to nervous systems and brains, i.e., they are organized and structured by evolution so as to enhance their capability to discover more effective adaptations. This is achieved by producing a pattern of variation that is highly differentiated and specialized across the genome. This variation is highly hierarchical and can be seen as a set of hypotheses based on past experience about the likelihood that future change in particular characteristics will be worthwhile, and about the type of changes that are likely to be beneficial.

Stewart discusses in detail the following issues [23]:

- Impediments to the evolution of genetic cognition.
- How to overcome impediments to the evolution of genetic cognition.
- The evolution of mutational systems.
- Evolution of genetic systems that include recombination.
- The evolution of adaptive hierarchies within genetic systems.
- The evolution of sexual reproduction.

Species that can produce variation through mutation, recombination, and sexual combination out-compete those which fail to do so. This type of explanation has been extensively discredited since the early 1960s by several theoretical biologists who pointed out that, in most cases, a characteristic would not prevail in a population unless it benefits the individuals that exhibit the characteristic, no matter how much the characteristic benefits the population as a whole. This has revealed the special difficulties in the evolution of both cognitive arrangements and cooperation, despite the fact that both cognition and cooperation could surely be beneficial to the population.

Stewart [24, 25] has identified ways that can effectively overcome the cognitive limitations and the associated limitations in the evolution of cooperative organization amongst individuals. These ways involve the formation of higher level organizations (hierarchy of organizations). The hierarchical individuals feedback to

members of the group the effects to the organization of the cooperative actions of the members (thereby overcoming the cooperation limitation) and the expected future benefits of their cognitive and other actions (thereby overcoming the cognitive limitation).

*John Edward Terrell* (1999) examined the key issues of adaptation in *evolutionary archaeology* and answered the question: "do we need the concept of adaptation to archaeology?" He pointed out that adaptation has two principal roles in archaeology, namely [26]:

- The role it plays in our inferential efforts (desire) to "flesh out" what remains in us from the past, a way of giving artifacts and signs of human habitation meaning, purpose, and historical significance. The basic principle here is that, if we can define the *purpose* of *adaptive functions* of a thing, we have in a sense found its human soul.
- The role closely related to the fundamental functionalist thesis that artifacts and other traces of the past carry with them evidence of their own design or purpose in the scheme of things. Using the meaning of adaptation as the "act or process of adapting", this role is actually dynamic. That is, we do not restrict ourselves to asking the question: "What was this? (i.e., what did this do?)" but we progress to the question "What difference did this make?"
- *Terrell* points out that this second role of adaptation to archaeology is more interesting. This is due to the general agreement of social scientists (based on the evolutionary hypothesis) that what people end up doing is shaped by what they can do and by how successfully they do it. He stated that our apparent complex behavior is primarily a reflection of the environment in which we live. To explain the observed variation in the archaeological record (i.e., what people have done), we must address this natural "*logic*" (or "*design)*" structuring our actions.

It is worthwhile to that the mechanism for survival of human beings living under adverse conditions (e.g., undernutrition, poverty/economic oppression, exposure to infection, heavy workloads, etc.) is the so-called *predictive adaptive response* (**PAR**) mechanism, which includes two levels of adaptation, viz., (1) short-term adaptive response for immediate survival and (2) predictive responses required to secure post-natural survival to reproductive age. This mechanism is evaluated in comparison with the so-called "*developmental programming*" in [27], where a life history analysis is made in which the DP versus PAR debate from disease or adaptation is called "trade-offs". Even under good conditions, the stages of human life history are replete with trade-offs for survival, productivity, and reproduction. Under adverse conditions, trade-offs lead to reduced survival, poor growth, constraints on physical activity, and poor reproductive outcomes.

## 8.5 Adaptation Measurement

A comprehensive discussion of the problem of measuring adaptation can be found in [1]. The basic biological adaptation criterion is Darwinian "*fitness*", which over the years has been defined as "the ability of an individual to produce viable offsprings, or of an interbreeding population to reproduce itself." From the middle of the twentieth century until the resent time, the principal measure of adaptation is the so-called "*selection coefficient*", which quantitatively measures the *relative reproductive efficacy* of various genotypes in breeding populations (demes). This rigorous quantitative measure, which has been extensively used in microevolutionary laboratory and field research, until recently was not in use in larger evolutionary processes, such as the sociocultural evolution of human kind. Currently, biologists recognize the complex relationship between adaptation at the micro-level (individuals) and at higher levels of trait groups, social organizations, demes, ecological communities, species, etc. Actually, the process of adaptation is in most cases multidimensional and hierarchical (multilevel). The measures of adaptation at these higher levels of human evolution involve the following measures of human well-being:

- Economic measures,
- Social measures,
- Political measures,
- Physiological measures, and
- Health measures.

*Peter Corning* [11] reviewed several approaches to defining and measuring adaptation in humans, of which we mention here the following subset [28–35]:

- Amount of energy capture (in various cultures),
- Efficiency of energy capture (cost–benefit ratio),
- Struggle for satisfactions,
- Adaptation in goal-oriented behavior,
- Adaptive potential,
- Primary goods,
- Optimality of population size,
- Environmental quality, and
- Moral and ethical measures (moralnet).

We mention that the *moralnet* of Raoul Naroll [31] offered a normative list of "*indicators*" that can monitor the ongoing condition in "*socionomics*", as he tried to provide a policy/planning tool for the "creation of a stable world order." His moral indicators include suicide, divorce, child abuse, alcoholism, drug abuse, mental illness, crime, social discrimination, etc. The first two measures (energy) will be discussed in Chap. 10, and some of the other measures in Chap. 13 (see also Sect. 1.4.2).

## 8.6 Complex Adaptive Systems

### *8.6.1 General Issues*

The field of *Complex Adaptive Systems* (**CAS**) is concerned with the study of high-level instantiations of natural and man-made systems that cannot be studied by traditional analytical techniques. A CAS is a complex, self-similar set of interacting agents, and its study requires a multidisciplinary methodology. Examples of complex adaptive systems are rain forests, ant colonies, ecosystems, the biosphere, the immune system, the brain, the stock market, the global economy, large computer networks, a human society or community, a manufacturing enterprise, etc. *John Holland*, one of the inventors of evolutionary computation and genetic algorithms, defines a *Complex Adaptive System* as many agents that act in parallel, act persistently, and react to the actions of other agents. A CAS is controlled by decentralized and distributed controllers. The performance of a CAS is the result of the competition and cooperation between the agents, which involves a large number of decisions being made all the time by the individual agents [36–38]. Four basic goals (purposes of existence) of a CAS are the following:

- To understand the ways in which complex adaptive systems work and the unified underlying principle of complex adaptive behavior in life and society.
- To compare natural and man-made manifestations of CAS and reveal their possible differences.
- To study and understand the interplay of behavior at different scales, and reveal what is universal or not, and when averaging is applicable or not.
- To study and evaluate computer simulations of simplified models of natural systems and their actual importance for human life applications.

The term "*complex adaptive systems*" involves three words: complex, adaptive, and systems:

- *Complex* means composed of many parts which are joined together.
- *Adaptive* has the meaning(s) explained in Sect. 8.2, i.e., it refers to the fact that all living systems adapt (dynamically) to changing environments in their effort to survive and thrive.
- *System* is the concept which implies that everything is interconnected and interdependent.

Contrary to nonadaptive complex systems (e.g., the weather system), CAS has the capability to internalize information and to learn and to evolve (modify their performance) as they adapt to changes in their environments. A CAS produces macroscopic global patterns that emerge from the agents' nonlinear and dynamic interactions at the microscopic (low) hierarchical level. These emergent patterns are more than the sum or aggregation of their parts. Here is exactly the reason why the conventional reductionist approach cannot describe how the macroscopic patterns emerge.

According to Holland, the following three reasons explain why a CAS is difficult to be studied by the conventional system's theory approach [39]:

- A CAS loses the majority of its features when the parts are isolated.
- A CAS is highly dependent on its history and so the comparison of instances and identification trends is difficult.
- A CAS operates far from global optimum, and equilibrium points that concern the system "end points" are difficult to grasp by conventional approaches.

Fundamental higher level features that differentiate a CAS from a pure multi-agent system (MAS) are as follows:

- Complexity,
- Self-similarity,
- Emergence, and
- Self-organization.

**Complexity** is characterized as the "edge of chaos" since it lies in the middle region between two extremes, viz., static order and chaos [36, 40, 41]. A discussion of complexity is provided in Sect. 8.9. The self-similarity property will be discussed in relation to fractals in Sect. 8.8. The emergence concept will be outlined in Sect. 8.10, and self-organization is the subject of the next chapter. An MAS is simply composed of multiple interacting agents, whereas in a CAS the agents and the overall system are adaptive, and the system is self-similar.

According to Holland, a unified theory of CAS must incorporate at least three mechanisms [39]:

- Parallelism,
- Competition, and
- Recombination.

On the basis of this, the basic structure of a CAS has the form shown in Fig. 8.1.

Also, in order for a system to be characterized as a CAS, it must have the following seven characteristics (properties and mechanisms):

**Properties**

- Aggregation (complexity emerges from interaction of smaller parts).
- Nonlinearity (agents interact dynamically and nonlinearly).
- Flows (network of agent interactions).
- Diversity (agent evolution goes toward filling diverse niches, whereby niche evolution has larger impact on the system than the evolution of agents).

**Mechanisms**

- Tagging (agents have ways to discriminate agents with particular properties).
- Internal models (agents change via their interactions, and the changes specify future action, i.e., agents adapt).
- Building blocks (reuse of components for multiple purposes).

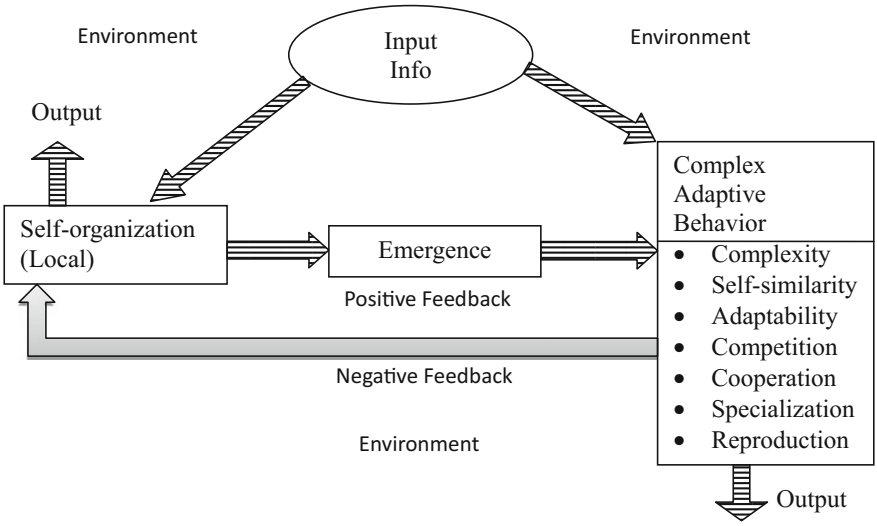



**Fig. 8.1** Basic structure of a complex adaptive system

### *8.6.2 A Concise Definition of CAS*

The definition of a CAS given by Holland (Sect. 8.61) is just one of the many alternative definitions available in the literature. Our purpose here is to provide a concise (nominal) definition of a CAS that attempts to capture all the features involved in the definitions given by Prigogine and Stengers [42], Jantsch [43], Maturana and Varela [44], and Holland [36]. This definition was composed by Kevin Dooley in 1996 [45] and merges into one master list all of the conceptual lists (and principles) involved in the above works.

**Definition of CAS** (*Dooley*)

This definition involves the following four aspects:

1. A CAS is composed by *agents* (as basic elements) which are semiautonomous units that try to maximize their fitness by evolving over time. Agents scan their environment and develop *schemas* (schemata), i.e., mental templates that interpret reality and define the proper responses to given stimuli. These schemas are subject to change and evolution via a selection–enactment–retention process in their effort to survive in their competition environment.
2. When a mismatch between observation and expected behavior is detected, an agent can take action in order to adapt the observation to fit an existing schema. An agent can also alter schemas as it likes to better fit the observation. Schemas can change via random or purposeful mutation and/or combination with other schemas. Changes of schemas lead to agents that are more robust to increasing variety or variation, adaptable to a broader range of conditions, and more predictable in performance.

3. Schemas define the way an agent interacts with its neighboring agents, whereby actions among agents involve a nonlinear exchange of information and/or resources. Agent tags help to identify what other agents are capable of transactions with a given agent and also facilitate the formation of meta-agents that distribute and decentralize functionality so as to allow diversity to thrive and specialization to occur. Agents can also reside outside the bounds of a CAS, and schemas associated with them can also define the rules of interaction and resource/information flow externally.
4. The agents' fitness involves in a complex way several local and global factors. Unfit agents are likely to initiate changes of schemas. The optimization of local fitness enables differentiation and novelty/diversity. Global fitness optimization leads to improved CAS coherence (as a system) and induces long-term memory.

## 8.7 List of Reference Works on Complexity, Complex Systems, and Complex Adaptive Systems

The field of CAS is about 25 years old and has its origin in the research work carried out by an interdisciplinary group of scientists (Murray Gell-Mann, John Holland, Brian Goodwin, and others) at the Santa Fe Institute (New Mexico, USA). This field emerged from the discipline of complexity and general complex systems, chaos theory, and adaptation in biological, ecological, and human societies and organizations.

Beginning in the early twentieth century, scientists started abandoning deterministic Newtonian mechanics and developed more general physical laws (e.g., Werner Heisenberg's uncertainty principle, Albert Einstein's relativity theory according to which time is relative, space is curved, and matter energy are interchangeable). From about the 1950s, leading researchers in biological systems have realized that the so-called *reductionism* (i.e., examining the parts of a system to understand their overall performance) is not sufficient for real-world systems (living organisms, ecosystems, economics, meteorological systems, human societies, etc.) because they are very complicated. This fact has motivated research work into general complex systems theory and has formed the foundations of complexity theory and complex adaptive systems of today.

Our purpose in this section is to provide a representative (non-exhaustive) chronological list of works in this field [46–98], in addition to those presented in [39–45, 99]. These works are classified into the following two major categories:

1. **Nonlinear Systems, Chaos, and General Systems**
   Maxwell (1873) [46], Lorenz (1963) [48], von Bertanlaffy (1968) [49], May (1976) [51], Nicolis and Prigogine (1977) [52], Mandelbrot (1977) [53], Ackoff (1978) [54], Smith (1982) [56], Devaney (1986) [57], Parker and Chua (1989) [58], Langton (1990) [60], Hayles (1991) [61], Kauffman (1991) [62], Lewin

(1992) [63], Devanay (1992) [64], Mitchell, Hraver and Crutchfield (1993) [65], Ott (1993), Wolfram (1994) [69, 70], Holland (1995) [36], Gell-Mann (1995) [72], Heylinghen (1996) [78], Fitzerald (1996) [76], Bar-Yam (1997) [80], Cilliers (1998) [83], Lissack (1999) [84], Segel (2000) [85], Buchanan (2000) [86], Bar-Yam (2000) [88], and Bar-Yam and Minai (2002) [91].

2. **Complex Evolutionary and Complex Adaptive Systems**
Holland (1975) [50], Smith (1982) [56], Goldberg (1989) [59], Holland (1992) [37], Kauffman (1993) [66], Gell-Mann (1994) [71], Morowitz and Singer (1995) [74], Per (1996) [75], Mitchell (1996) [77], Langton (1997) [79], Holland (1998) [40], Sigmund (1998) [81], Bonabeau (1998) [82], Smith and Bedau (2000) [87], Sawyer (2001) [89], Harris (2001) [90], Levin (2003) [92], Harkerma (2003) [93], Goldstone, Sakamoto (2003) [94], Bullock, Cliff (2004) [95], Holden (2005) [96], Harre (2006) [97], and Kauffman and Clayton (2006) [98].

## 8.8 Chaos and Nonlinear Systems

### *8.8.1 Fractals and Strange Attractors*

According to Huajie Liu (1999) [100], the discipline of chaos had its origin on *Kolmogorov's* theorem of conservative systems (1954) [101], later proved by Vladimir Arnold (1963) [102] and Jürgen Moser (1962) [103] which is now known as *Kolmogorov's* KAM theorem from the initials of the surnames of the above three scientists and in *Edward Lorenz's* model of weather via dissipative systems (1963) [48]. The KAM theorem gives the conditions which assure that chaos is bounded in phase space by a "*doughnut-like*" surface. Lorenz's weather model has led to the so-called "*Butterfly Effect*" according to which a butterfly flapping its wings in China could influence weather conditions (air currents) on a large scale and lead to windstorms in the USA (say a month later).

Chaos can appear in both space and time. In space, a chaotic system is known as "*fractal*", which is a geometrical shape which does not change when it is analyzed in smaller parts continuously (i.e., fractals are not smooth). Similarly, simple fractals when emerged (coming into existence and developed further) always reproduce the same shape (or structure), i.e., spatial fractals are *self-similar*.

Nature is full of fractals. We actually live within fractals. The human body, a tree and its branches, the sky on a semi-cloudy day, the ocean's surface waves, the coast of England, avalanches, and so on all are spatial fractals. Simple examples of fractals are given in Fig. 8.2 where $D$ is the spatial dimension of the fractal ($D = 1$ for a straight line segment, $D = 2$ for a plane shape, $D = 3$ for a 3-D space shape, etc.)

The dimension $D$ of a fractal is given by the following relation:

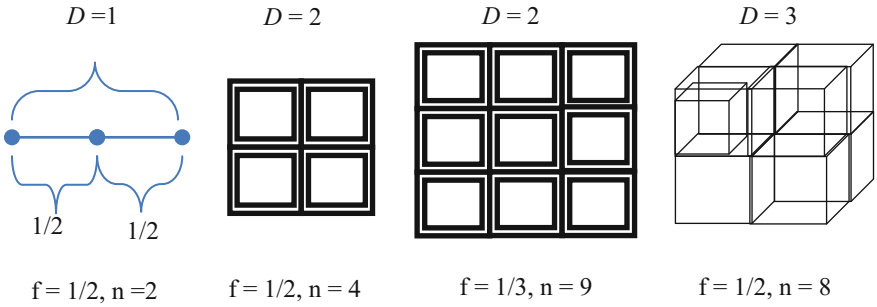



**Fig. 8.2** Spatial fractals (f = subdivision ratio, n = number of fractals)

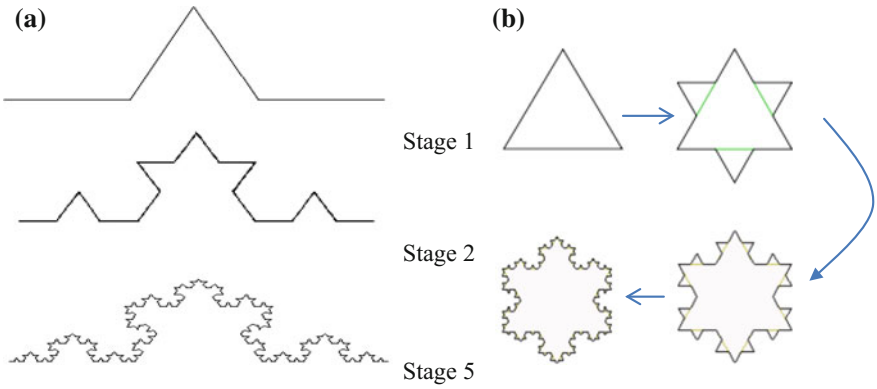



**Fig. 8.3** **a** Generation of Koch's "snowflake" fractal, **b** Application of Koch's fractal scaling to the sides of an equilateral triangle [50]

$$n = \left(\frac{1}{f}\right)^{D} \text{ or } D = \frac{\log n}{\log(1/f)} = \frac{-\log n}{\log f}$$

The "*snowflake*" fractal of Koch (Fig. 8.3) is constructed by replacing, at every stage, each straight line segment by a broken line segment of a length 4/3 as long. The results of stages 1, 2, and 5 are shown in Fig. 8.3a. The dimension of this fractal is

$$D = \frac{\log 4}{\log 3} = 1.26185.\ (1 < D < 2)$$

If we apply this scaling to each one of the sides of an equilateral triangle, we obtain, after three steps, the region shown in Fig. 8.3b. At each step, the perimeter of the region is increased by a factor 4/3. In the limit, the factor $(4/3)^{N}$ increases without bounds, but the area enclosed by the curve is finite, equal to

$$\frac{\sqrt{3}}{4}+\frac{\sqrt{3}}{4}\left(\frac{1}{3}+\frac{4}{3^3}+\frac{4^2}{3^5}+\frac{4^3}{3^7}+\cdots\right)=\frac{2\sqrt{3}}{5}$$

i.e., equal to 8/5 the area of the initial triangle.

Similarly, the tree fractal has the form shown in Fig. 8.4a. Figure 8.4b shows a realistic-looking plant fractal generated via a Lindenmayer Grammar (LG) when adding morphological perturbators.

In Fig. 8.4a, for the branches to not overlap, we must have in the following limit:

$$f\ \cos 30^0 = f^3\ \cos 30^0 + f^4\ \cos 30^0 + f^5\ \cos 30^0 + \cdots$$

i.e.,

$$f = f^3 + f^4 + f^5 + \cdots = f^3\left(1 + f + f^2 + \cdots\right) = \frac{f^3}{(1-f)}$$

Thus,

$f^2 = 1 - f$ or $f = \frac{\sqrt{5}-1}{2} = \frac{1}{\phi} = 0.618$

where $\phi$ is the "golden division ratio" (well known from geometry), which is defined as follows. Take a straight line segment (Fig. 8.5).

We say that the point C is a golden division point of (AB) if

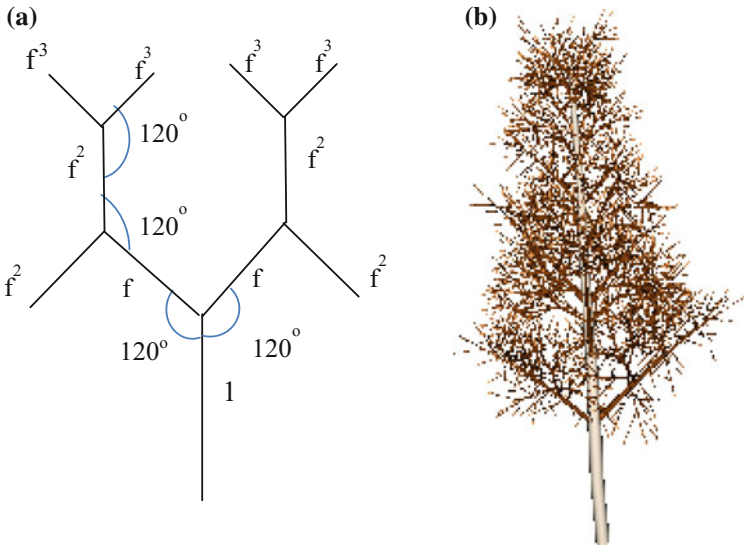



**Fig. 8.4** Branches of tree fractal. **a** Golden division-based tree fractal. **b** Realistic tree produced by LG

$\frac{1}{x} = \frac{x}{x+1}$ or $x^2 = x + 1$

which has the positive solution $x = \left(1 + \sqrt{5}\right)/2 = 1.618 = \phi$, and hence $1/\phi = 0.618$. Some other well-known expressions of $\phi$ are

$$\phi = \sqrt{1 + \sqrt{1 + \sqrt{1 + \cdots}}}, \phi^2 = 1 + \sqrt{1 + \sqrt{1 + \sqrt{1 + \cdots}}} = 1 + \phi$$

$$\frac{1}{\phi} = \frac{1}{1 + \frac{1}{1 + \cdots}} = \frac{1}{1 + \frac{1}{\phi}}, \text{ i.e., } \phi^2 = 1 + \phi$$

Actually, in addition to natural fractals, there exist a large number of man-made fractals designed by mathematicians, physicist, and computer scientists (e.g., those seen on animated movies, computational games, etc.). Four artificial fractals (cardial fractal, seahorse fractal, candle fractal, and tetrahedron fractal) and four natural fractals (butterfly fractal, broccoli fractal, and two flower fractals) are shown in Figs. 8.6, 8.7, 8.8, 8.9, and 8.10.

Now we turn our attention to the concept of *chaos in time* which appears due to the *"sensitivity of chaotic dynamic systems to initial conditions."* The motion equations of dynamic systems help us to compute their state at the next instant of time, and in all successive time instants, thus providing the *trajectory (orbit)* of the system in state space. Typically, we start from the state of the system at some initial time, i.e., from a given set of *initial conditions*. Consider a *chaotic system* (or *time chaos*) and two sets of initial conditions (i.e., two points in state space) arbitrarily close to each other. The two trajectories that correspond to these two sets of initial conditions are very close to each other at the beginning, but they do not stay close as time evolves. Actually, they diverge away from each other over time. This phenomenon, which as mentioned above, is called "*sensitivity to initial conditions*" and was discovered by *Edward Lorenz*, who called it the "*butterfly effect*". The probability of getting time chaos in nonlinear systems is extremely high. This is the rule rather than exception, although generations and generations of scientists were taught that systems are *integrable* (multi-periodic), i.e., insensitive to initial conditions. Thus, *chaos* is the *end of reductionism* as indicated in Sect. 8.7. This feature had been hypothesized already in 1873 by James Maxwell [46]. The term "*fractal*" was coined by Benoit Mandelbrot (1977) [53], who showed via computer graphics that *strange attractors* have a *fractal* property. Every chaotic dynamical system is a fractal-generating mechanism, and conversely, every fractal can be regarded as the outcome of long-time acting time chaos. Time chaos and space

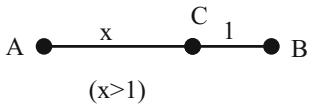



**Fig. 8.5** Definition of the golden division

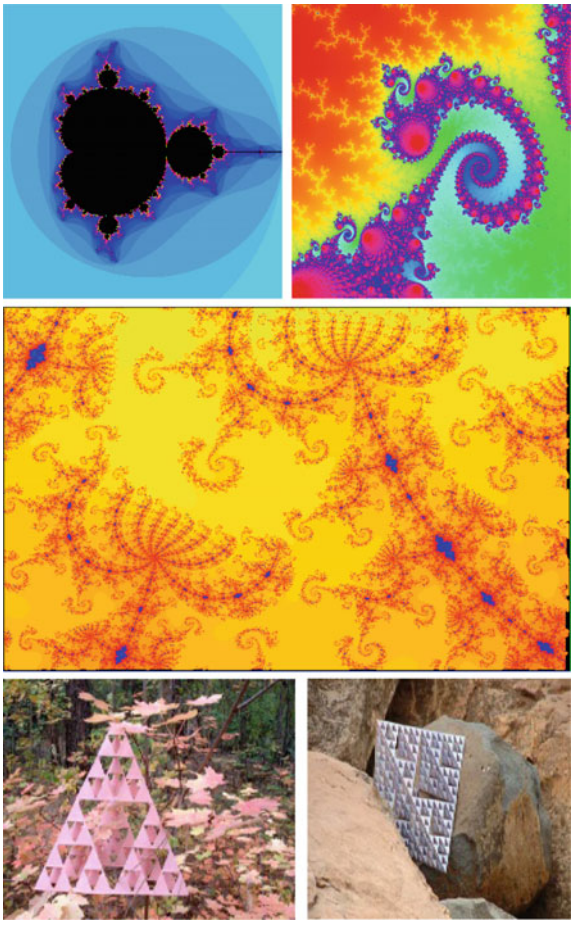


**Fig. 8.6** Artificial (computer generated) fractals. Seahorse fractal (http://www.superliminal.com/fractals/mbrot/mbrot.htm), Candle fractals (http://www.softsource.com/m_cndl.gif), Tetrahedron fractals (http://www.fractalnature.com). (The reader is informed that web figures and references were collected at the time of writing the book. So, some of them may not be valid now due to change or removal by their creators)

**Fig. 8.7** Butterfly fractal (http://www.mirrorofnature.org/BBookButterfllyFritillary.jpg)

**Fig. 8.8** Romanesco broccoli fractal (http://britton.disted.camosun.bc.ca/fractals/broccoli-fractal.jpg)

**Fig. 8.9** Camelia fractal

chaos are closely related. For example, let us take a simple region (e.g., a cube or a sphere, etc.) in the state space of a chaotic system. As each point of this region traverses its trajectory, the region itself moves and changes shape. As time passes, the region will tend to a fractal which becomes perfect (complete) at infinity.

Over the years, it has been realized that chaos may be a feature of very simple systems as well. We will illustrate this by examining Lorenz's atmospheric model

**Fig. 8.10** Dahlia fractal (http://cdn2.listsoplenty.com/listsoplenty-cdn/uploads/2010/01/Queen-Annes-Lace-a-fractal-flower-1024x967.jpg)

(attractor), which is derived using a simplified version of the Navier–Stokes fluid dynamics equations [104]:

$$\begin{aligned}\dot{x}_1 &= f_1(x_1, x_2, x_3) = -ax_1 + ax_2\\ \dot{x}_2 &= f_2(x_1, x_2, x_3) = -x_2 + rx_1 - x_1x_3\\ \dot{x}_3 &= f_3(x_1, x_2, x_3) = -bx_3 + x_1x_2\end{aligned}$$

An air layer is heated from bottom to top. The warm air that goes up interacts with the colder air that goes down and produces turbulent convention rolls. Here, $x_1$ is the rate of rotation of the convection rolls, $x_2$ is the temperature difference between the *upward* and the *downward* moving air masses, and $x_3$ is the deviation from linearity of the vertical temperature profile. We will study the performance of this system for $a = 10$ and $b = 8/3$. The parameter $r$ represents the temperature difference between the top and the base of the air layer. Increasing $r$, we add more energy to the system, which results in more vigorous dynamics.

The equilibrium points are given by the solution of the algebraic system:

$$-ax_1 + ax_2 = 0,\ -x_2 + rx_1 - x_1x_3 = 0,\ -bx_3 + x_1x_2 = 0$$

Obviously, the point $(x_1, x_2, x_3) = (0, 0, 0)$ is an equilibrium point (a solution). The first equation gives $x_1 = x_2$, and so the second equation becomes $x_1(-1 + r - x_3) = 0$. Thus, for $x_1 \neq 0$, we get $x_3 = r - 1$. Then, the third equation

becomes $x_1^2 = b(r-1)$. For $0<r<1$, this equation has complex-valued roots, and so the only equilibrium point is $(x_1, x_2, x_3) = (0, 0, 0)$. For $r = 1$, the equilibrium point is again $(x_1, x_2, x_3) = (0, 0, 0)$. Finally for $r > 1$, we have three equilibrium points:

$$P_0 = (0,0,0),\ P_1 = \left(\sqrt{b(r-1)}\right),\ \sqrt{b(r-1)}, r-1\Big),\ P_2 = \left(-\sqrt{b(r-1)}, -\sqrt{b(r-1)}, r-1\right)$$

To test the stability of these points, we find the dynamic system's Jacobian matrix at the point $P_0 = (0, 0, 0)$:

$$\mathbf{A} = \frac{\partial \mathbf{f}}{\partial \mathbf{x}} = \begin{bmatrix} -a & a & 0 \\ r - x_3 & -1 & -x_1 \\ x_3 & x_1 & -b \end{bmatrix} = \begin{bmatrix} -10 & 10 & 0 \\ r & -1 & 0 \\ 0 & 0 & -8/3 \end{bmatrix}$$

where $\partial \mathbf{f}/\partial \mathbf{x}$ denotes the partial derivative of the vector-valued function $\mathbf{f} = [1, f_2, f_3]^{\mathrm{T}}$ with respect to the vector-valued variable $\mathbf{x} = [x_1, x_2, x_3]^{\mathrm{T}}$. For $r > 0$, the matrix $\mathbf{A}$ has the eigenvalues $\lambda_{1,2} = (1/2)\left(-11 \pm \sqrt{81 + 40r}\right)$ and $\lambda_3 = -8/3$. For $0<r<1$, all these eigenvalues are negative and so the equilibrium point $(0, 0, 0)$ is a stable equilibrium point. If $r > 1$, then $\lambda_1 > 0$ and the point $(0, 0, 0)$ is an unstable equilibrium point. The eigenvalues of $\mathbf{A}$ in the cases of the two other equilibrium points $P_1$ and $P_2$ are of the same form (i.e., $\lambda_{1,2} = a \pm j\beta,\ \lambda_3 \leq 0$). The real part $a$ is negative for $1<r<r_0$ and positive for $r > r_0$ with $r_0 \approx 24.8$. Therefore, for $1<r<r_0$, there are two stable equilibrium points, whereas for $r > r_0$ all equilibrium points are unstable.

This means that for $1<r<r_0$ two neighboring trajectories return in a spiral way to the nonzero equilibrium points, while for $r > r_0$ they spiral outward.

The trajectories can be computed with the Runge–Kutta method (e.g., fourth-order RK method) and are sensitive to the parameters of the method (e.g., the step size). Figure 8.11a shows a view of the Lorenz attractor's trajectories, for $r = 28$, drawn by Jim Loy, using Paint Shop Pro. Figure 8.11b shows the trajectories of a second strange attractor.

Looking at the reference works list 1 of Sect. 8.7, we see that a large number of research efforts were made in the field of fractals, chaos, and chaotic systems, both of a theoretical and a computational nature. We mention here a few of the leaders in the field: I. Prigogine, R.L. Devaney, T.S. Parker, C.G. Langton, E. Ott, P. Coveney, Y. Bar-Yam, and R. Lewin. To summarize, complex nonlinear (chaotic) systems have the following features [80, 105, 106]:

- Very small differences in initial conditions lead to extremely different outcomes.
- Small inputs can cause catastrophically large outcomes and consequences.
- Global properties result from aggregate behavior of individuals (the whole is greater than the sum of its parts).

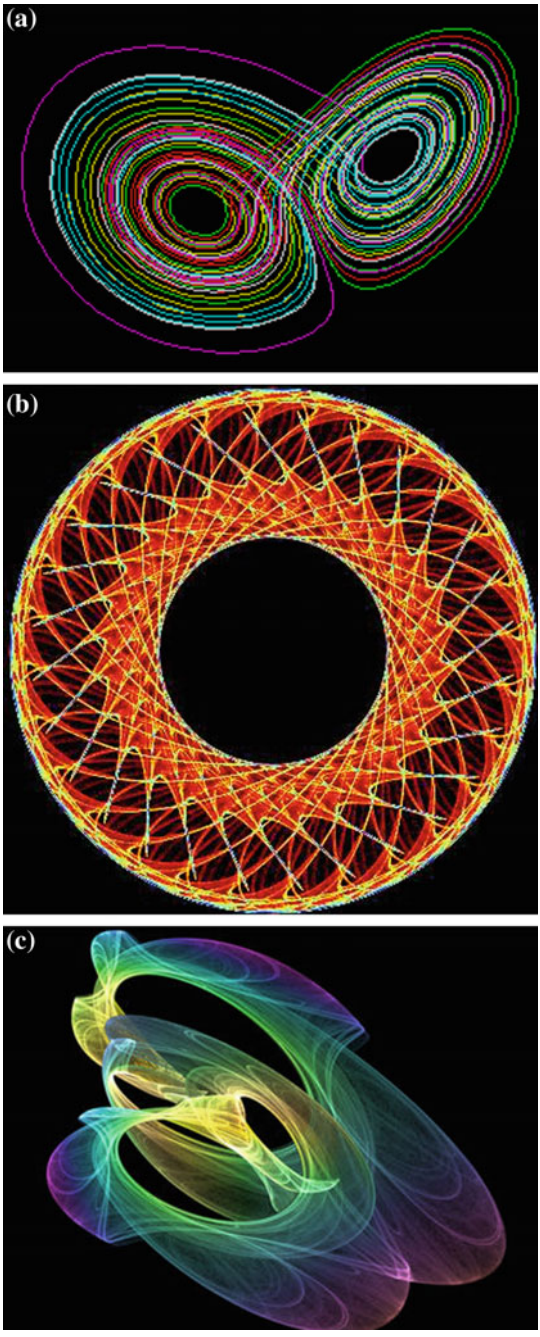



**Fig. 8.11** **a** A view of the Lorenz attractor trajectories obtained for a = 10, r = 28, b = 8/3, **b** & **c** Trajectories of two other strange attractors (**a** http://cu.imt.net/∼jimloy/fractals/lorenz.htm; **b** http://www.alunw.freeuk.com/attractor.jpg; **c** http://mathforum.org/mathimages/index.php/Lorenz_Attractor)

- In all cases, the fractals and chaos are manufactured by the nonlinearities via stretching and folding. Linear systems, if they start stretching, stretch forever; they never fold. It is the nonlinearity that folds.
- Fractals are self-similar, i.e., independent of scale.
- In most cases, there exist the so-called *attractors* (or *strange attractors*), i.e., states to which the system finally settles, depending on the particularities of the system.

*Strange attractors* result from trajectories that roam irregularly within a bounded region, without being repeated, but with neighboring regions that on, the average, diverge exponentially as time goes.

### *8.8.2 Solitons*

Another nonlinear natural phenomenon, besides that of a "*strange attractors*", is the "*solitary traveling-wave pulse*" observed in 1834 by *John Scott Russel*, a Scottish naval engineer, during his experiment to determine the most effective design for canal boats. These waves neither disperse nor break but continue traveling on indefinitely in the same shape and amplitude without fading away. He named this strange traveling-wave pulse a "*soliton*". Thus, a soliton is a special form of a surface water wave. He described this phenomenon as follows: "I was observing the motion of a boat which was rapidly drawn along a narrow channel by a pair of horses, when the boat suddenly stopped—not so the mass of water in the channel which it had put in motion; it accumulated round the prow of the vessel in a state of violent agitation, then suddenly leaving it behind, rolled forward with great velocity, assuming the form of a large solitary elevation, a rounded, smooth and well-defined heap of water, which continued its course along the channel apparently without change of form or diminution of speed. I followed it on horseback and overtook it still rolling on at a rate of some eight or nine miles an hour, preserving its original figure some thirty feet long and a foot to a foot and a half in height. Its height gradually diminished, and after a chase of one or two miles I lost it in the windings of the channel. Such, in the month of August 1834, was my first chance interview with that singular and beautiful phenomenon which I have called the Wave of Translation."

He viewed the solitary wave as a self- sufficient entity, a "thing" possessing many of the properties of a particle. In our day, it is considered as a constructive element in formulating the complex dynamic behavior of wave systems through science, including hydrodynamics, nonlinear optics, plasmas, shock waves, tornados, and elementary particles of both matter and thought.

The mathematical equation that describes the unidirectional propagation of solitons on a shallow canal was derived by Diederik Korteweg and Gustav de Vries in 1895 and is now known as "*KdV equation*". The *KdV* equation is a

two-dimensional, nonlinear partial differential equation (PDE) with variables of space $x$ and time $t$, of third order with respect to $x$, namely

$$\frac{\partial y(x,t)}{\partial t} + \frac{\partial^3 y(x,t)}{\partial x^3} - 6y(x,t)\frac{\partial y(x,t)}{\partial x} = 0$$

The constant "6" in front of the nonlinear term is used for convenience but is of no significance, since if $y(x,t)$ is a solution of the *KdV* equation, then the function:

$$y(x,t) = \frac{v}{2\cosh^2\left[\frac{1}{2}\sqrt{v}(x - vt - \phi)\right]},\ c > \mathbf{0}$$

where $v$ is the *phase speed* of the wave and $\Phi$ is a *phase difference constant*. This equation describes a right-moving soliton. One can observe that the peak amplitude is exactly half the speed, i.e., larger solitons have greater speeds. A snapshot of a single-soliton wave is shown in Fig. 8.12a. The *KdV* equation permits also a double soliton or multi-soliton. Figure 8.12b shows a snapshot of a double soliton (with separated centers and different amplitudes), which has the following form:

$$y(x,t) = 2\frac{\partial^2}{\partial x^2}\ln\left(1 + c_1 e^{\theta_1} + c_2 e^{\theta_2} + Bc_1 c_2 e^{\theta_1 + \theta_2}\right)$$

$$\theta_1 = a_1 x - a_1^3 t, \quad \theta_2 = a_2 x - a_2^3 t, \quad B = (a_1 - a_2)^2/(a_1 + a_2)^2$$

where $c_1$, $c_2$, $\alpha_1$, and $\alpha_2$ are the arbitrary constants.

Two natural soliton waves occurring in a body of water (a), or in the Strait of Gibraltar (b), are shown in Fig. 8.13.

Solitons take place when opposing linear effects of *dispersion* are perfectly balanced with the concentrating effects of *nonlinearity*. The soliton is lost if one of

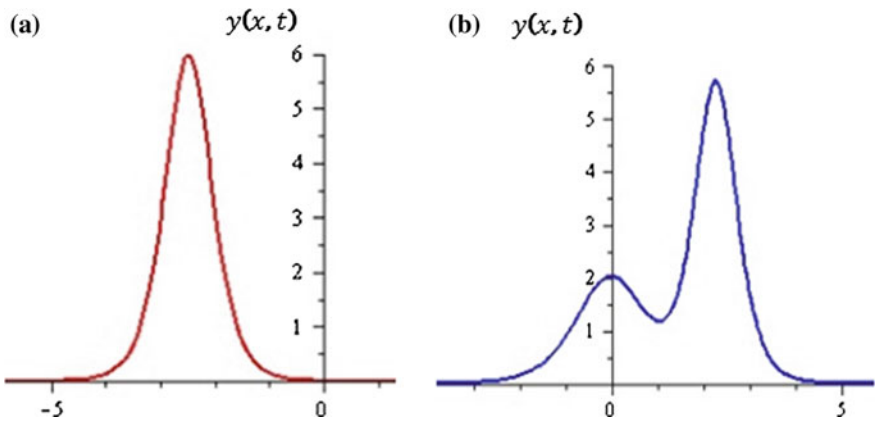



**Fig. 8.12** **a** Single soliton, **b** Double soliton

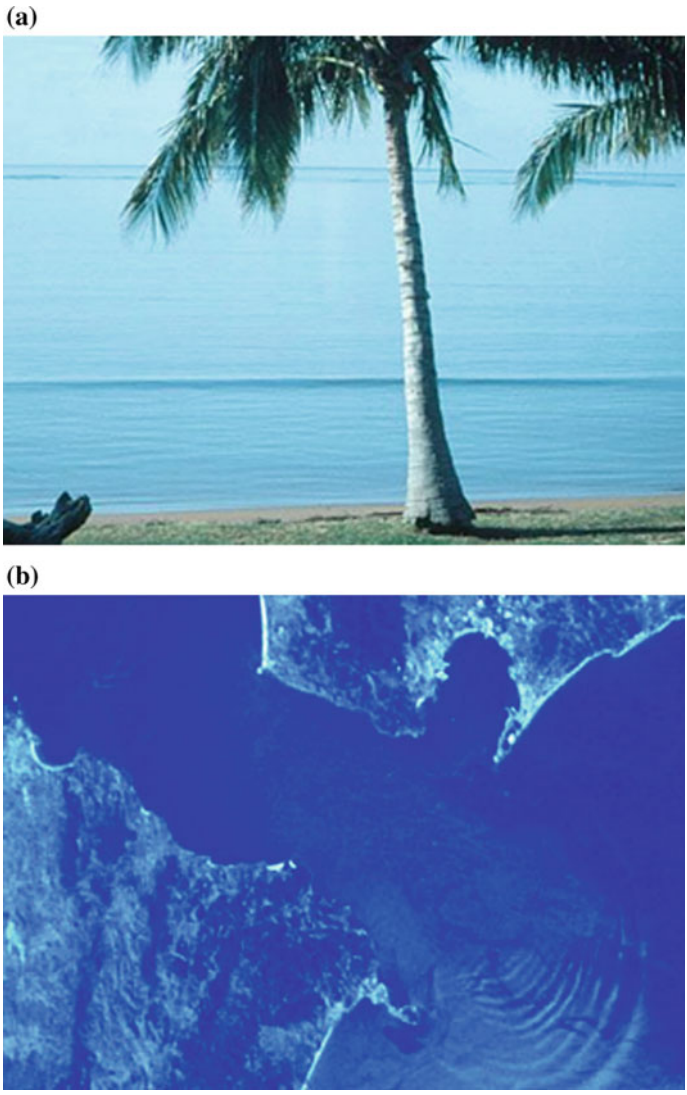



**Fig. 8.13** Two natural soliton waves: occurring on a body of water **a** and in the Strait of Gibraltar **b** (**a** http://www.technovelgy.com/graphics/content/Soliton-Kuaii.jpg, **b** http://www.lpi.usra.edu/publications/slidesets/oceans/images/ocean13.jpg)

the above two competing effects is lost. Solitons can occur in any medium involving the key ingredient of nonlinearity (e.g., solids, liquids, gases, fiber optics, etc.). Tidal bores on rivers are good examples of solitons. A more complex question is whether tsunami waves involve solitons. Typically, tsunami waves produced by sharp localized impulses, such as meteorite strikes into the sea, involve solitons. In

general, however, it is difficult to check models because tsunamis are almost impossible to observe in mid-ocean before they are modified by coastal effects. Sea-level measurements have shown that tsunami waves, unlike solitons in the strict sense, have both peaks and troughs. Another class of ocean waves in the category of so-called rogue waves (sometimes called extreme waves, abnormal waves, or Luller waves). These are very large and occur spontaneously far out at sea and are a threat even to large ocean liners. It must be noted that rogue waves are not tsunamis, which are initiated by earthquakes and travel at high speed, assuming greater height as they approach the shore.

## 8.9 Complexity

*Complexity* has not been, and cannot uniquely and precisely be defined. Its definition is also complex. Complexity is inherent in life, science, technology, society, and human performance in general. From the level of biomolecular interactions to human peace-and-war cycles, complexity is the major aspect and has to be faced: The basic question is whether there is a trend toward greater complexity over time among living beings, in nature, and in human activity. This can only be answered if we find a global measure of the world's complexity. This seems to be very difficult, if not impossible, but over the years many relevant measures of complexity have been discovered or developed. Complexity, like chaos, has its origin in the "*hard sciences*" (mathematics, physics, chemistry, biology). The soft (social and management) sciences joined in at a later stage in its development.

> *Francis Heylinghen* stated the following about complexity [78]: "*Complexity has turned out to be very difficult to define. The dozens of definitions that have been offered all fall short in one respect or another, classifying something as complex which we intuitively would see as simple, or denying an obviously complex phenomenon the label of complexity. Moreover, these definitions are either only applicable to a very restricted domain, such as computer algorithms or genomes, or so vague as to be almost meaningless.*"

However, although an exact and complete definition of complexity cannot be given, it is useful to identify the list of complexity's characteristics which may be improved and/or extended as our knowledge about the world, life, and humans increases. Such a list (surely not complete) involves the following properties of complex systems [88, 107]:

**P1** *Complex systems are made from several parts that interact nonlinearly.*

Nonlinearity is necessary for a system to be chaotic but not all chaotic systems are complex. There is a big difference between complexity and chaos. Complexity spans several scales, but chaos may occur only at one scale.

**P2** *The parts of a complex system are interdependent.*

For example, consider a "liquid quantity in a pot". If we take out a part (say 15%) of this quantity, nothing more happens than a reduction of the liquid volume by 15%. This is so because the system is noncomplex. But, if we remove 15% of an

animal's body (e.g., by cutting a leg), the final result will be very critical. This is because the body of an animal is a complex system, the parts of which are interdependent.

**P3** *Complexity exhibits a balance between chaos and non-chaos.*

Heylingen [78] has placed complexity between "*order and disorder*" or as it more often is called "*on the edge of chaos*". However, it is not fully clear what this does mean. For example, as illustrated in the Lorrenz attractor, a system may be non-chaotic for values of a control parameter less than a critical value, and chaotic for larger values of this parameter. Clearly, this critical value can be regarded as the edge of chaos. But perhaps complex systems (e.g., living organisms) may manage to operate as much as possible at such an "edge of chaos" (probably through self-organization), thus achieving a balance or coexistence of chaos and order (non-chaos).

A system that combines **chaos** and **order** was named by Dee Hock [108] **"*chaordic*"**. A concise definition of a chaordic system was given by Laurie Fitzerald [109] and is "Chaordic system is a complex and dynamical arrangement of connections between elements forming a unified whole, the behavior of which is both *unpredictable* (chaotic) and *patterned* (orderly)… simultaneously".

**P4** *Complexity involves several paradoxes.*

This coexistence (or interplay) of two properties (chaos and order) is actually a "paradox", which in simple systems does not occur. In complex systems, we may have many such paradoxes, i.e., coexistence, balance or interplay of properties that in simpler contexts are incompatible. Examples of such pairs of properties are as follows:

- Universal and unique,
- Selfish and altruistic,
- Orderly and flexible,
- Ordered and disordered,
- Persistent and dynamic,
- Controllable and uncontrollable,
- Softer and stronger,
- Random and predictable,
- Logical and paradoxical,
- Adaptable and nonadaptable, and
- Competitive and cooperative.

One can find many cases in which each one of the above pairs of properties coexists or balances. For example, the interplay between the last pair of properties (competition and cooperation) is a basic feature of all populations of organisms and social systems that are subject to the laws of evolution. We can also observe that this can be an interplay between system scales. Usually, competition on scale N is nourished by cooperation on the finer scale below (scale N + 1). Insect colonies (ants, bees, etc.) are spectacular examples of this. In many game-like systems, the best solution is found by a mix of cooperative and competition policies [110, 111].

This illustrates that evolution is not only governed by the law (cliché) of the "*survival of the fittest*".

**P5** *Complex systems involve the multicausal paradox.*

This is the most direct paradox exhibited by complex systems and refers to the fact that, in complex systems, there may exist more than one "cause" that lead to the same outcome, i.e., A causes B and C causes B are not mutually incompatible statements (processes). Understanding this and other paradoxes in complex systems enhances our knowledge capacity by enabling us to conceive of the diversity of possibilities in understanding the world and in designing technological systems that are not restricted to conventional perspectives.

**P6** *Complex systems are coherent.*

Michael Lissack and Johan Roos [112] were concerned with the role of complexity in management, and a central concept in their approach to complexity is the concept of *coherence*. Coherence is defined as "*an alignment* of *context, viewpoint, purpose* and *action that enables further purposive action.*" Lissack and Roos state that coherence plays a key role in leadership as follows [112]:

> "*Finding coherence, enabling coherence, and communicating coherence are the critical tasks of management in the era of knowledge society. We call this mastering complexity through coherence. We have to have an understanding at a level separate from the actions. Such an understanding will encompass both purpose and identity. By purpose we mean that reason for being or doing: Why am I doing what I am doing? By identity we mean an evolving, moving intersection of the inner and outer forces that make each of us who we are…*"

We close this section with a brief discussion to the measures of complexity. Scientists in all fields facing difficult problems involving complexity have asked over time the following major questions about their problems, processes, and systems:

- How hard is it to describe?
- How hard is it to create?
- What is the degree of organization?

The measures developed are grouped according to these questions and are the following [113]. These measures are typically similar and closely related quantities. Due to space limitations, we will not present the details of them, but for a few of them we have already presented their main characteristics.

1. *Difficulty of description:* This is typically measured in bits using the following measures: information, entropy, algorithmic complexity (or algorithmic information content), minimum description length, Fisher information, Renyi entropy, code length, Chernoff information, dimension, fractal dimension, and Lempel–Ziv complexity.
2. *Difficulty of creation:* This is typically measured in time, energy, money, etc., using the following measures: computational complexity, time computational

complexity, space computational complexity, information-based complexity, logical depth, thermodynamic depth, cost, and crypticity.

3. *Degree of organization:* Here, use is made of two groups of measures: effective complexity (difficulty of describing organizational structure: chemical, cellular, managerial, etc.) and mutual information (i.e., the information shared between the parts and the outcome of the organizational structure).

   (a). Measures for *effective complexity* are grammatical complexity, hierarchical complexity, conditional information, schema length, stochastic complexity, excess entropy, fractal dimension, metric entropy, etc.
   (b). Measures for *mutual information* include organization, correlation, channel capacity, stored information, and algorithmic mutual information.

According to Buchanan [86], the principal difference between *chaos* and *complexity* is their history, in the sense that chaotic systems do not rely on their history whereas complex systems do.

According to Roger Lewin [63], the science of complexity has to do with structure and order, especially in living systems, social organizations, the development of embryo, patterns of evolution, ecosystems, business and nonprofit organizations, and their interactions with the technological–economic environment.

According to Chris Langton, [60], "You can only understand complex systems using computers, because they are highly nonlinear and are beyond standard mathematical analysis."

Langton investigated the *edge of chaos* in *cellular automata*, trying to determine the conditions under which a simple cellular automaton could support "*computational primitives*, i.e., entities that can transmit, store, and modify information. He performed experiments on a cellular automaton with 128 cells in a circular structure, whereby each cell was taking as input its own state and the states of the two neighbors on each side. These five cells are its neighborhood, which is associated with transmission. The automaton's internal state is associated with storage and the transition function from state to state with modification of information. To examine how order and chaos influence computation, Langton introduced a parameter value $\lambda$ which gives the probability that a neighborhood state will push the call to a "*frozen*" (quiescent) state. For $\lambda = 0$, all neighborhood states go to the quiescent state and the whole automaton is frozen, i.e., *completely ordered*. As $\lambda$ increases from zero, the cellular automaton's transitions provide the so-called transient streams (cell transition sequences), which as $\lambda$ tends to 1 break down and *disperse*. Transients represent the ability of the cellular automaton to compute and, when graphed as two-dimensional series, indicate the qualitative concept of *complexity*. The above experiment showed that computation is possible at the *edge of chaos*, i.e., in a thin region of complexity between order and chaos where self-organization could be possible. Actually, what Langton has successfully done is to tune $\lambda$ to the critical value for which the cellular automaton is not too ordered and not too chaotic, but just capable of having complex behavior.

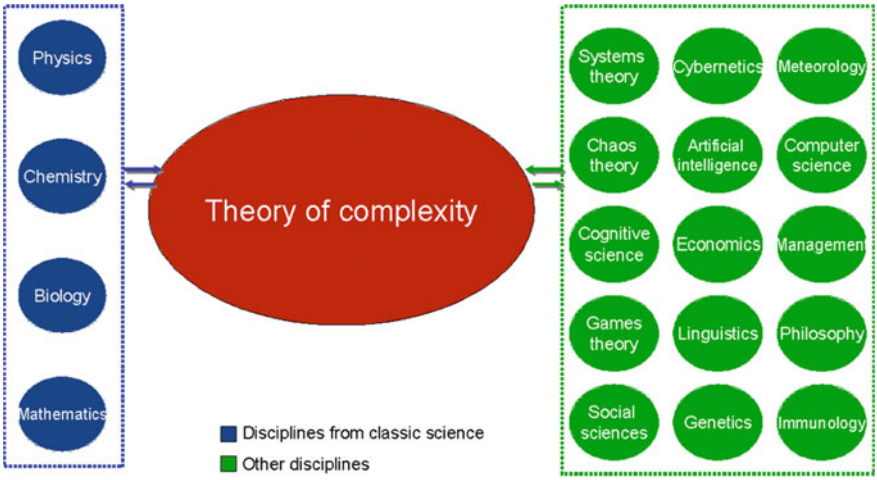



**Fig. 8.14** The disciplines that contribute to the modern general theory of complexity. Complex systems are encountered in all these disciplines (http://www.diegm.uniud.it/create/Immagini/sciences.gif)

Figure 8.14 gives a pictorial representation of the theory of complexity which incorporates and draws from classical science and modern scientific branches.

## 8.10 Emergence

The concept of *emergence* has a long history in which many philosophers and scientists have attempted to define it in several different ways. The debate about what, actually, is emergence started in the early twentieth century and has strongly reasserted itself during the most recent three decades. This reappearance of the emergence debate is mainly due to the development of *complexity science*, in general, and *complex adaptive systems*, in particular, which have led to the death of *positivistic reductionism* and the resurgence of *emergentism* (nonreductive materialism). We start our discussion with a few of the opinions and definitions of emergence given over the years.

**Samuel Alexander (1920)** [114] has viewed the concept of emergence within the framework of metaphysics, according to which the activity of a human is the outcome of a unique type of process of a physicochemical nature. In volume II of this work, [114], he states: "We are forced, therefore, to go beyond the mere correlation of the mental with these neural processes and to identify them. There is but one process, which, being of a specific complexity, has the quality of consciousness…" This means that there appear new qualities and corresponding patterns of high-level causality that cannot be expressed directly via the more

fundamental principles and entities. Actually, they are macroscopic patterns taking place within microscopic-level interactions. Emergent qualities are something new on Earth, but the world's fundamental dynamics remain invariant.

**Lloyd Morgan (1923)** [47, 115] describes emergent evolution as follows: "Evolution, in the broad sense of the word, is the name we give to the comprehensive plan of sequence in all natural events. But the orderly sequence, historically viewed, appears to present, from time to time something genuinely new. Under what I here call *emergent evolution* stress is laid on this incoming of the new." According to Morgan, "the emergent step… is best regarded as a qualitative change of direction, or critical turn point, in the course of events, and emergent events are related to the expression of some new kind of relatedness among pre-existent events." More specifically, Morgan's view of emergence gives us the basic features that must be possessed by a phenomenon, process, property, event, etc. to be characterized as emergent. These features are as follows [116]:

- It must be genuinely new, i.e., something that never happened before in the course of evolution.
- It is something tightly connected with the appearance of a new kind of relatedness among pre-existent events or entities.
- It changes the evolution mode, as the way pre-existent events run their course is altered in the context of new kind of relatedness.

**C.D. Broad (1925)** [117] was concerned with the so-called "*mechanistic-vital*" debate on the question: "Are the apparently different kinds of material objects irreducibly different?" and with the broader question: "Are the special sciences reducible to more general sciences (e.g., biology to chemistry) and finally to physics, the base-level science?" He concluded that *emergence* is the result of primitive high-level causal interactions, which are beyond those of the more fundamental levels.

**Stephen Pepper (1926)** [118] was concerned with the question: "Are there emergents?" His first argument was that "indeterminism is neither essential to, nor characteristic of, theories of emergent evolution." His second conclusion was that "a theory of emergent qualities is palpably a theory of epiphenomena." Finally, he classified emergence theories in theories of *emergent qualities* and theories of *emergent laws*. This classification yields actually the following three categories of emergent theories:

1. Theories of emergent qualities without emergent laws.
2. Theories of emergent qualities with emergent laws.
3. Theories of emergent laws without emergent qualities.

**Paul Meehl and Wilfrid Sellars** [119] remarked that only *class 1* is committed to epiphenomenalism, and that to make it consistent with determinism one must refuse to call "*laws*" the regularities between emergent qualities and the contexts in which they emerge, or, calling them "*laws*" refuse that they are emergent. After thorough

reasoning, Meehl and Sellars conclude that "*Peppers*" demonstration of the impossibility of non-epiphenomenal emergent is invalid.

**Crutchfield (1994)** [120] introduced the concept of *intrinsic emergence* which involves behaviors that possess the following properties:

- They are compatible with the model used by the observer concerned.
- Their occurrence cannot be foreseen in advance on the basis of the adopted model only.
- They are macroscopic, i.e., they occur persistently despite any changes in the observational scale.

Obviously, the results of the observer depend on the means he/she has available to observe the behavior at hand, or measure operations, and on his/her mental schemas. Thus, emergence is a concept that can be defined only relative to the observer. The question here is, of course, whether there really exist mathematical models that allow intrinsic emergence. The answer is *affirmative*.

**Eliano Pessa** [121] points out that the most celebrated *ideal models* of intrinsic emergence are the models that are based on the mechanism of *Spontaneous Symmetry Breaking* **(SSB)** in Quantum Field Theory [122]. Nonideal models of intrinsic emergence include cellular automata, fuzzy logic models, neural networks, and the models of artificial life. Our current knowledge on *ideal* and *nonideal* intrinsic emergence gives rise to many important but until now unanswered questions. Some examples of these questions are as follows:

- How are ideal and nonideal emergence models related?
- Is it possible to generalize the SSB mechanism?
- How can we describe the formation, within an ideal model, of finite volume subdomains, associated with suitable boundaries?

Some initial comments on the above question are provided by Pessa [121].

We now proceed to a short general discussion of some general issues on emergence [123–125]. **Emergence** is actually a *philosophical concept of art* that applies to substances and properties:

- That "rise" from more fundamental substances or properties.
- That "are new" or "irreducible", i.e., in some sense they do not dependent on more fundamental substances or properties.

This means that we have to face the phenomenon of two sets of properties/substances that are distinct but yet closely related. The second feature is what it makes emergent properties really interesting and singles them out any odd distinct properties. Emergence is generally categorized into the following two types:

- Epistemological emergence.
- Ontological emergence.

*Epistemology* is concerned with the question: "What do we know or can we know, and how do we come to know certain things?

*Ontology* is concerned with the question: "What kind of entities, substances, properties, etc., exist?

Therefore, emergent properties are either new properties from an *epistemological* point of view or new properties from an *ontological* point of view. In any case, they add something to our knowledge of the world. Of course, their common aspect is their dependence (in some way) on the more fundamental properties.

Epistemological emergence deals with what we can know about the behavior of complex systems and not what comes into existence, and so epistemological emergent properties would not be known on the basis only of our knowledge about the parts of the system and their interactions. Epistemological emergence is either *strong* **(C.D. Broad)** or *weak* **(S. Alexander)**. Strong epistemological emergence declares that not even an ideal cognizer would be able to know which properties will emerge from a complex system given its parts, their internal interactions, and their interaction with the environment. The two most common versions of epistemological emergence are as follows:

- *Predictive:* Emergent properties are features of complex systems which could not be predicted, from the standpoint of a pre-emergent stage, despite a thorough knowledge of the features of, and laws governing, their parts.
- *Irreducible:* Emergent properties and laws are systematic features of complex systems governed by true, law-like generalizations within a particular science such that these generalizations cannot be reduced to (captured in) the concepts of physics.

Epistemological emergence is *weaker* than ontological emergence. In ontological emergence, the physical world is seen as completely consisting of physical structures (simple or composite). It is noted that composite structures are not (always) mere sums or aggregates of the simple structures, because there are levels, or layered strata, or objects as complexity increases. Each higher layer is the outcome of the emergence of an "*interacting gamma of new qualities*". As J. Kim pointed out [126], in order to have *robust emergence* (natural relation), something more than *supervenience* and *functional irreducibility* is needed, because supervenience allows for different background/base relations, and so it is not a robust concept.

*Weak emergence* includes properties or states $P$ that are characterized by one or more features from the following list:

- The outcome of many lower level states (microstates) nonlinearly interacting.
- Unanticipated/unpredicted.
- Unintuitive or counter-intuitive.
- Qualitative different to any of the lower level states that lead to $P$.
- Nomologically different to any of the lower level states that lead to $P$.

Weak emergent properties are encountered in networks of biological signaling pathways [127], in microtubules [128], in conspecific attraction in fragmented

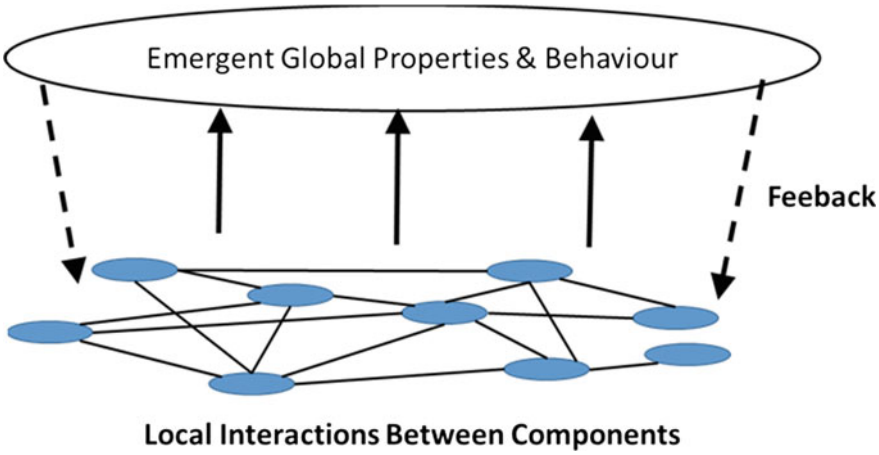



**Fig. 8.15** Pictorial illustration of emergence 8 http://www.tcd.ie/futurecities/assets/img/research/energy/Enabling%20Complex%20Adaptations%20in%20emerge.png9

landscapes [129], in functional structure tree growth models [130], etc. A short discussion on these properties, of course taking into account the various definitions of emergence involved in [127–130], is collectively provided in [125]. Some other interesting references on the concept of emergence in complexity science are the following [131–135]. Figure 8.15 is a schematic of the emergence of global properties and behavior from local interactions between system components.

## 8.11 More on Complex Adaptive Systems

*Complex adaptive systems* exhibit evolutionary and adaptation features similar to those possessed by systems; and so they operate and evolve like them. CASs actually bridge complex systems theory and Darwinian-type evolution principles and embrace a wide gamma of physical and social phenomena. Complexity theory (and adaptive complex systems theory) is a truly interdisciplinary theory embracing both hard-type and soft-type scientific branches.

According to Anderson [136], CASs can be effectively studied using the following eight scientific theories and approaches (see also [38]):

- Mathematical and computation complexity theory.
- Information theory and information-based measures.
- Ergodic theory: orbits, chaos, bifurcations, and attractors.
- Artificial entities such as cellular automata and computer-based simulators.
- Large random physical systems: spin glasses, neural nets, manifolds, etc.
- Self-organized criticality which occurs in random fractals, fluctuations at all scales, laws for avalanches distribution, and so on.

- Artificial intelligence: problem solvers, expert systems, classifier systems, genetic algorithms, etc.
- Welfare: the naturalistic study of CASs such as the brain, etc.

According to *W. Brian Arthur*, complexity of systems increases in three main ways [137]:

- Growth of coevolutionary variety, e.g., new individuals offer new niches, new opportunities for new individuals, etc.
- Structural deepening, i.e., systems exceed their limits due to the addition of new subsystems or novel functions.
- Capturing simpler entities and properly programming them.

According to Steels [138], "*evolving complex adaptive systems* **(ECAS)**" possess the following four properties:

- *Self-maintenance:* The system is actively establishing itself by drawing materials from the environment and by confirming a boundary between itself and the environment. This is the well-known process of *autopoiesis (self-creation)* [139].
- *Adaptability*: The system, in addition to self-maintenance, is able to adapt to small-scale changes in the environment in order to increase its chances for further existence.
- *Information preservation:* The information about the system is preserved, and so the system does not depend on the continued existence of its components to survive.
- *Spontaneous increase in complexity*: The system is capable of spontaneously increasing its own internal complexity, e.g., by introducing more components, more complex relations between parts, more complex behaviors of the parts, and so on. Sometimes, copies of the system come together to form a larger system, operating as a unique system at a higher level.

The properties of ECAS are possessed by certain types of chemical reactions, known as *autocatalytic* reactions or, as they are called, pre-life or uncoded life systems. In *living organisms,* the above four properties are achieved as follows:

- Self-maintenance is achieved by metabolic processes on materials drawn from the environment.
- Adaptability is achieved not only by chemical means but also by changing behavior.
- Preservation of information is achieved by coding the system in terms of genes. The code itself (in the form of DNA) is copied as opposed to the entire organism.
- More complexity is achieved via the genetic mechanism (mutation, combination, and natural selection).

Two other classes of systems that possess the four ECAS's properties are as follows:

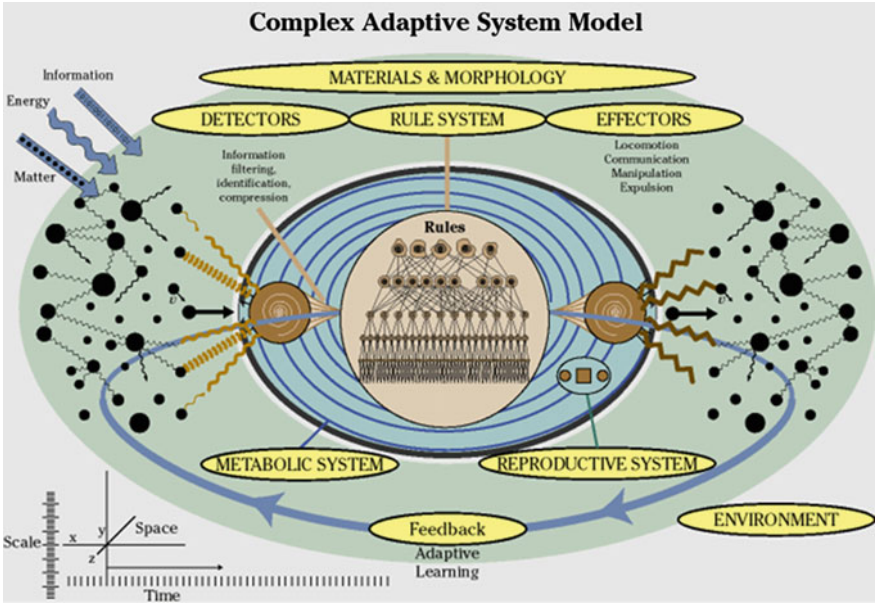



**Fig. 8.16** A self-explained model for complex adaptive systems in life and society (http://necsi.edu/projects/mclemens/casmodel.gif)

- *Intelligent systems* (via neural nets, symbolic capacity, responsiveness to environmental influences and control actuators, etc.).
- *Cultural systems* (which include languages systems, social systems, and religious systems).

Actually, there exist complex interrelations among these system classes because each one is based on some other, e.g., living organisms involve multiple autocatalytic networks, intelligent systems have emerged from living systems, and cultural systems have been developed via intelligent systems. Figures 8.16 and 8.17 show two general pictorial illustration models of CAS, which include biological as well technological systems.

As we have already mentioned in several places of this section (Sect. 8.6.2, etc.), the changes of the agents of a CAS are based on their interactions with one another and with the environment. This process is collectively called *coevolution*. To describe coevolution, Kauffman [66] has coined the concept of *fitness landscape*, which for each system involves all possible survival possibilities available to it. Figure 8.18 shows that a landscape comprises many peaks and valleys. The higher the peak is, the greater the fitness that it represents.

The adaptation/evolution of the system is actually a journey (search) within the fitness landscape aiming at locating the highest peak. The system can be "trapped" on the first peak it finds, or on some peak lower than the global maximum. If the

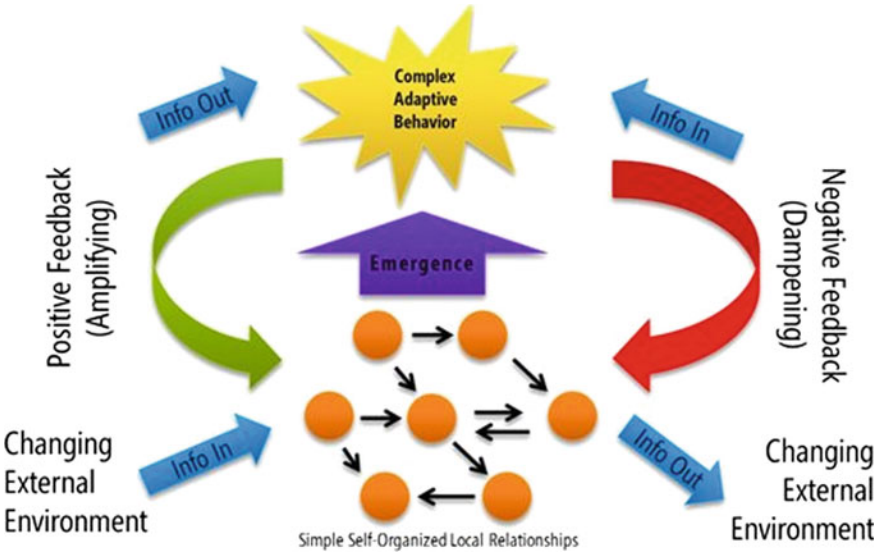



**Fig. 8.17** A second generic model for complex adaptive systems in life and society based on the emergence concept [(see Fig. 8.15) Wikipedia: http://blogs.msdn.com/cfs-filesystemfile.ashx/__key/communityserver-components-imagefileviewer/CommunityServer-Blogs-Components-WeblogFiles-00-00-01-37-81/7356.Complex-Adaptive-System.jpg_2D00_550x0.jpg]

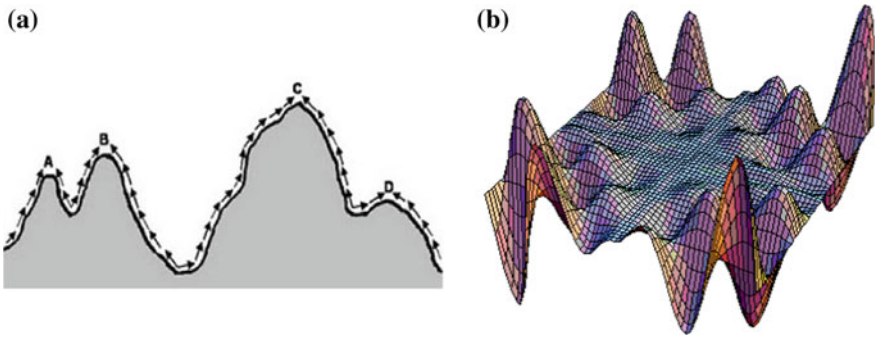



**Fig. 8.18** Fitness landscape of a CAS. **a** A simple 2-D landscape (arrows show the preferred evolution path). **b** A 3-D fitness landscape (genotypes move in the deviation of increasing fitness)

system changes its strategy, other interconnected subsystems (agents) will react and the landscape will undergo some change.

Here is exactly the point where Holland's *genetic algorithms* (GAs) [39] find their high value and importance. According to Holland, an adaptive system develops what he called "*adaptive plans*" in the way living organisms evolve genetically. Then he worked out and constructed computer programs that were able to specify "*what is to be done*" in order to solve a problem of the type shown in

Fig. 8.8. He then specialized further the "*adaptive plan*" to what he called "*genetic plan*", which was actually the basis for developing the *genetic algorithms* (GAs) and *evolutionary computation*. Without going into the details, GAs are parallel, computational representations of the biological processes of *selection, crossover (*recombination*), mutation,* and *inversion* on the basis of the fitness that is inherent in most processes of adaptation and evolution. GAs have been extensively and successfully used to solve problems of optimization, control, classification, learning, ecological modeling, etc.

Actually, a GA is a method for moving from one population of "*chromosomes*" (strings of bits 0, 1) to a new population by using a form of "*natural selection*", along with the operations (processes) of crossover, mutation, and inversion. Each chromosome consists of "*genes*" (i.e., bits), each gene being an instance of a particular "*allele*" (e.g., 0, 1). The selection, crossover, mutation, and inversion operators perform as follows:

- **Selection operator:** This operator chooses the population chromosomes that will be permitted to reproduce. The fitted chromosomes reproduce, on the average, more offspring than the ones with less fit.
- **Crossover operator:** This operator exchanges subparts of two chromosomes, basically in the way that two chromosomes are biologically recombined between two single-chromosome organisms.
- **Mutation operator:** This operator changes randomly the *allele* values of some locations in the chromosome and so creates variation.
- **Inversion Operator:** This operator reverses the order of a contiguous section of the chromosome, thus rearranging the order in which genes are arrayed.

Figure 8.19 shows the diagram (flow chart) of one cycle of a GA which is called a *generation*. For clarity, Fig. 8.20 presents an alternative flow chart of the GA evolutionary cycle.

A GA is usually repeated (iterated) from about 50 to 500 generations (or more). The whole set of generations is called a *run*. The steps of a basic GA are the following:

*Initial time:* $0 \leftarrow t$

*Step 0:* Select an arbitrary initial population $P(t)$ and compute the initial value of its fitness function $f(t)$

**Evolutionary cycle**

*Step 1:* Until termination

- Increase by 1 the time: $t \leftarrow t+1$
- Choose a subpopulation for the generation of offsprings $P^{'}(t) \leftarrow$ Selected parents $P(t)$
- Recombine the "genes" of the selected parents, i.e., mutate $P^{'}(t)$ with crossover rate $p_c$
- Disturb the paired population randomly, i.e., mutate $P^{'}(t)$ with mutation rate $p_m$
- Compute the new fitness function of $P^{'}(t)$

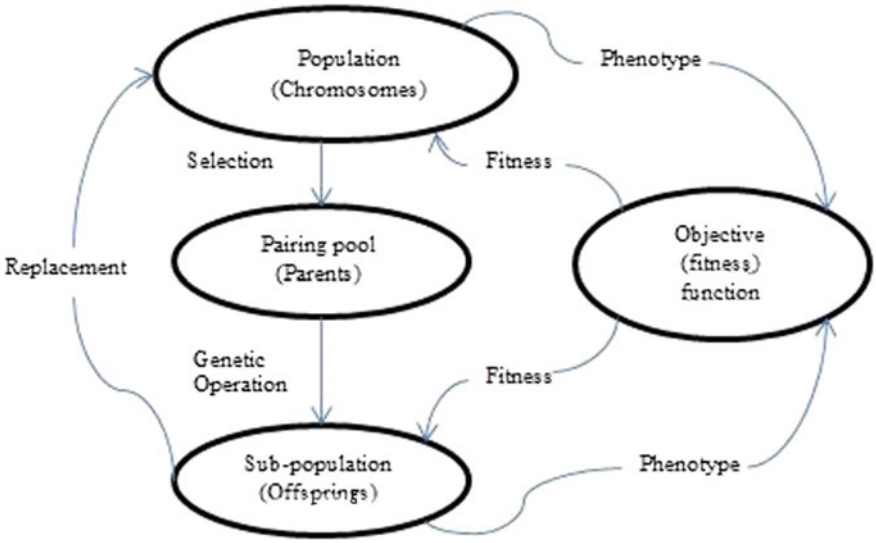



**Fig. 8.19** Flowchart of the evolutionary cycle of a genetic algorithm

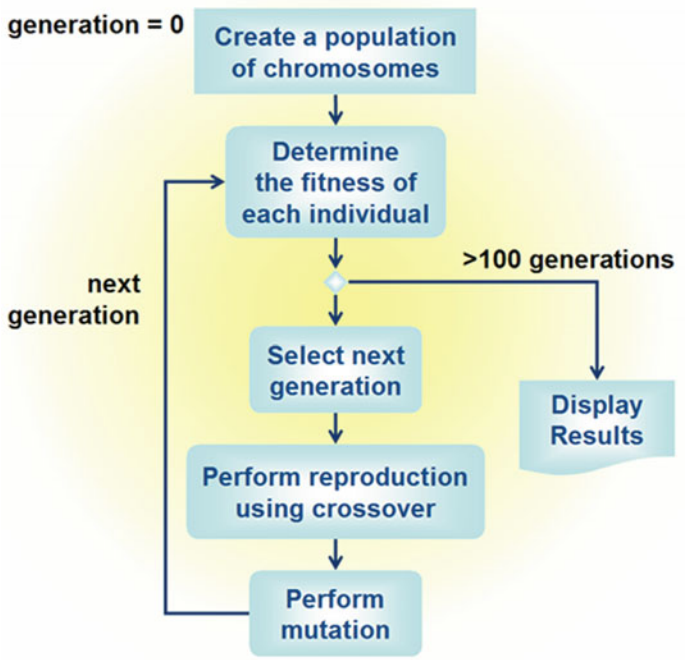



**Fig. 8.20** Alternative flow chart of the evolutionary cycle of a genetic algorithm (http://1.bp.blogspot.com/_nWD8gSvCXFk/TNnb__oQIRI/AAAAAAAACo0/vbSn02aQ2wU/s1600/rhjan07g01.png)

*Step 2*: Select the survived individuals from the actual value of the fitness function

*Step 3:* $P \leftarrow Survival\, P,\ P^{'}(t)$ If the termination criterion is satisfied end the algorithm. Otherwise, repeat from step 1.

The above GA is the basis for most of the applications of GAs. Of course there are also many other details that must be considered (e.g., the population size $N$, the string length $l$, crossover and mutation rates, and termination criterion). The success of the GA depends very much on these details. One of the most popular schemes for parent selection is the "*roulette wheel scheme*" in which each individual is given a "slice" of a circular roulette wheel equal in area to the individual's fitness. The roulette wheel is "spun", the ball comes to rest on one wedged-shaped slice, and the corresponding individual is selected. If the population size is $N$, then the roulette wheel will be allowed to spun $N$ times. The termination criterion may be the number of evolutionary cycles (computation runs), or the number of individuals in the various generations, or a predetermined value of the fitness function. The selection of the parameters $p_c$ (crossover rate or crossover probability) and $p_m$ (mutation rate or mutation probability) needs the solution of a complex optimization problem. Typical values of $p_c$ and $p_c$ for large $N$ are $p_c = 0.7$, $p_m = 0.001$, and for small $N$ are $p_c = 0.9$ and $p_m = 0.01$. To illustrate how the above basic GA works, we present the following example.

*Example* The problem is to apply the basic GA to the following case:

*String length:* $l = 8$

*Fitness function:* $f(x) = number$ of "1$s$" in the chain

*Crossover rate:* $p_c = 0.7$

*Mutation rate:* $p_m = 0.01$

Initial population $P(t)$ (randomly selected, e.g., by roulette wheel)

| Chromosome | String | Fitness |
|---|---|---|
| A | 00000110 | 2 |
| B | 11101110 | 6 |
| C | 00100000 | 1 |
| D | 00110100 | 3 |

**Solution**

We will give one cycle of the GA. The other cycles work in the same way. The subpopulation for producing offsprings via crossover is $\{B, D\}$. Crossover takes place after the first bit to give the following two offsprings:

$$E = 10110100$$

$$F = 01101110$$

The non-crossovered parents $A$ and $C$ produce offsprings that are exact copies of them.

At the next step, the offsprings $E$ and $F$ are subject to mutation at each position with probability $p_m$. Thus, suppose that the offspring $E$ is mutated at the sixth position to give

$$E' = 10110000$$

The offsprings $F$ and $C$ are not mutated, and the offspring $B$ is mutated at the first position to give

$$B' = 01101110$$

Thus, after the first cycle of the GA the new population $P'$ is as follows:

| Chromosome | String | Fitness |
|---|---|---|
| *E′* | 10110000 | 3 |
| F | 01101110 | 5 |
| C | 00100000 | 1 |
| *B′* | 01101110 | 5 |

We observe that, in the new population, the best chromosome B that has fitness six was lost. But, in total, the new population is better because its average fitness has increased from $\bar{f} = 12/4$ to $\bar{f} = 14/4$.

This cycle is repeated until the desired termination criterion is satisfied.

For understanding and explanation of the operation of Gas, which is complicated contrary to their simple description and programming, Holland [39], Goldberg [59], et al. developed several theories. The basic theory is Holland's *schemas (schemata) theory*. A brief sketch of this theory based on the three-dimensional space of Fig. 8.21 is the following.

Suppose that the solution of a problem can be represented by three-bit strings. Then, we have $2^3 = 8$ possible strings 000, 001, 101, 100, 010, 011, 111, and 110 which are placed at the eight vertexes of the cube as shown in Fig. 8.9, where two neighboring strings differ only by one bit (with the 000 string at the origin).

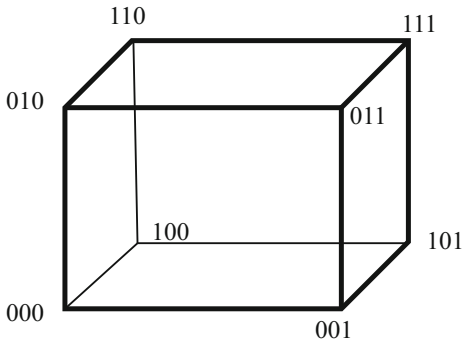



**Fig. 8.21** Three-dimensional (cubic) search space

A "*schema*" is defined to be a set of bit strings that can be described by a template of 1s, 0s and asterisks "*" which represent "*don't cares*" ("*wild cards*"). Each schema $S$ corresponds to a hyperplane (or subset) of the search space, i.e., to all strings that are matched to all positions except the positions "*". For example, the schema $S = 0$ ** denotes the set of all 3-bit strings that begin with 0. In Fig. 8.9, this schema is the front plane (face) of the cube. Clearly, each schema matches exactly $2^r$ strings, where $r$ is the number of the asterisks. Every binary string is a *chromosome* that corresponds to one vertex of the cube. Of course, not every possible subset of the set of length L-bits constitutes a schema (actually the large majority cannot). There exist $2^L$ possible strings of length $L$, and so $2^{2L}$ possible subsets of strings, but there are only $3^L$ possible schemas.

The central idea here is that schemata are (implicitly) the *building blocks* that are processed by the GA via selection, mutation, and single-point crossover. The GA is a search based on populations. A population of sampled points (instances evaluated) provides information for a large number of hyperplanes. In gas, there is inherent parallelism. Each time a simple string is evaluated, many different hyperplanes are evaluated in an indirect parallel fashion, but it is the cumulative evaluation outcome of the population of points that gives statistical information for any particular subset of hyperplanes. Since a schema represents a set of strings, we can associate a fitness value $f(S,t)$ with the schema $S$ and the average fitness of the schema. Then, $f(S,t)$ determines all the fitted strings of the population. Holland estimated the number of fitted strings of a schema $S$, for the case of analog fitness in the reproduction phase. The result is known as *Holland's schema theorem*, which describes the growth of a schema from one generation to the next. This theorem implies that short, low-order schemas, the average fitness of which remains above the mean, will receive exponentially increasing number of points (samples) over time. The hypothesis, which GAs work in this way, is called the "*building-block hypothesis*". That is, a GA searches for the (near) optimal performance via the competition of short, low-order, and high-performance schemas which are called "*buildings blocks*". Note that the order $\tau(S)$ of schema $S$ is the number of defined (i.e., non asterisk) bits in $S$. Actually, the *schema theorem* provides a lower bound for the expected number of instances, $a(S,t)$, at time $t+1$ (i.e., in the next generation), namely

$$a(S,t+1) \geq a(S,t)\frac{f(S,t)}{F(t)}\left[1 - p_c\frac{l(S)}{L-1} - p_m\tau(S)\right]$$

where $F(t)$ is the average fitness of the current generation, $a(S,t)$ is the number of the instances at the current generation, $L$ is the length of chromosomes, $l(S)$ is the *defining length* of the schema $S$ (defined as the distance of its mostly separated positions) and determining how much compact is the information contained in $S$), $p_c$ is the probability (rate) of the single-point crossover, $p_m$ is the mutation probability (rate), and $\tau(S)$ is the schema's order. We observe that indeed in order to have a high growth rate it is not sufficient to have only high average fitness, but also

*short* (small length l), *low order* $\tau(S)$, above the average fitness $(f(s,t) > F(t))$ schemas. Holland has proved that the number of schemas (hyperplanes) that can be processed in a single cycle is of order $N^3$, where $N$ is the population size. This result does not hold for any population size $N$, because in order for $N$ to be reasonable it must be selected taking into account the length $L$ of the chromosome. Actually, it was shown that for $L \geq 64$ we must have $2^6 \leq N \leq 2^{20}$ which is quite large region of allowable population sizes. Proofs of Holland's theorem and many other details on GAs can be found in [39, 59, 77, 140–142].

We close our discussion on CASs by outlining three basic tools that have been used for studying complex adaptive systems. These are the following [143, 144]:

- **Aggregation of models:** These are very useful because aggregation simplifies the model without significant error. Aggregation models include Markov chains and GAs.
- **Test problem generation:** GAs, or more generally CASs, work on a certain environment. Test problem generation involves, for example, the production of random environments with several minor adjustments. In this way, the problem reduces to that of matching the performance of the algorithm to the environmental features.
- **Agent-based simulation:** This kind of simulation provides a method for emulating the nonlinear properties of real-world complex systems. Among these techniques, we have simulation of games (particularly for social and economic systems) and agent model simulation, which is most appropriate for the study of systems exhibiting adaptive behavior.

## 8.12 Concluding Remarks

In this chapter, we have provided a conceptual overview of adaptation, complexity, and complex adaptive systems, including biological, hard science, soft science, and computer science issues.

The *adaptation process* was of primary concern to biologists and scientists over the years, and vigorous debates emerged, especially after the publication in 1872 of Darwin's theory about the origin and evolution of species on Earth. At the center of this debate was the interpretation of terms such as *function, capacity to learn, pre-adaptation, mind, and brain, teleonomy versus teleology* [145], etc. *Julian Huxley* points out that "Adaptation and function are two aspects of one problem" [2]. The chapter presented the different types of adaptation, a historical exposition of the principal views of adaptation and their variants, and the widely recognized adaptation mechanisms. *Teleonomy* is a term coined by *Collin Pittendrigh* (1958) in the framework of cybernetics and self-organizing systems [146], and extensively used by many researchers in evolutionary biology. Teleonomy is the hypothesis that adaptations emerge without the existence of a prior purpose, but by the process of

natural selection on genetic variability. This is opposite to the *teleology* of Aristotle that accepts the existence of intention, and purpose, which was revisited in 1977 by *Ernest Nagel* who studied the "*biological goal directness*" [147].

Complexity is an inherent property of life and technology. Therefore, the capability of modern science and technology to improve human behavior is critically dependent on our understanding of systems as overall entities, not just as merged components. Complexity of systems is the outcome of emergence arising from structures and functions, information and actions, evolution of patterns of behavior, multiscale and multiple possibilities, etc. From the level of biomolecular interactions to the tactics and strategies used in modern organized human society, locally and globally, complexity is shown to be a common unifying feature of system operation and action. A useful collection of book reviews in the area of complexity, complex systems, the evolution of complexity, complex adaptive systems, and their applications was assembled by Cowan, Pines, and Meltzer in [137]. A dynamic source of information on the new science of complexity is provided by Codynamics [148], and a comprehensive list of references, with their websites, on "Complex Adaptive Thinking" is offered by Brian McIndoe (2008) in [149].

## References


1. P.A. Corning, Biological adaptation in human societies: a basic needs approach. J. Bioecon. **2**, 41–86 (2000)
2. J. Huxley, *Evolution the Modern Synthesis* (Allen and Unwin, London, 1942)
3. C. Darwin, *The Origin of Species* (John Murray, London, 1872). http://www.sacred-texts.com/aov/darwin/origin/index.htm
4. Reaction to Darwin's Theory: Wikipedia. http://en.wikipedia.org/wiki/Reaction_to_Darwin's_theory
5. Adaptation: Wikipedia. http://en.wikipedia.org/wiki/Adaptation
6. R. Swenson, Thermodynamics, Evolution and Behavior, in *The Encyclopedia of Comparative*, ed. by G. Greenberg, M. Haraway (Garland Publishers, New York, 1997). http://www.entropylaw.com/thermoevolution1.html
7. C. Hubert, Adaptation. http://christianhubert.com/writings/adaptation.html#16
8. P.R. Ehrlich, P.H. Raven, Butterflies and plants: a study in coevolution. Evolution **18**, 586–608 (1964)
9. J. Maynard Smith, *The Theory of Evolution* (Penguin, New York, 1975)
10. T. Dobzhansky, On some fundamental concepts of evolutionary biology. Evol. Biol. **2**, 1–34 (1968)
11. T. Dobzhansky, *Genetics of the Evolutionary Process* (University. of Columbia Press, N.Y., 1970), pp. 1–6, 79–82, 84–87
12. T. Dobzhansky, Genetics of natural populations XXV. Evolution **10**, 82–92 (1956)
13. D.L. Hardesty, *Ecological Anthropology* (J. Wiley, New York, 1977)
14. R.C. Lewontin, Adaptation. Sci. Am. **239**(3), 213–230 (1978)
15. R.C. Lewontin, Adaptation, in *Conceptual Issues in Evolutionary Biology*, ed. by E. Sober (Harvard University Press, Cambridge, MA, 1984)
16. R. Swenson, Emergent Evolution and the Global Attractor: The Evolutionary Epistemology of Entropy Production Maximization, in *Proceedings of the 33rd Annual Meeting of the*

*International Society for the Systems Sciences*, vol. 33, ed. by P. Leddington (No. 3, 1989), pp. 46–53
17. D.E. Koshland Jr., The seven pillars of life. Science **295**, 2215–2216 (2002)
18. R. Brandon, Adaptation and representation: the theory of biological adaptation and function. Interdisciplines. http://www.interdisciplines.org/adaptation/papers/10
19. R. Brandon, *Adaptation and Environment* (Princeton University Press, Princeton, NJ, 1990)
20. R. Brandon, The Principle of Drift: Biology's First Law. J. Philos. **103**(7), 319–335 (1996)
21. L.E. Orgel, F.H.C. Crick, Selfish DNA: the ultimate parasite. Nature **284**, 604–607 (1980)
22. W.F. Doolittle, C. Sapienza, Selfish genes, the phenotype paradigm and genome evolution. Nature **284**, 601–603 (1980)
23. J.E. Stewart, The evolution of genetic cognition. J. Soc. Evol. Syst. **20**, 53–73 (1997)
24. J.E. Stewart, Metaevolution. J. Soc. Evol. Syst. **18**, 113–147 (1995)
25. J.E. Stewart, Evolutionary Transitions and Artificial Life. Artif. Life vol. **3** (1997)
26. J.E. Terrell, Adaptation. in *Proceedings of the Symposium on 'Key Concepts in Modern Evolutionary Archaeology*' (64th Annual Meeting of the Society for American Archaeology, Chicago, 1999)
27. B. Bogin, M.I. Vareta Silva, L. Rios, Life history trade- offs in human growth: adaptation or pathology? Am. J. Hum. Biol. **19**(5), 631–642 (2007)
28. D.L. Hardesty, The ecological perspective in anthropology. Am. Behav. Sci. **24**(1), 107–124 (1980)
29. E.A. Smith, B. Winterherlder (eds.), *Evolutionary Ecology and Human Behavior* (Aldine De-Gruyter, New York, 1992)
30. E.E. Ruyle, Genetic and cultural pools: some suggestions for a unified theory of biological evolution. Hum. Ecol. **1**, 201–215 (1973)
31. R. Naroll, *The Moral Order: An Introduction to the Human Situation* (Sage Publications, Beverly Hills, CA, 1983)
32. B. Colby, Well-being: a theoretical paradigm. Am. Anthropol. **89**, 879–895 (1987)
33. J. Rawls, *A Theory of Justice* (Harvard University Press, Cambridge, MA, 1972)
34. A.K. Sen, *Welfare and Measurement* (The MIT Press, Cambridge, MA, 1982)
35. B. World, *Social Indicators of Development* (The John Hopkins University Press, Baltimore, 1996)
36. J.H. Holland, *Hidden Order: How Adaptation Builds Complexity* (Addison Wesley, Reading, MA, 1995)
37. J.H. Holland, Complex adaptive systems. Daedalus **121**, 17–30 (1992)
38. J. Brownlee, Complex adaptive systems. CIS Tech. Report **070302A**, 1–6 (2007)
39. J.H. Holland, *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence* (MIT Press, Cambridge, MA, 1975)
40. J.H. Holland, *Emergence: From Chaos to Order* (Addison-Wesley, Redwood City, Calif, 1998)
41. M.M. Waldrop, *Complexity: The Emerging Science at the Edge of Order and Chaos* (Simon and Schuster, New York, 1992)
42. I. Prigogine, I. Stengers, *Order Out of Chaos* (Bantam Books, New York, 1984)
43. E. Jantsch, *The Self—Organizing Universe* (Pergamon Press, Oxford, 1980)
44. H. Maturana, F. Varela, *The Tree of Knowledge* (Shambhala, Boston, 1992)
45. K. Dooley, A nominal definition of complex adaptive systems. Chaos Netw. **8**(1), 2–3 (1996)
46. J.C. Maxwell, *Teaching Nonlinear Phenomena* (King's College, London, 1873)
47. C.L. Morgan, The case of emergent evolution. J. Philos. Stud. **4**(15), 431–432 (1929)
48. E.N. Lorenz, Deterministic non-periodic flow. J. Atmos. Sci. **20**(3), 448–464 (1963)
49. L. von Bertalanffy, The History and status of general systems theory. Acad. Manag. J. Gen. Syst. Theory **15**(4), 407–426 (1972)

50. J.H. Holland, J.S. Reitman, Cognitive Systems Based on Adaptive Algorithms. in *Pattern-Directed Inference Systems,* eds. by D.A. Waterman, F. Haynes, Roth (Academic Press, New York, 1978)
51. R.M. May, Simple mathematical models with very complicated dynamics. Nature **261**, 459–467 (1976)
52. G. Nicolis, I. Prigogine, *Self-Organization in Non-Equilibrium Systems: From Dissipative Structures to Order through Fluctuations* (Wiley, New York, 1978)
53. B. Mandelbrot, *The Fractal Geometry of Nature* (Freeman, New York, 1977)
54. R.L. Ackoff, *The Art of Problem Solving* (Wiley, New York, 1978)
55. R.L. Ackoff, Some unsolved problems in problem solving. Oper. Res. Q. **13**, 1–11 (1962)
56. M. Smith, *Evolution and Theory of Games* (Cambridge University Press, Cambridge, 1982)
57. R.L. Devaney, *An Introduction to Chaotic Dynamical Systems* (Addison-Wesley, Redwood City, CA, 1986)
58. T.S. Parker, L.O. Chua, *Practical Numerical Algorithms for Chaotic Systems* (Springer, New York, 1989)
59. D.E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning* (Addison-Wesley, Reading, MA, 1989)
60. C.G. Langton, Computation at the edge of chaos: phase transitions and emergent computation. Physica D 12–37 (1990)
61. N.K. Hayles, *Chaos Bound: Orderly Disorder in Contemporary Literature and Science* (Cornell University Press, Ithaca, NY, 1991)
62. S.A. Kauffman, Antichaos and adaptation. Sci. Am. **265**, 78–84 (1991)
63. R. Lewin, *Complexity: Life at the Edge of Chaos* (MacMillan, New York, 1992)
64. R.L. Devaney, *A First Course in Chaotic Dynamical Systems* (Addison-Wesley, Reading, MA, 1992)
65. M. Mitchell, P.T. Hraber, J.P. Crutchfield, Revising the edge of chaos: evolving cellular automata to reform computations. Complex Syst. **7**, 89–130 (1993)
66. S.A. Kauffman, *The Origins of Order: Self-Organization and Selection in Evolution* (Oxford University Press, New York, 1993)
67. E. Ott, *Chaos in Dynamical Systems* (Cambridge University Press, Cambridge, 1993)
68. K. Kelly, Out of Control: The New Biology of Machines, Social Systems and the Economic World (Addison-Wesley, Boston, 1994). http://www.kk.org/outofcontrol/
69. S. Wolfram, *Cellular Automata and Complexity: Collected Papers* (Perseus, Reading, MA, 1994)
70. S. Wolfram, A New Kind of Science (Wolfram Media, 1994). http://www.wolframsscience.com/(2004)
71. M. Gell-Mann, *The Quark and the Jaguar: Adventures in the Simple and the Complex* (W.H. Freeman, San Francisco, 1994)
72. M. Gell-Mann, What is Complexity? Complexity vol. **1** (1995)
73. P. Coneney, R. Highfield, *Frontiers of Complexity: The Search for Order in a Chaotic World* (Fawcett Columbine, New York, 1995)
74. H.J. Morowitz, J.L. Singer, *The Mind, The Brain, and Complex Adaptive Systems* (Addison Wesley-Longman, Reading, MA, 1995)
75. B. Per, *How Nature Works: The Science of Self-Organized Criticality* (Copernicus, New York, 1996)
76. L.A. Fitzerald, *Organizations and Other Things Fractal: A Primer on Chaos for Agent of Change* (The Consultancy, Denver, CO, 1996)
77. M. Mitchell, *An Introduction to Genetic Algorithms* (MIT Press, Cambridge, MA, 1996)
78. F. Heylighen, What is Complexity, Brussels, Free University (1996). http://pcp.lanl.gov/HEYL.html2004
79. C. Langton, Modeling Complex Adaptive Systems (1997). http://www.anderson.ucla.edu/research/marschak/1997-98/abstracts/31oct97.htm
80. Y. Bar-Yam, *Dynamics of Complex Systems* (Perseus Books, Reading Mass, 1997)

81. K. Sigmund, Complex adaptive systems and the evolution of reciprocation. Ecosyst. Biomed. Life Sci. Earth Environ. Sci. **1**, 444–448 (1998)
82. E. Bonabeau, Social insect colonies as complex adaptive systems. Ecosyst. Biomed. Life Sci. Earth Environ. Sci. **1**, 427–430 (1998)
83. P. Cilliers, *Complexity and Postmodernism: Understanding Complex Systems* (Routledge, London, 1998)
84. M.R. Lissack, Managing the Complex: Mastering Corporate Complexity. Doing It, Not Just Talking About It: The Role of Coherence. in *Proceedings of the Annual Colloq. on Complex Systems and the Management of Organizations* (NESCI, Boston, March 1999). http://www.learning.org.com/98.11/0314.html
85. L.A. Segel, Diffuse feedback from diffuse information in complex systems. Complexity **5**, 39–46 (2000)
86. M. Buchanan, *Ubiquity: Why Catastrophes Happen?* (Three River Press, New York, 2000)
87. R.M. Smith, M.A. Bedau, Is echo a complex adaptive system? Evol. Comput. **8**, 419–422 (2000)
88. Y. Bar-Yam (ed.), Unifying Themes in Complex Systems I. in *Proceedings of the 1st International Conference on Complex Systems* (Perseus Press, New York, 2000)
89. R.K. Sawyer, Emergence in Sociology: Contemporary Philosophy of Mind and Some Implications for Sociological Theory: 1. Am. J. Sociol. **107**(3), 551–585 (2001)
90. D. Harris, *Echo Implemented: A Model for Complex Adaptive Systems Computer Experimentatation* (Sandia National Labs, USA, SAND, 2001), pp. 2001–2097
91. Y. Bar-Yam, A. Minai (eds.), Unifying Themes in Complex Systems II. in *Proceedings of the 2nd International Conference on Complex Systems* (Perseus Press, New York, 2002)
92. S.A. Lewin, Complex adaptive systems: exploring the known and the unknowable. Am Math. Soc. **40**, 3–19 (2003)
93. S. Harkema, A complex adaptive perspective on learning within innovation projects. Learn. Organ. **10**(6), 340–346 (2003)
94. R.L. Goldstone, Y. Sakamoto, The transfer of abstract principles governing complex adaptive systems. Cogn. Psychol. **46**, 414–446 (2003)
95. S. Bullock, D. Cliff, Complexity and Emergent Behavior in ICT Systems. Hewlett-Packard Labs HP-2004-187 (2004). http://www.hpl.hp.com/techreports/2004/HPL-2004-187.html
96. L.M. Holden, Complex adaptive systems: concept analysis. J. Advanced Nurs. **52**, 651–657 (2005)
97. R. Harre, Resolving the emergence-reduction debate. Synthése **151**(3), 499–504 (2006)
98. S. Kauffman, P. Clayton, On emergence, agency, and organization. Biol. Philos. **21**(4), 501–521 (2006)
99. L. Von Bertalanfly, *General Systems Theory: Foundations, Development, Applications* (George Braziller, New York, 1968)
100. H. Liu, *A Brief History of the Concept of Chaos* (Peking University, Department of Philosophy, Peking, 1999). http://members.tripod.com/~haaje/Paper/chaos.htm
101. A.N. Kolmogorov, *On Stability of Conditionally Periodic Motions in Conservative Dynamical Systems* (Proceedings of the International Congress of Mathematicians, Amsterdam, 1954)
102. V.I. Arnold, Proof of a Theorem of A.N. kolmogorov on the preservation of conditionally periodic motions under a small perturbation of the hamiltonian, uspehy math. Mark **18**(5), 13–40 (1963) [Translation to English in: Russian Mathematical Surveys, vol. **18**, 9–36 (1963)]
103. J.K. Moser, On invariant curves of area-preserving mappings of an annulus. Courant Inst. Math. Sci. Math. Phys. K1.II, (New York, University 1962), 1–20
104. T. Vincent-Walter, J. Grantham, *Non Linear and Optimal Control Systems* (Wiley, New York, 1997)
105. S.H. Strogatz, *Nonlinear Dynamics and Chaos* (Addison Wesley, Reading, MA, 1994)
106. E. Stepp, Fractal Frequently Asked Questions and Answers (Marshall University, Huntington, WV, 1995). http://www.faqs.org/faqs/fractal-faq/

107. M. Baranger, Chaos, Complexity, and Entropy: A Physics Talk for Non-Physicists (Center for Theoretical Physics, Department of Physics, MIT, Cambridge, MA, U.S.A., 2002). http://necsi.org/projects/bouranger/cce.pdf
108. D.W. Hock, in *The Chaotic Organization: Out of Control and Into Order, 21st Century Learning Initiative* (1996)
109. L.A. Fitzerald, What is Chaos? Denver. http://www.orgmind.com/whatis.html
110. Cooperative Versus Competitive Games, Ed. Games http://thegamesjournal.com/articles/FamilyPastmes.shtml
111. C. Montet, *Game Theory and Economics* (Palgrave Macmillan, London, 2003)
112. M.R. Lissack, J. Roos, *The Next Common Sense: Mastering Corporate Complexity through Coherence* (Nicholas Brealcy, London, 1999)
113. S. Lloyd, Measures of complexity: a non-exhaustive list. IEEE Control Syst. Magaz. **71**(4), 7–8 (2001)
114. S. Alexander, *Space, Time and Deity*, vol. I, II (Macmillan, London, 1920)
115. C.L. Morgan, *Emergent Evolution* (Williams and Norgate, London, 1923)
116. C. Nino El-Hami, S. Pihlstrom, *Emergence Theories and Pragmatic Realism.* http://www.helsinki.fi/science/commens/papers/emergentism.pdf
117. C.D. Broad, *The Mind and Its Place to Nature* (Loutledge and Kegem Paul, London, 1925)
118. S.C. Pepper, Emergence. J Philos. **23**, 241–245 (1926)
119. P.E. Meehl, W. Sellars, The Concept of Emergence, In *Minnesota Studies in the Philosophy of Science: Vol.I, The Foundations of Science and the Concepts of Psychology and Psychoanalysis*, eds. by H. Fregl, M. Soriven (University of Minnesota Press, Minnesota, 1956), 239–252
120. J.P. Crutchfield, The calculi of emergence: computation. Dyn. Induction, Physica, D **75**, 11–54 (1994)
121. E. Pessa, What is Emergence? in *Emergence in Complex*, *Cognitive, Social, and Biological Systems*, eds. by G. Minati, E. Pessa (Kluwer/Plenum, New York, 2002)
122. C. Hzyksan, J.B. Zuber, *Quantum Field Theory* (McGraw-Hill, Singapore, 1986)
123. T. O'Connor, H.Y. Wong, Emergent properties. Stanford Encycl. Philos. (October 23, 2006). http://plato.stanford.edu/entries/proprerties-emergent/
124. T. O'Connor, Causality, mind and free will. Philos. Perspect. **14**, 105–117 (2000)
125. A. Matthies, A. Stephenson, N. Tasker, The Concept of Emergence in Systems Biology. A Project Report. http://www.stats.ox.ac.uk/_data/assets/pdf_file/0018/3906/Concept_of_Emergence.pdf
126. J. Kim, Being Realistic About Emergence. in *The Re-Emergence of Emergence: The Emergentist Hypothesis from Science to Religion*, eds. by P. Clayton, P. Davies (Oxford University Press, Oxford, 2006), 189–202
127. U.S. Bhalla, R. Lyengar, Emergent properties of networks of biological signaling pathways. Science **283**, 381–387 (1999)
128. J. Tabony, Self-organization and other emergent properties in a simple biological system of microtubules. ComPlexUs **3**(4), 200–210 (2006)
129. R.J. Fletcher, Emergent properties of conspecific attraction in fragmented landscapes. Am. Nat. **168**(2), 207–219 (2006)
130. C. Eschenbach, Emergent properties modeled with the functional structural tree growth model almis: computer experiments on resource gain and use. Ecol. Model. **186**(4), 470–488 (2005)
131. M. Christen, L. Franklin, The Concept of Emergence in Complexity Science. *Proceedings of the Complex Systems Summer School,* Santa Fe Institute (2002). http://www.ini.uzh.ch/node/11635
132. P. Cariani, Emergence and Artificial Life, In *Artificial Life* II eds. by C. Langton, D. Farmer, S., Rasmussen (Addison-Wesley, Redwood City, CA, 1991), 775–797
133. S. Forrest (ed.), *Emergent Computation* (North Holland, Amsterdam, 1990)
134. A.J. Dyan, Emergence is coupled to scope. Not Level, Complex. **13**, 67–77 (2007)

135. P. Clayton, Conceptual Foundations of Emergence Theory, Ch.1 in *The-Re-emergence of Emergence*: The Emergentist *Hypothesis from Science to Religion*, eds. by P. Clayton, P. Davies (Oxford University Press, Oxford, 2000)
136. P.W. Anderson, The Eightfold Way to the Theory of Complexity-A Prologue. in *Complexity, Metaphors, Models, and Reality*, eds. by G.A. Cowan, D. Pines, D. Meltzer (Addison-Wesley, Reading, MA, 1994), 7–16
137. G.A. Covan, D. Pines, D. Meltzer (eds.), *Complexity*: *Metaphors*, *Models, and Reality* (Addison-Wesley, 1994)
138. L. Steels, Evolving Complex Adaptive Systems. http://arti.vub.ac.be/∼steels/origin/subsection331.html
139. F.J. Varala, H.R. Maturana, B. Uribe, Autopoiesis: the organization of living systems, its characterization and a model. Biosystems **5**, 187–196 (1974)
140. E. Cantu-Paz, *Efficient and Accurate Parallel Genetic Algorithms* (Kluwer, Boston/Dordrecht, 2000)
141. Z. Michalewicz, *Genetic Algorithms + Data Structures = Evolution Program* (Springer, Berlin/London, 1994)
142. L. Davis, *Handbook of Genetic Algorithms* (Van Nostrand Reinhold, New York, 1991)
143. S. Chan, *Complex Adaptive Systems* (MIT Press, Cambridge, MA, 2001)
144. E. Mittleton-Kelly, Organizations as Co-Evolving Complex Adaptive Systems. in *Proceedings of the British Academy of Management Conference* (London, Sept 1997), 8–10
145. R.M., Young, *Mind, Brain and Adaptation in the Nineteenth Century*, *Cerebral Localization and its Biological Context from Gall to Ferrier* (Clarendon Press, Oxford, 1970)
146. C.S. Pittendrigh, Adaptation, Natural Selection and Behavior. in *Behavior and Evolution*, *eds. by* A. Roe, G. Graylord Simpson (Yale Univ. Press, Yale 1958)
147. E. Nagel, Teleology revisited: goal directness processes in biology. J Philos. **74**, 261–301 (1977)
148. E.W. Buck Lawrimore, The new science of complexity vs. old science. Codynamics. http://www.codynamics.net/science.htm
149. B.W. McIndoe, Complex Adaptive Thinking (*JECHO References*) (2008). http://brianmcindoe.com/jechoref.html

# Chapter 9
# Self-organization

*If biologists have ignored self-organization, it is not because self-ordering is not pervasive and profound. It is because we biologists have yet to understand how to think about systems governed by two sources of order. We have to see that we are the natural expression of a deeper order…*

Stuart Kauffman

*As a philosopher I am interested in all kinds of phenomena of self-organization, from the wind patterns that have regulated human life for a long time to the self-organizing patterns within our bodies, to the self-organizing processes in the economy, to the self-organizing process that is the Internet.*

Manuel De Landa

**Abstract** Self-organization is an inherent process of life and society that refers to the capability of biological, natural, and society systems to change their structure by their own during their operation, such as to show more order or pattern without the help of external agents. This chapter starts with the ontological question "what is self-organization" and provides representative alternative answers given by eminent workers and thinkers in the field. It continues by discussing the four fundamental mechanisms of self-organization observed in nature viz. synergetics, export of entropy, positive/negative feedback interplay, and selective retention, followed by an examination of the concept of self-organized criticality (edge of chaos). Then, this chapter discusses the contribution of cybernetics to the study of self-organization, and the relation of self-organization with "complex adaptive systems (CAS)" providing a description of five self-organization features that are transferred to CASs. This chapter continues with the presentation of six examples of natural and artificial self-organizing systems, namely ecological systems, magnetization, convective instability cells, linguistic systems, knowledge networks, and self-organizing neural network maps. The conclusions provide some additional remarks about complexity and the future of man-made self-organizing systems.

**Keywords** Self-organization (S-O) · Natural S-O · S-O mechanisms
Synergy · Entropy export · Positive/negative feedback interplay
Requisite variety · Interdependence · Selective retention · Self-organized criticality
Complex adaptive system (CAS) · Society vertical/horizontal S-O



S. G. Tzafestas, *Energy, Information, Feedback, Adaptation, and Self-organization*,
Intelligent Systems, Control and Automation: Science and Engineering 90,
https://doi.org/10.1007/978-3-319-66999-1_9

Bifurcation · Pitchfork bifurcations · Stationary bifurcations · Ecological S-O Robotic S-O · Self-organizing map

## 9.1 Introduction

Self-organization is a concept that refers to the capability of biological, natural, and society systems to change their structure by themselves during their interaction process with the environment. This means that self-organization is not environment determined but self-determined and self-adaptive. In other words, one can say that a system is self-organizing if it tends to become more organized on its own, i.e., if it shows more structure or order or pattern without the help or influence of an external agent. Clearly, the self-organization concept is one of the most useful concepts in science and society, but at the same time a very vague concept, because all terms used to define it, viz., organization, structure, order, pattern, etc., are not uniquely defined or interpreted. In some cases, self-organization is interpreted as emergence, but this is not correct because we can have self-organization without emergence and emergence without self-organization, although both of them are features of complex adaptive systems (see Fig. 8.1). The idea that natural systems have a tendency to become more orderly without external intervention was first stated by the philosopher Descartes who argued that "ordinary laws of nature tend to produce organization" (see his "Discourse on Method"). Also, Kant argued that "the principle of unity of nature is a regulative principle according to which nature is constructed so as to correspond to our needs for order" (see his "Critique for Judgment"). Many authors have used other terms for defining self-organization, which sometimes are related to human behavior. One of these terms is *autopoiesis* coined by *Humberto Maturama* and *Francisco Varela* [1]. The term autopoiesis comes from the Greek composite word αυτο-ποίηση (autopoiesis = self-making/self-creating). Another term is extropy, which is the opposite of entropy. If we adopt the entropy interpretation of Boltzmann as disorganization (disorder), then extropy means organization (order). A general field where self-organization has been extensively studied is cybernetics. More information on this is provided in Sect. 9.5. Another new technological field closely related to self-organization is "Artificial Life" (ALife) [2, 3].

The purpose of this chapter is:

- To investigate the question "what is self-organization?" and present a set of definitions given by eminent researchers in the field.
- To outline and discuss the four fundamental mechanisms of self-organization observed in nature (synergy, entropy export, positive/negative feedback interplay, and selective retention).
- To examine the concept of "self-organized criticality", a term equivalent to the "edge of chaos".
- To discuss the contribution of cybernetics to the study of self-organization via a listing of well-known cyberneticists and their major results.

- To study the relation of self-organization with "complex adaptive systems" (CAS) providing a description of the five self-organization features that are shared with CAS.
- To present six representative examples of natural and artificial self-organizing systems, namely: ecological systems, magnetization, heated liquids (convective instability cells), linguistic systems, knowledge networks, and self-organizing maps.

## 9.2 What Is Self-organization?

*Self-organization* is inherent in life, nature, and society. However, only after the 1950s has the scientific study of self-organization assumed concrete shape. According to the Longman Dictionary, the word organization has three linguistic meanings [4]:

- The way in which different parts of a system are arranged and work together.
- Planning and arranging something so that it is successful or effective.
- A group such as a club or business that has formed for a particular purpose.

These meanings are used in our current scientific, information, technological, cultural, and economic society, and cover both cases: external and internal organization of a system. In general, all these definitions imply that organization is some kind of order and excludes randomness produced by any cause at any level. The alternative definitions presented here are the following.

### *9.2.1 Definition of W. Ross Ashby*

In modern times, the term "self-organizing" was first used in 1947 by W. Ross Ashby, a cybernetician (psychiatrist, neuroscientist, and mathematician) [5–7]. According to him "a system shows self-organization if its behavior shows increasing redundancy with increasing length of the protocol" Ashby used the term *redundancy* (R) in the Shannon's sense, i.e.,:

$$R = 1 - H/H_{\max}$$

where $H$ is the actual uncertainty (entropy) and $H_{\max}$ the maximum uncertainty of the system. He argued that: "Since redundancy $R$ can only increase if either H is decreasing or $H_{\max}$ is increasing, and since $H_{\max}$ can only change by redefining the system (i.e., by externally changing the number of states), one can say that a system is *self-organizing*, only if the increase in the redundancy $R$ is the outcome of a corresponding decrease in the randomness $H$." This essentially means that non-utilized, potential, channel bandwidth provides a measure of self-organization.

Ashby used Shannon's *Tenth Theorem* which states: "If an error correction channel has capacity C, then equivocation of amount C can be removed, but no more," to formulate his "*Law of Requisite Variety*", which states: "Any quantity K of appropriate selection demands the transmission or processing of quantity K of information. There is no getting of selection for nothing." Shannon's theorem was developed in the context of telephone and other similar communication channels, regarding a case with a lot of "message" and little "error". In biology, we face the case where the "message" is small, but the disturbing errors are many and large.

Both "Shannon's Tenth Theorem" and Ashby's "Law of Requisite Variety" are applicable to regulatory biological systems, such as the brain, through the fact that "the amount of regulatory or selective action that the brain can achieve is absolutely bounded by its capacity as a channel."

### 9.2.2 Definition of Francis Heylinghen

According to Heylinghen: "Self-organization is the spontaneous emergence of global structure out of local interactions" [8]. "*Spontaneous*" here means that no internal or external agent is in control of the process; for a sufficiently large system, any individual agent can be removed or replaced without any effect on the resulting structure. The self-organization process is fully parallel and distributed over all the agents, i.e., it is truly collective. This implies that the organization that is achieved is inherently robust to faults and perturbations.

### 9.2.3 Definition of Chris Lucas

He stated that: "Self-organization is the evolution of a system into an organized form in the absence of external constraints. It is a move from a large region of state space to a persistent smaller one, under the control of the system itself" [9]. Here, the term "organized form" is meant in the sense described before (i.e., nonrandom form).

### 9.2.4 Definition of Scott Camazine

According to him "Self-organization in biological systems is a process in which pattern at the global level of a system emerges solely from numerous interactions among the lower level components of the system, and the rules that specify interactions, among system components, are executed using local information, without reference to the global pattern" [10]. This definition implies that the pattern is an emergent property of the system and not a property imposed on the system by an external ordering influence.

### 9.2.5 Definition of A.N. Whitehead

He stated that: "Self-organization of society depends on commonly diffused symbols evoking commonly diffused ideas, and at the same time indicating commonly understood action" [11]. He argued that the human mind is functioning symbolically when some components of its experience elicit consciousness, beliefs, emotions, and usages, respecting other components of its experience. The former set of components involves the "symbols", and the latter set constitutes the "meaning" of the symbols. He remarks that "symbolism plays a dominant part in the way in which all higher organisms conduct their lives. It is the cause of progress and the cause of error."

### 9.2.6 Definition of M. B. L Dempster

Dempster studied the distinction between *autopoietic* (self-producing) and *sympoietic* (collectively producing) systems. These two contrasting lenses offer alternatives views of the world, forcing recognition of system properties frequently neglected. Taking into account Andrew's remark that it is difficult, probably impossible, to find a precise definition of what is understood by a self-organizing system, he did not attempt to give such a precise definition, while stating that: "On an intuitive level, self-organization refers to exactly what is suggested: systems that appear to organize themselves without external direction, manipulation, or control" [12].

Self-organization in human society occurs at various levels (vertical self-organization) and various activities or processes (horizontal self-organization). From top level to bottom level, vertical self-organization involves [7]:

- Human–non-human environments
- Society establishment
- Groups and communities
- Individuals.

On the horizontal dimension, we have:

| | |
|---|---|
| • Culture | • Economy |
| • Ideology | • Industry |
| • Politics | • Agriculture |
| • Religion | • Education, etc |

All processes are interdependent and influence each other. This implies that coevolution occurs within and between vertical and horizontal processes.

According to *Takatoshi Imada* [13], in the 1960s attempts were made to develop a theory based on the logic of a system and its control. Contrary to this view of a societal system as the aggregate of individuals where self-organization is the sum of the practices of a system driven by control, or self-control in particular, in the 1980s

a new view gained popularity, adopted based on the logic of creative individuals and fluctuations. This new view looks at the practices of individuals departing from the standard logic of a system, making the existing system fluctuate and transforming its structure. In [13], Imada integrated these two antithetical approaches into a structure of the self and through self-reflection. This opened new ways for designing planning and control actions and developing a spontaneously performative action theory. More information on society self-organization and societal complex adaptive systems will be given in Chap. 13.

## 9.3 Mechanisms of Self-organization

The fundamental natural mechanisms by which self-organization is achieved are the following:

- Synergetics
- Export of entropy
- Positive/negative feedback interplay
- Selective retention.

The *synergetics* mechanism (from the Greek συν-έργεια = synergia = act together) was discovered by the German physicist *Hermann Haken* [14], who studied lasers and other similar phenomena and was surprised by the apparent cooperation (synergy) between the interacting components. The elements (agents, components) of a complex system at the beginning interact only locally (i.e., with their close neighbors), but, due to the direct or indirect connection and interaction of the agents, the changes gradually propagate to faraway regions, leading finally to an obvious synergy at the system level. Examples of such collective patterns resulting from many interacting components include (besides lasers), chemical reactions, molecular self-assembly, crystal formations, spontaneous magnetization, etc. This synergy in laser-light production is explained as follows. When atoms or molecules receive an energy input, they emit the surplus energy as "photons" at random times and directions. This leads to *diffuse light*. But under certain conditions, the atoms can be synchronized and emit photons at the same time in the same direction, with the outcome of a highly coherent and *focused beam of laser light* [15].

The achievement of a *synergetic state* is, in general, a "trial-and-error" or "mutual adaptation" process. System's components (agents, etc.) handle permissible or plausible actions (or sometimes select them randomly) and maintain or repeat those actions that bring them nearer to their goals. This process is actually a natural-selection process, but it differs from Darwinian evolution since the system agents are functioning simultaneously until they mutually fit, i.e., they *coevolve* (mutually adapted) so as to minimize friction and maximize synergy.

The mechanism for the *export of entropy* self-organization was revealed by Prigogine and Nicolis [16]. They developed and promoted the theory of *dissipative structures* (i.e., systems that continuously decrease their entropy). Dissipation (i.e.,

entropy export) is the mechanism that leads to self-organization. This means that a self-organizing system imports high-quality (usable) energy from the environment and exports entropy back to it. Prigogine formulated a new worldview. He saw the world as an irreversible "becoming", which produces novelty without end. This is the opposite of the Newtonian reduction to a static framework, i.e., to the "being" view. This point of view is compactly expressed by *Prigogine's* quote: "The irreversibility of time is the mechanism that brings order out of chaos". Speaking about chaos, *James Gleick* [17] wrote that "Where chaos begins, classical science stops". In other words, this means that chaos is our third great revolution in physical sciences after relativity and quantum mechanics.

*Prigogine and Stengers* [18] state that "order creation" at the macro-level is a way of dissipating (exporting) entropy caused by energy flux at the micro-level. For example, a whirlpool is formed spontaneously in a draining bathtub because in this way the potential energy of the standing water is dissipated better than a laminar (smooth) or turbulent (chaotic) flow [19].

As we have seen in Sect. 8.8, a nonlinear system has, in general, a multiplicity of attractors. Each one of these attractors corresponds to a self-organized configuration. Therefore, the study of self-organization is equivalent to the study of the system attractors' properties and dynamics. If the system starts out in a basin state, it will settle down to the corresponding attractor, but, if it starts between different "*basins*", it has the freedom to choose the basin and the attractor in which it will end up. This depends on the unpredictable fluctuations that may exist. The self-organized configuration is, of course, more stable than the configuration from which the system started. We call this phenomenon "*order from noise*" [20], but thermodynamicists [18] call it "*order through fluctuations*" or "*order out of chaos*".

The "*interplay between positive and negative feedback*, i.e., the self-organization mechanism, works in the same way as described in the previous chapter in connection with *adaptability* to the environment. Here, however, this constitutes an internal (esoteric) business of the system aiming at (and leading to) increased organization and order. Actually, self-organization takes place via existing feedback loops between system components (elements) and between components and the structures that are formed at the higher hierarchical levels. A necessary condition for this to occur is that the system is "nonlinear", as happens in living organisms, biochemistry (autocatalysis), and the behavioral systems in human society. Typically, self-organization starts with positive feedback. An initial fluctuation towards organization (order) is amplified and spreads quickly, until it affects the entire system. Once all elements of the system have "*aligned*" their behavior with the configuration created by the initial fluctuation, and the system has reached an equilibrium state, further growth of self-organization is not possible. This is because at this stage only changes that weaken the self-organized (dominant) configuration are possible, and the same mechanisms that reinforced that configuration will suppress the deviation (i.e., they will apply negative feedback) and return the system to its stable configuration. In more complex situations, there may exist several interlocking positive- and negative-feedback loops, i.e., changes in some directions are reinforced, and changes in other directions are suppressed. The final result of this process is very difficult to predict.

The *selective-retention* mechanism of self-organization ensures that the outcome of the interactions of the system components is not arbitrary but shows a "preference" for certain situations over others [8]. This is analogous to Darwinian evolution, which is based on the assumption that the environment acts on a population of organisms that compete for resources (in order to survive). The winners of this competition (those that most fit to obtain the resources) will be selected, the others are eliminated.

The second assumption of Darwinian evolution is that selection is carried out by the common environment of the competing organisms. However, in selective retention, there is no need to have a population of competing organisms (configurations). It works well even in "population-of-one" situations. A configuration (state) can be chosen or eliminated no matter if other candidate configurations are present. A single system can happen via a sequence of states or configurations. Some of them are selected (retained), while others are eliminated. Actually, the competition in selective retention is taking place between subsequent states of the same system, and more importantly, there is no need to assume the existence of an environment external to the state(s) under selection. Selective retention can occur in both living and nonliving systems. For example, a stone "prefers" to be in a stable state at the foot of a hill, instead of being in an unstable state on the top. A "cloud" of gas molecules in a vacuum will spontaneously diffuse, but a crystal in the same vacuum will maintain its crystalline form. The first configuration (i.e., the cloud) disappears; the second (i.e., the crystalline structure) is retained. An animal in an ecosystem prefers a situation that assures more food or minimizes the risk of being attacked by a predator.

## 9.4 Self-organized Criticality

*Self-organized criticality* (SOC) is an alternative name for the capability of complex systems to maintain a balance between "*order and disorder*", which is also called "*the edge of chaos*" (see Sect. 8.9). It is a common property of living beings to live at such an edge of chaos via self-organization. Our purpose in this section is to discuss a little more this feature of self-organized systems. Throughout the years, many biologists, nonlinear-systems researchers, and cyberneticians have attempted to explain the phenomenon of self-organized criticality and especially why and how a system moves on its own to such a state existing in the order-chaos spectrum.

Criticality, in general, is a state at which the properties of a system change suddenly, e.g., the critical gain in a control system determines the boundary (edge) of the stable region; higher gain leads the system to the unstable region. Another example of criticality is the case in which a structure moves from non-percolating to percolating or vice versa (where the system is subject to phase change). Percolation is a structure (or matrix) of parts in which a property appears that connects the opposite sides of a disconnecting structure by developing a path or disconnecting them into a fully connected structure by introducing an obstruction

(non-percolation). The edge at which this percolation/non-percolation change occurs is exactly the edge to which a self-organized system goes and obeys a *power distribution* law of effects, i.e., the smaller the effect, the more frequently it is occurring. This is actually the typical *self-similarity property* of all self-organized systems.

Examples of natural systems with self-organized criticality include: floods caused by interconnected valleys, forest fires in areas susceptible to lightning bolts, snow avalanches occurring on snowy hillsides, etc. Three such examples are illustrated in Fig. 9.1(a–c).

Self-organized criticality is the capability of a system to work in a manner by which it can approach closely to a critical point and then sustain itself at that point. Actually, there exist many alternative theories for explaining this movement of natural and biological systems to a self-organized critical state. Three of them are the following [19]:

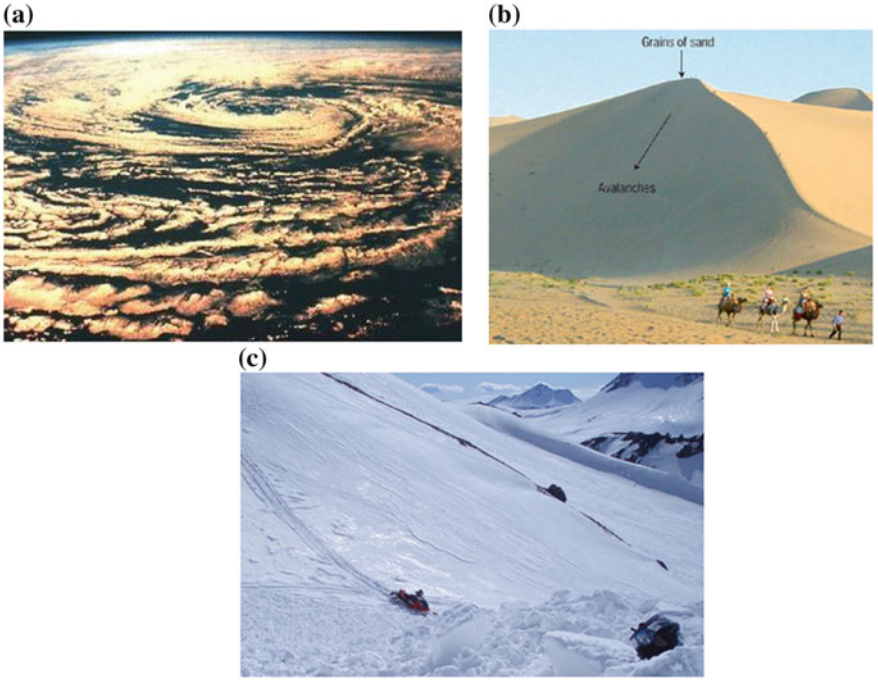



**Fig. 9.1** Three natural examples of self-organized criticality: **a** cyclone, **b** sand dune, **c** snow avalanche (http://www.newciv.org/pic/nl/artpic/10/1929/cyclone.jpg; http://www.nature.com/nmat/journal/v4/n6/images/nmat1405-f1.jpg; http://en.vedur.is/media/ofanflod/myndasafn/frodleikur/medium/P1010396%5B1%5D.JPG). The reader is informed that Web figures and references were collected at the time of writing the book. So some of them may no longer be valid due to change or removal by their creators

- **Stuart Kauffman** His explanation is based on the so-called "*coupled-fitness landscape*" which is a Boolean network of N cells, each one having S states with an overall of C possible connection paths to other cells [21, 22]. This N-cell system is mapped into a C-dimensional "landscape" that involves (topographically) all possible system states. According to Kauffman, the connectivity C is an index of how *orderly* or *chaotic* is a system. When C is very small, the system is "stuck" in its present state, and if C is very large, the system has chaotic behavior. If C has just the right size, the system can go to very high fitness peaks and achieve very good proficiency. The C-dimensional landscape may represent a genome, a population, etc.
- **Rod Swenson** His explanation is based on the law of maximum entropy production as explained in Sect. 8.3, which implies that ordered flow produces entropy faster than disordered flow.
- **P. Bak** He and his colleagues argued that a system self-organizes to a critical state without any "fine-tuning" process, but via a driving and dissipating process. Self-organized criticality seems to be the underlying concept for temporal and spatial scaling in dissipative nonequilibrium systems [23, 24]. Bak and colleagues explained that "power law distributions of phenomena" exhibit a self-organizing criticality performance. They studied several examples of systems in which this occurs. For example, they simulated the "sandpile" model which consists of sand poured on a table continuously until the occurrence of "mini avalanches". Similar results on the self-organized criticality have been derived by many other researchers (e.g., [25–27]).

## 9.5 Self-organization and Cybernetics

Self-organization was a topic of study in cybernetics right from its beginning. The term *cybernetics* comes from the Greek world "κυβερνώ/κυβερνήτης" (kyverno/kyvernitis = govern/governor). The founder of Cybernetics was *Norbert Wiener,* who defined it as follows: "*Cybernetics is control and communication in the animal and the machine*" (1947) [28]. The initial work of Wiener was related to the control of anti-aircraft fire. The gun should aim not at the present position of the aircraft, but at the point to which the aircraft will move during the flight time of the shell. He estimated this new position of the aircraft by collecting data about the discrepancies between predicted position and actual measured position and then feeding it back to the predictor. The result of his study is the celebrated *Wiener filter/predictor*. The field of cybernetics has attracted interdisciplinary interest. Scientists and engineers from a multitude of fields (physics, mathematics, operational research, biology, medicine, environment, psychology, anthropology, management, neurology, economics, sociology, ecology, computers, control, etc.), either individually or in multidisciplinary research groups, have derived important results in many directions. A comprehensive list of cybernetics and systems thinkers (cyberneticists or

cyberneticians) who have made substantial contributions to the field is provided in [29], and historical remarks about their role in [30]. Among them, in addition to *Norbert Wiener*, the father of cybernetics, those who have studied self-organization and closely related topics are the following:

- **W. Ross Ashby** One of the founders of cybernetics. He developed homeostasis, requisite variety law, and the self-organization principle [6, 31].
- **Henri Atlan** He studied self-organization in cells and networks and developed the theory of *random organization* according to which, at the birth of the universe, there was an order/disorder/organization dialogic triggered by calorific turbulence (disorder), in which under certain conditions (random encounters) organizing principles made possible the creation of nuclei, atoms, galaxies, and stars. The dialogue between order, disorder, and organization exists in a wide variety of forms, and via countless feedback processes is constantly in action in the physical, biological, and human worlds [32]. He contributed substantially to the development of *Biocentric Culture* governed by the "Vital Unconscious and Biocentric Principle".
- **Warren McCulloch** He developed mathematical models of learning and self-organizing neural networks. Together with *Walter Pitts,* he proposed the first model of a neural network, composed of functioning elements (neurons) and synaptic weights. This "artificial neuron" is known as McCulloch–Pitts neuron and is the foundation of most modern types of artificial neural networks [33].
- **Ilya Prigogine** He studied the thermodynamical approach to self-organization and coined the concepts of irreversibility and dissipative structures [16, 18, 34] discussed at many points in this book.
- **Heinz von Foerster** One of the founders of cybernetics. He was the first to study self-organization and self-reference and was the creator of second-order cybernetics [20, 25].
- **Humberto Maturana** He developed, together with *Francisco Varela*, the theory of *autopoiesis* and substantially contributed to complex systems theory [1].
- **Nikolas Luhmann** He applied the theory of autopoiesis to social systems [35].
- **Herbert A. Simon** He made major contributions to management, cognitive psychology, and complex systems theory [36]. Three important quotes of Simon are the following:
  - "I don't care how big and fast computers are; they are not as big and fast as the world."
  - "Learning is any change in a system that produces a more or less permanent change in its capacity for adapting to its environment."
  - "The social sciences, I thought, needed the same kind of trigger and the same mathematical underpinnings that had made the "hard" sciences so brilliantly successful."
- **Francis Heylinghen** He made important contributions to adaptation and self-organization [8, 37–39].
- **James Gleick** He reintroduced and reformulated chaos theory and contributed to the growth of interest in the modern science of complexity [17].

- **H. Haken** He studied physical phenomena that exhibit self-organization and revealed the synergetics mechanism between interacting components that leads to global patterns [14, 40].
- **Manfred Eigen** He studied the origin of life, including chemical self-organization and biological evolution. He was particularly interested in extremely fast chemical reactions induced in response to very short pulses of energy. The concept of "*hypercube*", i.e., an autocatalytic chemical cycle involving other cycles, as an explanation for the self-organization of prebiotic systems, was coined by him in 1971 in cooperation with *Peter Schuster* [41].
- **Gregory Bateson** He studied the parallelism between mind and natural evolution and developed the "double-bind" theory of the complexity of communication [42]. Some relevant statements of Bateson are the following:
  - "Logic is a poor model of cause and effect."
  - "Logic can often be reversed, but the effect does not precede the cause.
  - "It is impossible, in principle, to explain any pattern by invoking a single quantity."
  - "We do not know enough how the present will lead into the future."
- **Benoit Mandelbrot** He was the founder of *fractal geometry*, which, as we have seen in Sect. 8.8, describes the emergence of similar shapes or patterns at different scales that obey the power law of distributions of self-similarity [43] (see also: [44]).

To summarize the above aspects, we give in Fig. 9.2 the four main areas involved in the concept of self-organization.

Figure 9.3 shows two examples of self-organized crystal formation and an example of dissipative structure.

(a) Barium carbonate crystals (http://www.nanowerk.com/spotlight/id646.jpg),
(b) Crystals on cadmium (http://www.natureasia.com/asia-materials/article_images/227.jpg),
(c) Dissipative structures (http://www.filefestival.org/SITE_2007/RESOURCES/CONTENT/ILYA02.JPG).

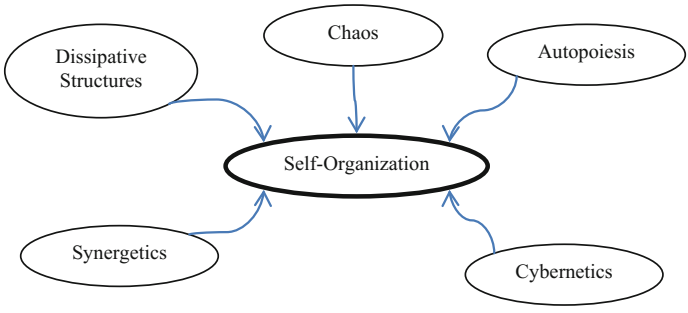



**Fig. 9.2** The basic areas of self-organization

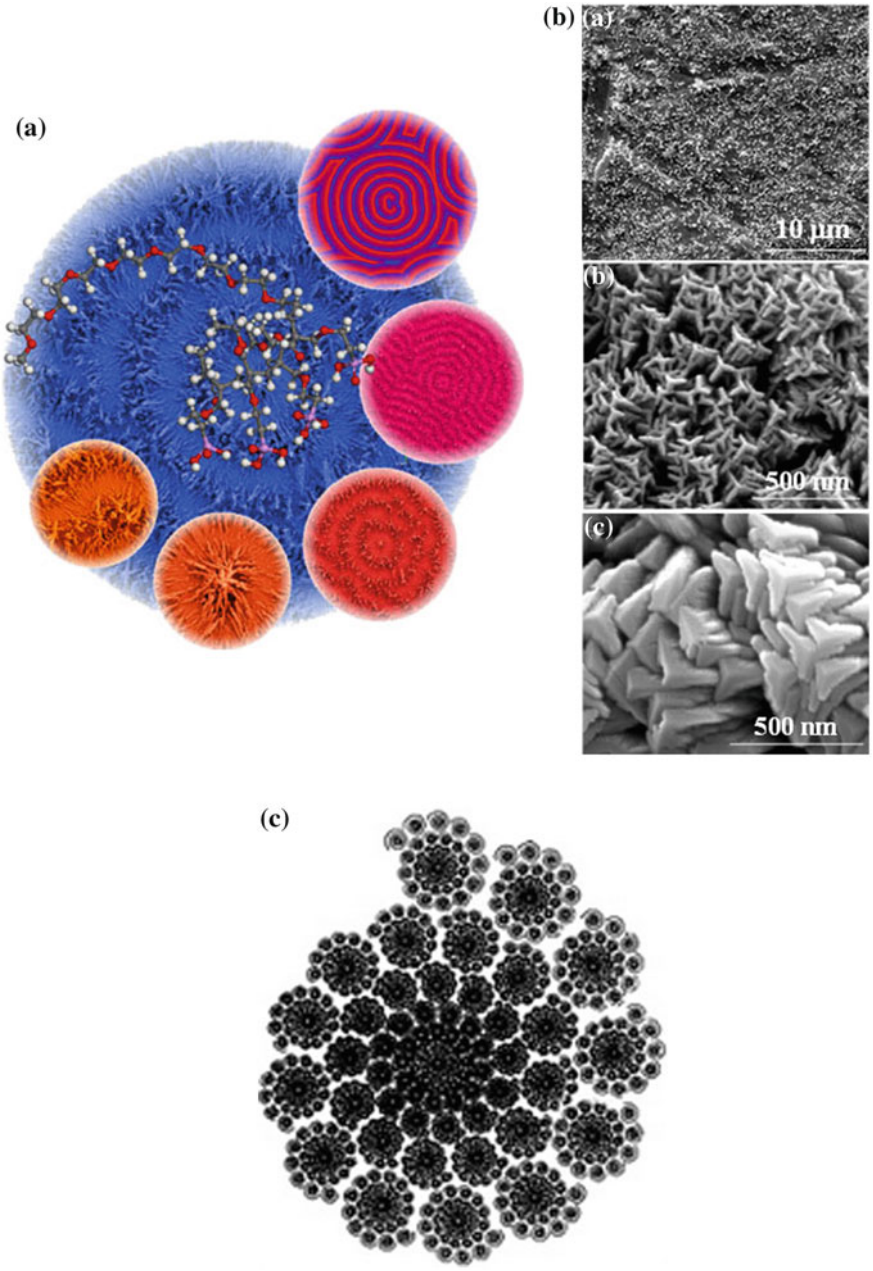



**Fig. 9.3** Natural self-organizing systems

## 9.6 Self-organization in Complex Adaptive Systems

Looking at the properties complex adaptive systems listed in most definitions available in the literature (see Chap. 8) one can verify that self-organization is one of the most fundamental common features of CASs together with adaptation, emergence, and complexity. The self-similarity property is a property of self-organization that is transferred to CASs. Other properties of self-organization that are transferred to CASs are the following:

- Interdependence
- Interaction
- Selective variety
- Modularity
- Clustering.

A short discussion of them follows.

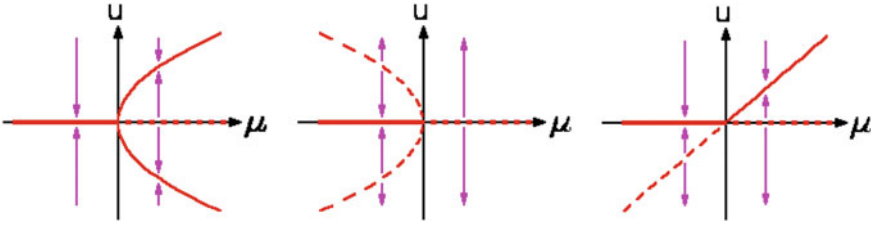
**Interdependence** This is a term indicating that the elements (parts, components, agents) of a complex system are related by interdependence relationships obeying the physical, biological, informational, systemic, psychological, economic, ecological laws, etc., depending on the scale and nature of them. Examples of such elements are molecules cells, systems states, animals, circuit elements, trees, human, etc.

Two elements that cooperate with each other are interdependent and interconnected. Complete dependence (as, e.g., in a crystal where the state of one molecule determines the states of all the others) may imply *full order*. Complete independence (like the molecules of a gas) implies *full disorder,* in which case the state of a molecule cannot give any information about the states of the other molecules. Interdependence is a concept very common in human societies, enterprises, economies, and countries interconnected to achieve common goals. It should be noted that to achieve self-organization, a system must be neither too loosely interconnected (in which case most elements are independent), nor too strongly interconnected (in which case most elements influence one another). For example, in Boolean networks, the optimum self-organization is obtained when there are two connections per element (unit) [9].

**Interaction** This is a property closely connected (but not identical) to the interdependence of two or more elements (objects, agents) that are acting so as to have an effect upon one another. These interactive actions, when combined and integrated, produce very important emergent outcomes. The interaction is usually a purposeful process aiming at maximizing the fitness, utility, and productivity of both the individual elements and the entire system or organism. Even if no specific purpose exists, the elements of the system act according to the input excitation that is received from the environment. This perception-action (causal) process (or rule) is effected initially by some type of elements, but then through the interdependence and interaction that is extended to other element types utilizing some learning or

evolutionary variation. The interaction process is implemented through communication and feedback among the system elements. Examples of interactions taking place in self-organized systems include: interaction of drugs in medications (pharmacodynamic or pharmacokinetic interactions), physical interactions (elementary particles' interactions via exchanging gauge bosons, interactions of charged particles via electromagnetic-field mediation, gravitational interactions), sociocultural interactions between individual persons, groups or larger societies and populations and genetic interactions (combined mutations affecting or not affecting the genotype), etc. [45, 46].

**Selective variety** This is the self-organization principle according to which the wider the repertoire of configurations a system has available for selection, the higher the probability that one or more of them will be selectively retained. This principle is the theoretical expression of the "selective-retention" mechanism discussed in Sect. 9.3. It implies that to increase the probability of achieving self-organization (and speed-up the process), a larger variety of configurations should be available for the system to pass through. After self-organization, one configuration dominates all others, which means that the system symmetry existing in the disorganized situation is lost. Of course, it must be noted that there does not exist well-known criteria for the preference of one stable configuration over another. Thus, it appears that the system has made an arbitrary decision via which it has changed the repertoire of possibilities. Actually, it is this unpredictability that (in some sense) produces the observed novelty. This selectivity phenomenon is also called a *bifurcation* (or *branching*) in the possible configuration. When a control parameter μ (called "*bifurcation parameter*" or "*order-parameter*") increases to a certain critical value $\mu_c$, there are two possible outcomes for the dependent variables u, i.e., to go upwards or downwards. In general, a bifurcation is a change in the number of candidate operating conditions of a nonlinear system that occurs as μ is quasi-statically varied. The depiction of the equilibrium points and limit cycles of a system plotted against the bifurcation parameter is known as the "bifurcation diagram". A bifurcation can be *super-critical*, *sub-critical,* or *trans-critical* depending on the direction of bifurcation, as shown in Fig. 9.4 [47]. The trans-critical bifurcation appears when, in the combined space of phase space and bifurcation




(**a**) Super-critical (**b**) Sub-critical (**c**) Trans-critical

**Fig. 9.4** Pitchfork bifurcations: **a** Supercritical, **b** Sub-critical, **c** Trans-critical. (https://elmer.unibas.ch/pendulum/pbif.gif)

parameter space, two different manifolds of fixed points cross each other. At the crossing point, the unstable fixed point becomes stable, and, vice versa, the stable fixed point becomes unstable.

It is noted that a trans-critical bifurcation is unlikely to occur in a space higher than two-dimensional (because in such a space two lines are unlikely to cross each other). When the nominal operating point exists, both before and after the critical parameter value, we say that a stationary bifurcation occurs "from a known solution". If the nominal solution disappears beyond the critical parameter value, we say that a stationary bifurcation occurs "from an unknown solution". The former is usually called simply a "*stationary bifurcation*".

To illustrate how a bifurcation occurs, we consider the following simple nonlinear system:

$$\dot{u} = \mu u - u^3$$

where $\mu$ is the bifurcation parameter. Clearly, the origin $u_0 = 0$ is an equilibrium point for any value of $\mu$. But, as μ increases from $\mu = 0$, the origin loses stability, since, at $\mu = \mu_c = 0$, a bifurcation occurs, and two more equilibrium points are possible. This pair of equilibrium points is found by solving the equation $(\mu - u^2)$ $u = 0$ for $u \neq 0$, i.e., they are $u_1 = +\sqrt{\mu}$ and $u_2 = -\sqrt{\mu}$. We say that this pair of equilibrium points is bifurcated (branched) from the origin for the critical value $\mu_c = 0$ of $\mu$ (i.e., the equilibrium point "breaks"), as shown in Fig. 9.4a. If the nonlinear system under consideration is

$$\dot{u} = \mu u + u^3,$$

the direction of bifurcation is reversed resulting in the subcritical bifurcation shown in Fig. 9.4b.

In complex systems, there may be more than two alternatives solutions (configurations) for selection at the bifurcation point. The increase in the number of possible configurations that follows the increase in the order parameter μ can be regarded as an increase in general variability, which facilitates the self-organization process. This is a special case of "order-from-noise" or "order-out-of chaos" processes [8, 37, 39].

**Modularity** This is a general principle for managing complexity. That is, to manage a large number of systemic interconnections, a complex system is broken into discrete subassemblies which communicate with one another via standard channels within a standardized structure or architecture. Modularity is an inherent feature of many living organisms, but today it is extensively used in man-made complex systems and social systems. According to *F.A. Hayek* [48]: "Complexity is a function of the minimum number of elements of which an instance of the pattern must consist in order to exhibit all the characteristic attributes of the class of patterns in question." According to *Herbert Simon* [36]: "A complex system is one made up of a large number of parts that interact in a non-simple way," and so

complexity is a matter of both the number of distinct parts contained in the system and the nature of the interconnection or interdependencies between these parts. *Simon* states that the *criterion of decomposability* (i.e., of grouping the system's elements in a smaller number of subsystems) in modular design can be provided by a person or drawn from the systems available (ready- made) in nature. In a non-decomposed system, the correct working of a given part depends on the performance of other parts with high probability, but in a decomposed system, this effect occurs with much lower probability. Therefore, a decomposable system may continue to work (of course, suboptimally) even if some subsystems are damaged or are incomplete. In other words, decomposable systems have the important feature of *fault/failure tolerance* and *robustness*.

In engineering and societal systems, the interaction between the parts can be considered to be an issue of information exchange or communication. For example, in computer systems, the decomposition of a system into modules can be done through partitioning of information into *visible design rules* and *hidden design rules* [49, 50].

- **Visible design rules** These rules consist of three parts, viz.,: *architecture* (identify the modules/functions/structure of the system), *interfaces* (ways of interaction and communication and fitting of modules), and *standard test* (conformity of modules to design rules and measure of relative performance of modules).
- **Hidden design rules** These rules are embedded within the modules without the need of being communicated to other modules, but only within the boundaries of the module.

According to *Richard Langlois* [50], the three parts (architecture, interfaces, standard test) are collectively called "*modularization*". Regarding modularity in social systems, the design rules of interaction are the so-called "*social institutions*", which (among others) determine how much a society is a modular system [51]. Modularity in social systems has been a topic of study since the 1960s. For example, *Adam Smith* [52] coined a decentralized concept which, he believed, would lead to economic growth enforced by learning, evolution, and further division of labor. Smith stated that his decentralized system is "the obvious and simple system of natural liberty."

**Clustering** This is a concept similar to modularization and has been developed in the field of complex networks (electrical networks, social networks, political networks, etc.). A *cluster* is a group of elements (components, agents, etc.) that are interacting and usually have similar goals, beliefs, values, etc. Clustering means that, if the element A is connected to B and B to C, then there is a high probability that A is also connected to C (this probability is always higher than the corresponding probability in a random (non-clustered) network. To understand better the concept of clustering, we consider the case of social networks (groups, societies, etc.). Here, clustering can be interpreted as, e.g., "the friends of my friends are (likely to be) my friends" [8]. Expressing this type of clustering in another way, we

have a social cluster or community if, e.g., everyone knows everyone, because when one meets regularly his/her friends, he/she has the chance to meet their friends as well. In general, if an entity A interacts frequently with an entity B, and B interacts with C, then with high probability A will interact (sooner or later) with C as well. If A and B have similar (or identical) goals, it is quite likely to act in a synergetic way, and the same is true for B and C, and consequently, between A and C. Scientific societies and worker syndicates operate in this cluster-wise manner.

## 9.7 Examples of Self-organization

It is generally accepted, on the basis of deep studies and extensive experiments, that most natural systems are self-organizing systems. Therefore, in their effort to understand how these self-organizing systems work, and what is their performance, scientists and engineers have designed and built appropriate computer-based models and simulators that involve embedded multiple agents, local interactions, multiple connections, positive and negative feedback loops, and synergetic features. The states and outputs of these simulators are monitored, analyzed, and evaluated over time using suitable human–machine interfaces. Therefore, these models and simulators provide the technological tools for the qualitative and quantitative study of self-organized systems. Some examples of such models and simulators include flocks, herds, swarms, plant communities, predator–prey interactions, plant-herbivore communities, social-insect colonies, fractal river basins, salmon propagation in rivers, immune systems, cellular automata, etc. [19, 53–62]. Especially, Holland's *ECHO* (ecological) CAS Model [59] has found great utility and many applications [60, 63].

Our purpose in this section is to provide a brief outline of a few natural and man-made self-organizing systems, namely:

- Ecological systems
- Magnetization
- Heated liquids
- Linguistic systems
- Knowledge networks
- Self-organizing maps.

**Ecological systems** This example illustrates very well the difference between the classical "*top-down*" from the "*bottom-up*" (self-organizing) modeling of life in ecological systems [19]. In the top-down methodology, the phenomena are studied using parameters from the higher hierarchical levels. For example, predators are studied as homogeneous populations that uniformly impact homogenous prey population. Trees are not studied individually but as *patches of trees*. This top-down approach violates two of the basic aspects of biology, namely, *individuality* and *locality,* which implies that population evolution is the result of activity

at the level of the individual and the range of locality. Actually, individual members in a population have clear differences (e.g., body size, reproduction rate, etc.) that may have cascading and amplifying effects at higher levels. Tree gaps that result when a tree falls in the tropics may produce severe ecological changes in the region of the gap (due, e.g., to a gap in the canopy). Clearly, seeds of the forest do not have an equal chance of germinating in the gap. Ignoring this fact (i.e., locality) may mask the factors that affect spatial and temporal ecological dynamics. For example, seeding found in a high-rainfall region may have better-growing conditions than those that exist in the dry soil. The possibility of increased moisture-detention is very high, which may result in the creation of new landscape patterns. This is an example of the validity of the ecological principle that "*pattern affects process*", which is one of the self-organization mechanisms in ecological systems [64, 65].

**Magnetization** This is a simple self-organizing system used by many authors to illustrate the basic physical mechanism of self-organization [37]. Consider a potential magnetic material (e.g., a piece of iron), which consists of a huge number of microscopic magnetic "dipoles" (known as "spins"). At high temperatures, these ferromagnetic dipoles move quite randomly (i.e., they are disordered), and the orientations of their magnetic fields are random and cancel each other, resulting in a non-magnetized overall configuration (state) of the material (see Fig. 9.5a). But, if the temperature is lowered, the "spins" are spontaneously aligned and point in the same direction (Fig. 9.5b). The outcome of this alignment is that now the magnetic fields add up, producing a strong, overall magnetic field.

This preference of the spins is due to the fact that dipoles pointing in the same direction attract each other (the north pole of one magnetic dipole attracts the south pole of another dipole), while dipoles with opposite direction repel each other. This spontaneous alignment (magnetization) process shows that "self-organization" is occurring. In other cases, such as "crystallization", the self-organization involves not only the orientations but also the positions of the molecules which are evenly arranged.

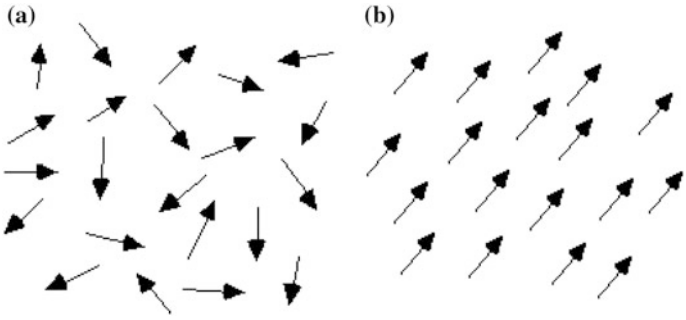



**Fig. 9.5** Self-organization leading to magnetization **a** disordered spins, **b** ordered spins

**Heated Liquids** A liquid contained in an open container is heated evenly from below (via a hot plate) [37]. Hot liquid is lighter than cold liquid, and so it tends to move upwards. Similarly, the cold liquid tries to sink to the bottom (convective instability). These two opposite movements take place in a self-organized way in the form of parallel "rolls" with an upward flow on one side of the roll and a downward flow on the other side. Initially, the molecules of the liquid have a random movement, but finally, all "hot" molecules are moving upward on the one side of the roll and "cool" molecules are moving downward on the other side as shown in Fig. 9.6.

This self-organizing process was first observed by *Bénard* and is known as the "*Bénard phenomenon*". In this example, the molecules after self-organization keep in perpetual motion, whereas the magnetic dipoles, in the magnetization example, after self-organization, do not move (the spins are "frozen").

### *9.7.1 Linguistic Self-organization*

Here again, a super macro-global structure is the result of local interactions. Self-organizing issues in linguistics include [66]:

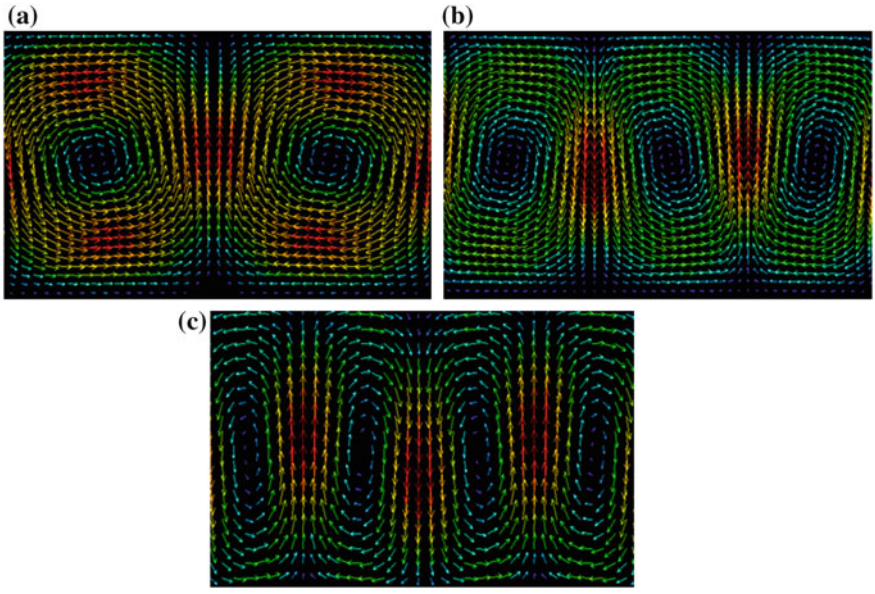



**Fig. 9.6** Three steps in the self-organization movement process of the liquid molecules: **a** Rayleigh number: $R_a = 2084$, **b** $R_a = 2603$, **c** $R_a = 9215$ (http://hmf.enseeiht.fr/travaux/CD0001/travaux/optmfn/hi/01pa/hyb72/rb/rb.htm)

- Decentralized generation of lexical and semantic conventions in populations of agents.
- Formation of conventionalized syntactic structures.
- Conditions for selection of systematic reuse.
- Shared inventories of vowels or syllables in groups of agents.

To study how self-organization takes place in linguistics, suitable operational models are constructed that explicitly involve the set of assumptions and show how their consequences (conclusions) are calculated.

Self-organization in linguistics occurs in the following [66]:

- **In the emergence of language** (information senders and receivers, compositionality, the ability to sustain cultural progress cumulatively).
- **In language acquisition** (through the ability to see others as intentional agents, or through joint attention of actions).
- **In articulatory phonology** (speech production via a coordinated set of gestures known as "constellations").
- **In diachrony and synchrony** (dynamic or self-organizing models of language evolution).

### *9.7.2 Knowledge Networks*

*Knowledge networks* [8] belong to the area of *information science* (see Sect. 5.2) and refer to the documentation items in libraries and databases worldwide. Documents (papers, books, reports) are typically produced by authors and researchers working in defined fields and building further on the result of other authors. This knowledge-producing system is actually a self-organized system, because it is not controlled centrally, but is generated spontaneously by local interactions of the individuals or groups that produce the new knowledge. The networks are formed by the researchers, the concepts used, and the publications linked directly or indirectly by corresponding relations (e.g., citations, collaboration, and information exchange). The new knowledge (patterns) are generated via the nonlinear interactions of multiple autonomous agents (scientists, groups, organizations), and the overall system is a "*heterogeneous network*" involving three different kinds of nodes, viz.: *agents* (individual scientists, groups, organizations), *containers* (documents, papers, etc.), and *concepts* (keywords, abstract knowledge items). Knowledge networks can be viewed as complex adaptive systems and can be designed and operated by CASs techniques.

### 9.7.3 Self-organizing Maps

*Self-organizing maps*(**SOMs**) [67, 68] represent a special class of artificial neural networks that are based on competitive learning, in which the network's output neurons compete among themselves to be fired or activated such that only one output neuron is activated (on) at each time. The name self-organizing map is due to the fact that the impact patterns are mapped on a topographical map where the spatial locations (coordinates) of the neurons indicate the various inherent features of the input patterns. An SOM transforms an input signal pattern of arbitrary dimension into a one-or two-dimensional discrete map and performs this transformation in a topologically adaptive way. The topographical mapping of the input patterns can be done as suggested by Kohonen and shown in Figs. 9.7, 9.8. This mapping is known as a Kohonen SOM or Kohonen model. It is clear that each neuron has a set of neighbors.

Each input pattern consists of a localized region or "spot" of activity against a quiet background. Since the location and nature of "spots" are different from one input pattern to another, to ensure that the self-organization process will be properly established, all neurons must receive a sufficient number of different realizations of the input pattern. The formation of the SOM starts by initializing the network's synaptic weights randomly, with small values provided by a random-number generator. After this random initialization, the SOM formation is formed via three basic processes:

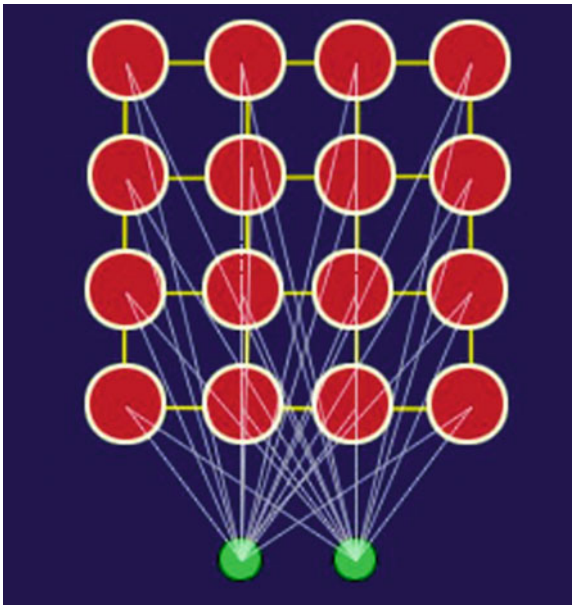


**Fig. 9.7** A first representation of a Kohonen self-organizing map. (http://www.ai-junkie.com/ann/som/som1.html)

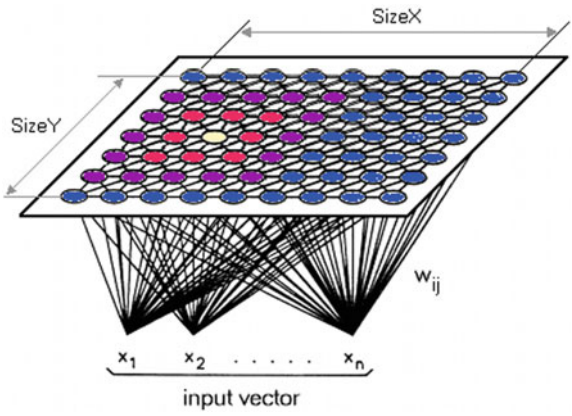



**Fig. 9.8** The winning node is the pink one (http://www.ai-junkie.com/ann/som/som1.html http://www.lohninger.com/comimg/kohonen1.gif)

- **Competition** A fitness (discriminant) function is computed for each input pattern by the neurons. The neuron with the largest fitness value is declared the winner of the competition.
- **Cooperation** The winning neuron specifies the location of a neighborhood of fired neurons, which are allowed to cooperate.
- **Synaptic adaptation** The excited neurons increase the value of the fitness function through suitable adjustments of their synaptic weights. In this way, the winning neuron enhances its response to similar patterns subsequently entered into the neural network/SOM.

The SOM starts from a completely disordered initial state and leads to an organized representation of activation patterns drawn from the input space. Kohonen [67, 68], pointed out that this is performed in these two phases:

- Ordering (self-organizing) phase
- Convergence phase.

Mathematical details of Kohonen's SOM algorithm can be found in standard textbooks on neural networks (e.g., [69–71].

## 9.8 Concluding Remarks

In this chapter, we have provided a tour to the basic concepts and principles of self-organization, which occurs in natural, biological, and societal systems. Self-organization is performed (and needed) in complex systems and, together with adaptation, in complex adaptive systems. We have seen that the mechanisms by which self-organization is realized are synergy, entropy export, positive/negative feedback interplay, and selective retention. Other self-organizational properties that are also possessed by complex adaptive systems are interdependence, interaction, selective variety, modularity, and clustering.

Complexity implies a lack of symmetry which exists in both full disorder and full order, but the midpoint between order and disorder which specifies the complexity depends on the level of representation. Something that appears to be complex in one representation may not seem complex in a representation on another scale [72]. A fractal is self-similar, i.e., its shape is independent of scale, something which is not valid in the case of a simple system like a building which, to an outside observer, is seen to be different at several scales: the entire building, the doors/windows, the rooms, and the bricks represent four different scales. Typically, an observer picks up those distinctions (features) that are in some sense the most important and creates categories of similar processes (neglecting the existing differences among the members of each category). Thus, the increase or decrease of complexity depends on which distinctions the observer is introducing [5, 6, 38].

We have seen that an increase in variety (which is called *differentiation*), or an increase in the *connectivity* (which is called *integration*) of a complex system facilitates and speeds-up the process of self-organization. Evolution and adaptation are based on (and produce) differentiation and integration along with several dimensions, viz., space, spatial scale, time, and timescale, leading to the so-called structural, hierarchical, functional, and functional hierarchical differentiation/integration, respectively [8, 37, 39].

To recapitulate, a self-organizing system has the following (surely not exhaustive) features:

- Autonomy (absence of external ordering or controlling agent).
- Self-configuration (autonomous arrangement of system's constituent past).
- Dynamic performance (time-evolving operation).
- Spontaneous order (emerging from local interactions).
- Synergy (mutual coevolution adaptation of local agents).
- Perturbations (noise/fluctuations, order-from-noise).
- Complexity ("*paradox*" *phenomena*).
- Nonlinearity (multiple "attractors", bifurcations).
- Dissipation (far from equilibrium, extropy).
- Self-organized criticality (edge-of-chaos operation).
- Selectively variety (selective retention).
- Positive/negative feedback interplay.
- Self-similarity (power-law distribution).
- Commonly understood action (at all levels).
- Redundancy (robustness to faults and damages).
- Self-maintenance (reproduction/repair).
- Symmetry-breaking (heterogeneity).
- Differentiation and integration.
- Modularity and clustering.
- Self-reference (the system's behavior is evaluated with respect to the system itself).

We close our discussion by noting that self-organization is typically achieved through distributed (non-centralized) control. This means that there is not a unique external or internal controller that drives the system towards self-organization. Rather, all parts of the system contribute smoothly to the resulting self-organized configuration.

Today, most industrial and other man-made controllers are centralized, and those which are decentralized or distributed work in a deterministic (reductive/cause-effect) way. But as *Jim Pinto* argues [73], drawing from [17], the advent of self-organizing industrial controllers (i.e., controllers working with mechanisms and principles of natural-like self-organization) will mark the end of deterministic and centralized controllers. The main reason for this is the fact that conventional centralized or decentralized controllers and DCSs cannot be sealed (i.e., they do not have the necessary self-similarity property). Therefore new peer-to-peer I/O-based self-organizing controls are the controls of the future. In these controllers and systems, the overall behavior must be the result of the interactions of the individual elements (components, agents, computer programs), which both decompose and integrate the control/performance problem. This means that all man-made self-organizing systems should have *autonomy* (partly or fully) and be able to operate successfully without the need for an external designer. Examples of such man-made (engineered), self-organizing systems (with features like the ones of natural systems) include many robotic systems such as *robot swarms* and *robot groups* [74], etc. These systems are characterized by prediction (anticipatory control), adaptation (adaptive control), robustness (robust control), and general intelligence (intelligent control) as described in Chap. 7. *Artificial-life* (**Alife**) systems form a class of man-made systems that exhibit properties and behaviors characteristic of living organisms, i.e., they synthesize life-like behaviors within computer and control science and engineering. As the founder of Alife, *Chris Langton* stated [2]: "by extending the empirical foundation upon which biology is based *beyond* the carbon-chain life that has evolved on Earth, Artificial Life can contribute to theoretical biology by locating "*life-as-we-know-it*" within the larger picture of "*life-as-it-could-be*"… Only when we are able to view *life-as-we-know-it* in the larger context of *life-as-it-could-be* we will really understand the nature of the beast."

Thus, *Alife* is a relatively new field employing a synthetic approach to the study of life-as-it-could-be. Alife differs substantially from artificial intelligence. The most important philosophical aspects of this area are coming from biology, not from psychology, and it complements traditional/theoretical biology in two ways, namely: (i) it deals with the synthesis of life-like behavior (further to the analysis of biological processes for which biology is concerned), and (ii) it aims at exploring the possibilities of life-as-it-could-be. That is, Alife explores and studies the total range of mechanisms that can aid such a synthesis, independently of their similarity or not of what we see in the actual biosphere. An important reference work on Alife is [3]. Two references on self-organization available on the Web are [75, 76]. Some references on history, principles, simulation, and global patterns of self-organization are [77–81].

## References


1. H. Maturana, F. Varela, *The Tree of Knowledge* (Shambhala, Boston, 1992)
2. C. Langton (ed.), Artificial Life, in *Proceedings 1st A Life Conference*, Santa Fe (Addison – Wesley, Redwood City, CA, 1989)
3. S. Levy, *Artificial Life: The Quest for a New Creation* (Jonathan Cape, London, 1992)
4. Organization (noun), Longman Dictionary of Cotemporary English. http://www.idoceonline.com/Organizations-Topic/organization
5. W. Ross, Ashby, principles of the self-organizing dynamic system. J. Gen. Psych. **37**, 125–128 (1947)
6. W.R. Ashby, Principles of the self-organizing system, in *Principles of Self-organization* (Pergamon Press, Oxford, pp 255–278, 1962). (Also in: *E.C.O. Special Issue*, vol 6(1–2), pp 102–126, 2004)
7. A. Schlemm, Self Organization in Society. http://www.thur.de/philo/ensoges/htm
8. F. Heylighen, Complexity and self-organization. in *Encyclopedia of Library and Information Sciences,* eds. by M.J. Bates, M.N. Mack (Taylor and Francis, London, 2008)
9. C. Lucas, Self-organization FAQ, (May 1997). http://psoup.math.wisc.edu/archive/sosfaq.html
10. S. Camazine, J.L. Deneubourg, N.R. Franks, J. Sneyd, G. Theraulaz, E. Bonablau, *Self-organization in Biological Systems* (Princeton University Press, Princeton, N.J., 2001)
11. A.N. Whitehead, *Symbolism: Its Meaning and Effect* (MacMillan, London/ Oxford, 1927)
12. M.B.L. Dempster, A self-organizing systems perspective on planning for sustainability, *Master Thesis*, University of Waterloo, Waterloo, Ontario, Canada, 1998. http://www.bethd.ca/pubs/mesthe.pdf http://www.nesh.ca/jameskay/ersserver.u.waterloo.ca/jjkay/grad/bd
13. T. Imada, *Self-organization and Society* (Springer, Berlin, 2008)
14. H. Haken, *Information and Self-organization: A Macroscopic Approach to Complex Systems* (Springer, New York, 2000)
15. H.J. Zeiger, P.L. Kelley, Lasers, in *The Encyclopedia of Physics* ed. by R. Lerner, G. Trigg (VCH Publishers, Chichester, U.K., pp 614–619, 1991)
16. I. Prigogine, G. Nicolis, *Self-organization in Non-Equilibrium Systems* (Wiley, New York, 1997)
17. J. Gleick, *Chaos: Making a New Science* (Cardinal, London, 1987)
18. I. Prigogine, I. Stengers, *Order out Chaos* (Bantam Books, New York, 1984)
19. E.H. Deker, Self-organizing systems: a tutorial in complexity, department of biology, University of New Mexico, Albuquerque, US. http://www4.ncsu.edu/∼debrown/sos.html
20. H. Von Foerster, On Self-Organizing Systems and Their Environments, in *Self-Organizing Systems,* ed. by M.C. Yovits, S. Cameron (Pergamon Press, London, pp 31–50, 1960)
21. S.A. Kauffman, Co-evolution to the edge of chaos: coupled fitness landscapes, poised states, and co evolutionary avalanches. J. Theoretical Biology **149**, 467–505 (1991)
22. S.A. Kauffman, *At Home in the Universe* (Oxford University Press, Oxford, 1995)
23. P. Bak, C. Tang, K. Wiesenfeld, Self-Organized Criticality. Phys. Rev. A **38**, 364–374 (1988)
24. P. Bak, *How Nature Works: The Science of Self-Organized Criticality* (Springer, Berlin, 1996)
25. R.V. Sole, S.C. Manrubia, Are rainforests self-organized in a critical state? J. Theor. Biol. **173**, 31–40 (1955)
26. H.J. Jensen, *Self-Organized Criticality* (Cambridge University Press, Cambridge, 1998)
27. R.V. Sole, O. Miramontes, Information at the edge of chaos in fluid neural networks. Physica D **80**, 171–180 (1995)
28. N. Wiener, *The Mathematical of Self-Organizing Systems: Recent Developments in Information and Decision Processes* (MacMillan, New York, 1962)
29. Cybernetics and Systems Thinkers, *Principia Cybernetica Web*. http://pespmcl.vub.ac.be/CSTHINK.html#Foerster

30. History of Cybernetics and Systems Science. http://pespmc1.vub.ac.be/cybhist.html, http://www.answers.com/topic/history-of-cybernetics
31. W.K. Ashby, *Design for the Brain, The Origin of Adaptive Behavior* (Chapman and Hall, London, 1966)
32. H. Atlan, Immune information, self-organization and meaning. Int. Immunol. **10**(6), 711–717 (1988). http://worldcat.org/identities/lccn-n82-129180
33. W. McCulloch, W. Pitts, A logical calculus of the ideas immanent in nervous activity. Bull. Math. Biophys. **5**, pp 115–133 (1943). http://philosophy.uwaterloo.ca/MindDict
34. I. Prigogine, *The End of Certainty* (The Free Press, New York, 1997)
35. N. Luhman, *Essays on Self-Reference* (Columbia University Press, Columbia, 1990)
36. H.A. Simon, The architecture of complexity. Proc. Amer. Philos. Soc. **106**, 467–482 (1962) (also: In: *The Sciences of the Artificial*, MIT Press, Cambridge, 1981)
37. F. Heylinghen, The science of self-organization and adaptivity, in *Encyclopedia of Life Support Systems* (EOLSS Publishers, Oxford, 2001). http://www.edss.net
38. C. Gershenson, F. Heylinghen, When can we call a system self-organizing? in *Advances in Artificial Life*: *ECAL-2003, Dortmund, Germany*, ed. by W. Banzhaf, et al. (Springer-LNAI, Berlin, 2003), pp. 606–614
39. F. Heylighen, Principles of systems and cybernetics: an evolutionary perspective, in *Cybernetics and Systems '92* R, ed. by Trappl (World Scientific, Singapore, pp 3–10, 1992)
40. H. Haken, Synergetics: An Introduction: Non-equilibrium Phase Transition and Self-organization in Physics, Chemistry and Biology, Springer, Berlin, 1983
41. M. Eigen, P. Schuster, *The Hypercycle: A Principle of Natural Self-organization* (Springer, Berlin, 1979)
42. G. Bateson, *Mind and Nature: A Necessary Unity, Series on Advances in Systems Theory, Complexity and Human Sciences* (Hampton Press, N.J., Cresskill, 1979)
43. W.H.B. Mandelbrot, *The Fractal Geometry of Nature* (Freeman Press, New York, 1983)
44. I. Havel, Scale dimensions in nature. Intl. J. Gen. Syst. **23**(2), 303–332 (1995)
45. Interaction-Wikipedia. http://en.wikipedia.org/wiki/interaction
46. B.L. Dress, et al, Derivation of genetic interaction networks from quantitative phenotype data. Genome Biol. **6**(4), R.38 (2005). http://genomebiology.com/2005/6/4/R38
47. E.H. Abed, H.O. Wang, A. Tesi, Control of Bifurcation and Chaos, in *The Control Handbook*, ed. by W.S. Levine (CRC Press/IEEE Press, New York, 1996), pp. 951–966
48. F.A. Hayek, *Studies in Philosophy, Politics, and Economics* (The University of Chicago Press, Chicago, 1967)
49. C.Y. Baldwin, K.B. Clark, Managing in an age of modularity. Harvard Bus. Rev. **75**(5), 84–93 (1997)
50. R.N. Langlois, Modularity in Technology, Organization and Society, *Working Paper 1999-05*, Department of Economics, University of Connecticut, August, 1999
51. R.N. Langlois (ed.), *Economics as a Process, Essays in the New Institutional Economics* (Cambridge University Press, Cambridge, 1986)
52. A. Smith, *An Enquiry into the Nature and Causes of the Wealth of Nations* (Clarendon Press, Glasgow Edition, Oxford, 1976)
53. M. Mitchell, S. Forrest, Genetic algorithms and artificial life. Artif. Life **1**, 267–289 (1994)
54. R.D. Boids, Flocks, Herds and Schools: A distributed Behavioral Model, 1999. http://www.red.com/cwr/boids.html
55. D. Hiebeler, The Swarm Simulation System and Individual- Based Modeling, *Working Paper 94-12-065*, Santa Fe Institute, 1994. http://cam.cornell.edu/hiebeler/swarm-paper.html
56. T. Smith, M. Huston, A theory of the spatial and temporal dynamics of plant communities. Vegetation **83**, 49–69 (1989)
57. E. McCauley, W.G. Wilson, A.M. de Roos, Dynamics of age-structured and spatially structured predator-prey interactions: individual–based models and population-level formulations. Am. Nat. **142**, 412–442 (1993)
58. C. Furusava, K. Kaneko, Origin of complexity in multicellular organisms. Phys. Rev. Lett. **84** (26Pt1), 6130–6133 (2000)

59. J.H. Holland, *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence* (MIT Press, Cambridge, MA, 1975)
60. P.T. Hraber, T. Jones, S. Forrest, The ecology of echo. Artificial Life **3**, 165–190 (1997)
61. G. Hartvigsen, S.A. Levin, Evolution and spatial structure interact to influence plant-herbivore population and community dynamics. Proc. Royal Soc. London[B] **264**, pp 1677–1685 (1997)
62. M.R. Cross, Salmon breeding behavior and life history evolution. Ecology **72**, 1180–1186 (1991)
63. D.L. Harris, Echo Implemented: A Model for Complex Adaptive Systems Computer Experimentation, *Sandia National Laboratories*, *SAND 2001–2017, 2001*
64. C.B. Huffacer, Experimental studies on predation: dispersion factors and predator-prey oscillations. Hilgargia **27**, 343–383 (1958)
65. A.S. Watt, Pattern and Process in the Plant Community. J. Ecol. **35**, 1–22 (1947)
66. Self-organization: Wikipedia. http://en.wikipedia.org/wiki/Self-organization
67. T. Kohonen, Self-organized formation of topologically correct feature maps. Biol. Cybern. **43**, 59–69 (1982)
68. T. Kohonen, Self-Organizing Maps, Springer, Berlin, 1997
69. S. Haykin, *Neural Networks: A Comprehensive Foundation* (Prentice Hall, Upper Saddle River, N.J., 1999)
70. I. Aleksander, H. Morton, *An Introduction to Neural Computing* (Chapman and Hall, London, 1990)
71. J.A. Anderson, *Introduction to Neural Networks* (MIT Press, Cambridge, MA, 1995)
72. D.A. Perry, Self-Organizing Systems across Scales. Trends in Evolution and Ecology **10**, 241–244 (1995)
73. J., Pinto, The Advent of Self-Organizing Industrial Controls: The End of Centralized, Deterministic Control Systems. http://www.jimpinto.com/writings/selforg.html
74. J. Halloy et al., Social integration of robots into groups of cockroaches to control self-organizing choices. Science **318**(5853), 1155–1158 (2007)
75. An Introduction to Self-organization http://www.rpi.edu/~eglash/eglash.dir/selforg%20intro.htm
76. Cellular Automata. http://llk.media.mit.edu/projects/emergence/contents.html
77. History of self-organization in science and society. http://www.rpi.edu/~eglash.dir/selforg/selforg%20history.htm
78. B. Farley, W. Clark, Simulation of self-organizing systems by digital computer. Trans. IRE, Professional Group Inf. Theory **4**(4), 76–84 (1954)
79. H. Von Foerster, Jr., W. Zopf (eds.), 9 V, Principles of self-organization, *Information Systems Branch*, US Office of Naval Research, 1962
80. F. Heylinghen, Relational closure: a mathematical concept for distinction- making and complexity analysis, in *Cybernetics and Systems '90,* ed. by R. Trappl (World Science Publishers, pp 335–342, 1990)
81. M. Kawata, V. Toquenaga, Artificial individuals and global patterns. Trends Ecol. Evol. **9**, 417–421 (1994)

# Chapter 10
# Energy in Life and Society

*The history of energy demonstrates that human societies have not escaped its implacable laws. The more complex a society, the more it will need significant quantities of energy to maintain itself.*

Joël de Rosnay

*Energy is efficiently used when the quality of the source is matched to the quality demanded by the task.*

Ralph Torrie

**Abstract** This chapter is concerned with the use and impact of energy on life and society. All activities of life and society are energy-based and energy-handling processes. The energy for all life on Earth comes from the Sun. Living organisms consume the available high-quality energy and return lower quality energy as specified by thermodynamics. Nonliving entities also consume energy over time, but life processes are more efficient in consuming energy. The three dominant stages of energy domestication in human societies are the survival stage, the stage of increased energy depletion, and the present stage of more efficient use of Earth's energy resources (exhaustible and non-exhaustible). This chapter starts with a discussion of the three primary biochemical pathways, i.e., full series of energy-handling chemical reactions that take place in living organisms, namely, photosynthesis, respiration, and metabolism (catabolism, anabolism). Then, it examines the energy flow (food chains, food webs) in ecosystems including the efficiency of this flow. This chapter continues with a number of issues of the energy role in human society, namely the evolution of energy resources, the relation of energy with economy, the management of energy such that to achieve energy saving, the demand management which leads to "peak demand" minimization, and the use (consumption) of energy including relevant statistical data for the different parts of the Earth. The above issues and problems show the critical role of energy both for the life and the society, by providing the fuel needed for their existence, activity, and sustainability.

**Keywords** Energy · Life · Society · Biochemical pathways · Calvin cycle Carbohydrates · Carboxylation · Reduction · Regeneration · ATP NADPH · Respiration · Fermentation · Primary/organismal metabolism Catabolism · Anabolism · Ecosystem · Abiotic/biotic ecosystem Autotrophs · Heterotrophs · Food chains · Energy flow efficiency



S. G. Tzafestas, *Energy, Information, Feedback, Adaptation, and Self-organization*, Intelligent Systems, Control and Automation: Science and Engineering 90, https://doi.org/10.1007/978-3-319-66999-1_10

Energy/population pyramid · Renewable/nonrenewable energy resources
Thermoeconomics · Energy-matter-econonomic cycle · Energy management
Consumption of energy

## 10.1 Introduction

Life and society's activities are first of all energy-based and energy-handling processes. The principal characteristic that distinguishes the living from the nonliving is the fact that life needs energy to maintain itself. Of course, the same principles of physics and chemistry govern the energy-handling process of both living and nonliving objects. Heat and molecules flow by necessity from regions of high concentration to regions of low concentration (heat dissipation, diffusion). All of the energy for all life comes from the Sun. In nonliving entities, the energy handling process involves passive erosion, corrosion, and general dissolution. Living organisms consume the available high-quality energy (exergy) and return lower quality energy as specified by thermodynamics. Actually, the life processes are more efficient in consuming exergy than nonliving ones.

The flow of energy via life (i.e., the food chain) is an *in-out* process, the first step of which is always the process of photosynthesis. In photosynthesis, the Sun's radiant energy is converted into *carbohydrate* (sugar) molecules that are used by all living beings as a fuel for life, motion, and reproduction.

The work force of living organisms is provided by "proteins", the precise shape of which determines their features. *Metabolism* is performed through the antagonistic processes of *anabolism* (building up) and *catabolism* (breaking down), which in aerobic animals depends on oxygen use. The metabolic rate of animals is higher when they are youngest and slows down gradually with increasing age. Zero metabolic rate means the organism is *not alive* (i.e., dead).

The three broad stages of the domestication of energy in human societies are as follows: the *survival stage* (a long period of human survival via the direct use of Earth's energy income); the stage of *increased energy depletion* (a period started about 150–160 years ago); and the present stage of *more efficient use* of Earth's energy resources, exploiting systematically solar and wind energy [1, 2].

This chapter discusses the fundamental issues of life and human society. Section 10.2 examines the three most important *biochemical pathways* (photosynthesis, respiration, and metabolism). Section 10.3 is concerned with the energy flow in ecosystems (*food chains*), including the efficiency of this flow. Sections 10.4–10.7 examine a number of issues of the energy role in the human society, namely, the evolution of energy resources, the relation of energy with economy (*thermoeconomy*), the management (saving) of energy, the demand management, and the consumption of available energy, including relevant statistical facts.

## 10.2 Energy and Life: Biochemical Pathways

A *biochemical pathway* is defined as the full series of energy-handling chemical reactions taking place in living organisms. The three most important biochemical pathways, which will be briefly reviewed here, are [3–16] as follows:

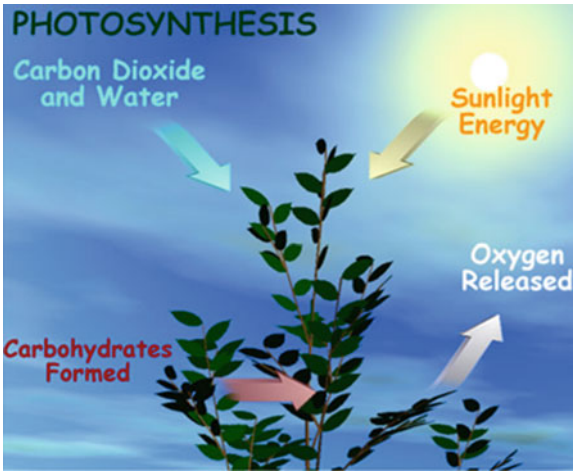



**Fig. 10.1** Illustration of the photosynthesis process (http://techalive.mtu.edu/meec/module19/images/Photosynthesis.jpg (The reader is informed that web figures and references were collected at the time of writing the book. Since then some of them may no longer be valid due to change or removal by their creators))

- Photosynthesis
- Respiration
- Metabolism

### *10.2.1 Photosynthesis*

Photosynthesis is the process by which organisms involving the *pigment chlorophyll* use sunlight energy to convert *water* ($H_2O$) and *carbon dioxide* ($CO_2$) into *glucose* ($C_6H_{12}O_6$) and *oxygen* ($O_2$) (Fig. 10.1). Actually, the light energy is transformed into the chemical energy stored in the molecular bonds of glycose. Glycose is the basic fuel and basic building material for almost all of life. The process of photosynthesis is described by the following chemical equation:

$$6CO_2 + 6\,H_2O + \text{Light Energy} \rightarrow C_6H_{12}O_6 + 6O_2$$

The released oxygen ($O_2$) is the by-product of this reaction. Glucose, like all other sugars, is converted into *starch*, *protein*, *cellulose*, *fats*, and the other chemical compounds contained in the body of living organisms, as will explained later.

The photosynthesis process can be broken up into two principal classes of reactions, namely:

- **Light reactions**: Light energy excites photosynthetic pigments to higher energy levels, which results in the production of two high-energy compounds, namely, *adenosine triphosphate* (**ATP**) and *nicotinamide adenine dinucleotide phosphate* (**NADPH**). These two compounds do not appear in the photosynthesis equation because they are consumed during the subsequent dark reactions in the production of glucose.
- **Dark reactions**: These reactions involve a series of enzymatic reactions that make *carbohydrates* (sugars) from $CO_2$. The dark reactions do not employ light directly, but they use ATP and NADPH, which are the outcome of the light

reactions. Therefore, the dark reactions depend indirectly on light and often take place in the light. The principal part of the dark reactions is called the *Calvin cycle*, from *Melvin Calvin* who discovered them. The Calvin cycle takes $CO_2$ and converts it into organic molecules (starch). It consists of 13 different biochemical reactions, each catalyzed by a separate enzyme, which can be summarized in the following three steps:

- Carboxylation,
- Reduction, and
- Regeneration.

In *carboxylation*, a molecule of $CO_2$ is combined with a molecule of **RuBP** (*ribulose bisphosphate*), the five-carbon sugar phosphate, to make two molecules of *phosphoglyceric acid* (*PGA*), a three-carbon organic acid. The enzyme that catalyzes the carboxylation reaction (i.e., the conversion of $CO_2$ to PGA) is **RuBISCO** (*Ribulose bisphosphate carboxylase/oxygenase*). RUBISCO is the most abundant enzyme on Earth.

In *reduction*, ATP and NADPH (synthesized during the light reactions) provide the energy required for synthesizing high-energy carbohydrates from the PGA produced during carboxylation. In *regeneration*, the carbohydrates produced at the reduction stage are subject to a sequence of enzymatic reactions which regenerate RuBP, used initially in the carboxylation. Thus, the carboxylation, reduction, and regeneration reactions start and end with the same RuBP, and so they form a cycle (the *Calvin cycle*). After the completion of six Calvin cycles, six molecules of $CO_2$ are used, and a molecule of glucose ($C_6H_{12}O_6$) is produced (Fig. 10.2).

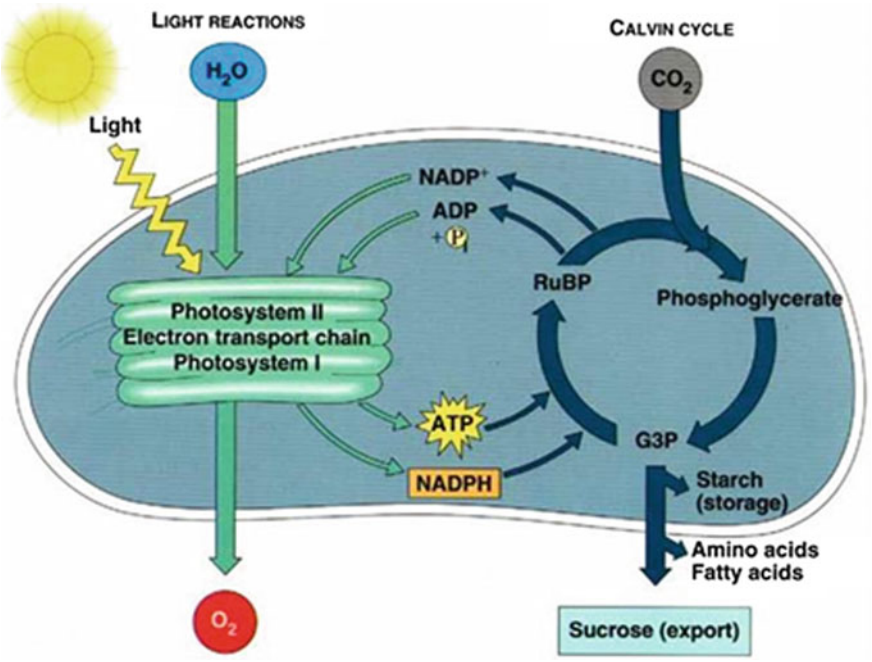



**Fig. 10.2** Illustration of the Calvin cycle (http://mandevillehigh.stpsb.org/teachersites/laura_decker/calvin_cycle_files/image002.jpg)

The photosynthesis process takes place in the *green leaves* of plants, *marine algae*, and some types of bacteria called *cyanobacteria* (where «*cyano*» means blue from the Greek "*κυανό*" = blue). The green color of the leaves is due to the chlorophyll pigments in the *chloroplasts* (little bacteria-sized "organelles", which are inside the leaf). Leaves appear to be green, because this color of light (frequency spectrum) is reflected (i.e., not used), whereas the red and blue colors are absorbed and used providing the greater part of energy for the process of photosynthesis.

The wavelengths absorbed by chlorophyll lie between 400 and 700 nm (1 nm (nanometer) = $10^{-9}$ m). The oxygen released by photosynthesis is the oxygen contained in the water $H_2O$. This was proven by *Melvin Calvin* via an isotopic tracer experiment where the heavy oxygen isotope was used, i.e., $H_2{}^{18}O + CO_2$ yields ${}^{18}O_2$, and $H_2O + C\ {}^{18}O_2$ yields $O_2$. Similarly, the dark reactions sequence was determined using the carbon radioactive isotope ${}^{14}C$. From a thermodynamic point of view, photosynthesis is not a very efficient process, since only 1% of the sunlight energy reaching the surface of a leaf is used in photosynthesis. Of the remaining (99%), 4% is converted to heat energy, 5% is transmitted through the leaf, 15% is reflected, and 75% is evaporated.

Actually, photosynthesis takes place inside the *chloroplast*, specifically in the *granum* (Fig. 10.4). Cutting open a plant leaf, we see that it consists of the following principal parts (more parts are shown in Fig. 10.3):

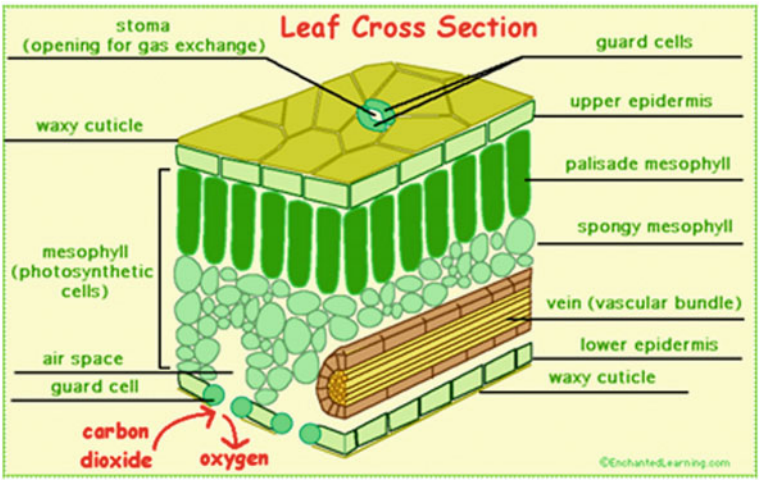



**Fig. 10.3** Sketch of a leaf's cross section (http://image.tutorvista.com/content/feed/u2044/leafcross%20section.GIF)

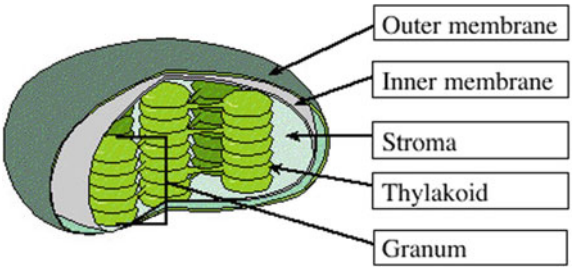



**Fig. 10.4** Structure of the chloroplast (http://schools.look4.net.nz/science/biology/plant/photosynthesis/Chloroplasts.jpg)

- *Epidermis*: This dermal tissue protects the leaf, gives shape to it, and keeps sufficient water inside it for the photosynthesis.
- *Cuticle*: A hard, waxy, watertight material of varying thickness (according to climate conditions), which allows suitable beading of water inside the leaf.
- *Mesophyll cells*: These cells are located between the upper and lower epidermis. Mesophyll includes the *vascular tissue* and the *ground tissue*. In Fig. 10.3, the mesophyll is divided into the upper section, called *palisade parenchyma*, and the lower section, called *spongy parenchyma.* The mesophyll cells contain one to many chloroplasts (up to 50 or more according to the species, age, and health of each plant). The palisade parenchyma cells contain three to four times the number of chloroplast contained in the spongy parenchyma cells.
- *Stoma*: Stoma (from the Greek στόμα (stoma) = opening) is the means via which air brings $CO_2$, needed for the photosynthesis into the mesophyll cells, and $O_2$ (the by-product of photosynthesis) flows to the atmosphere. The $CO_2$ produced by the plant respiration (see Sect. 10.1) is also given off by the stomata (plural of stoma) or stomas. The combination of the leaf's protection cells (cuticle and epidermis cells) is called "stomate".

In the ocean, photosynthesis is performed by *phytoplankton* (or *phytoplankton*), which is made from very small living cells near the water surface. These cells use the sunlight energy to make energy-providing molecules (ATP) and build material for themselves and all other *marine life.* Another group of bacteria is called *chloroxybacteria* which are prokaryotes. This is represented by the single gene *Prochloron*, which contains chlorophyll and carotenoids as photosynthetic pigments. Prochloron acts very much like the chloroplasts of higher (eukaryotic) plants. It is estimated that green plants, algae, and cyanobacteria/chloroxybacteria (in toto, called together *primary producers*) produce every year about 100 billion tons of sugar for life on Earth.

### *10.2.2 Respiration*

Chemically, *respiration* is the exact opposite of photosynthesis and occurs in all living cells (not only in the cells that involve chlorophyll). By respiration, food (organic) molecules (e.g., sugars) are oxidized (and broken down) to derive ATP from the molecular bonds that are broken. Energy is released when **ATP** is converted to **ADP** (*Adenosine diphosphate*).

The equation for respiration is:

$$C_6H_{12}O_6 + 6O_2 \rightarrow 6CO_2 + 6H_2O + 36ATP$$

We see that the products of this reaction are carbon dioxide and water, and the energy is released in the form of ATP. The theoretical maximum yield of cellular respiration is 36 ATP per molecule of glycose.

It is noted that plants during the night also consume oxygen for their respiration like animals, but, overall, every 24 h they release much more oxygen via photosynthesis than the oxygen they consume. Structurally, ATP consists of the *adenine*

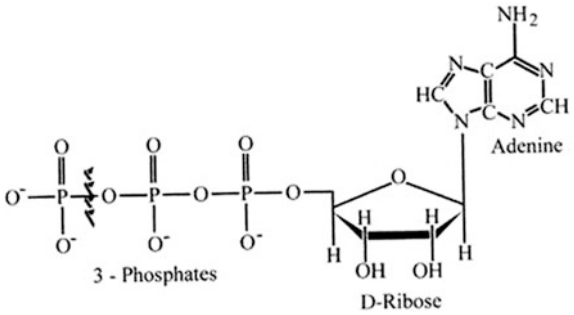



**Fig. 10.5** The ATP molecule

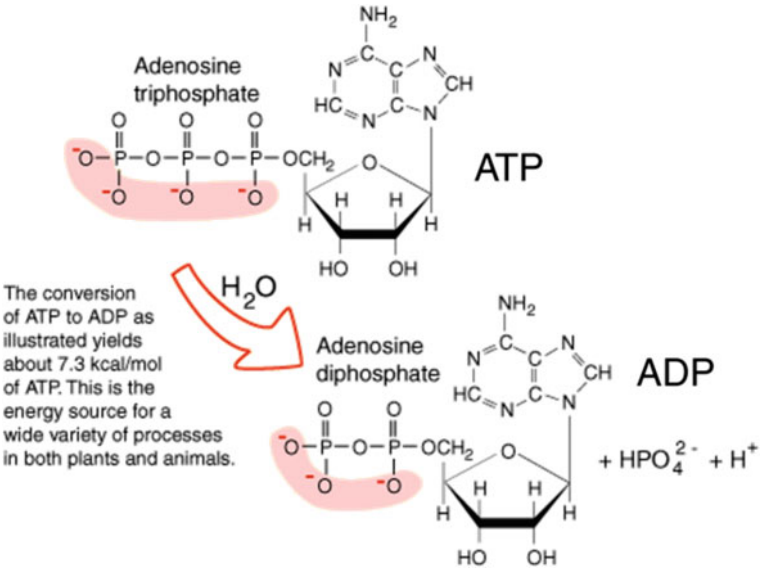



**Fig. 10.6** The ATP to ADP conversion via hydrolysis (http://hyperphysics.phy-astr.gsu.edu/hbase/biology/imgbio/atptoadp.gif)

*nucleotide base* (ribose sugar, adenine base, and the phosphate group $PO_4^{-2}$), plus two other *phosphate groups;* its chemical formula is shown in Fig. 10.5.

Energy is stored in the *covalent bonds* between phosphates, with the larger quantity (about 7 kcal/mole) in the pyrophosphate bond (the bond between the second and third phosphate groups). **ADP** is the product of *ATP* dephosphorylation via ATPases (hydrolysis) and has one phosphate group less than ATP, i.e., its chemical formula is as shown in Fig. 10.6.

The process (hydrolysis) is the following:

$$\mathrm{ATP} + \mathrm{H_2O} \rightarrow \mathrm{ADP} + \mathrm{H_3PO_4} + \text{energy}\ (30.5\,\mathrm{kJ/mol})$$

One more dephosphorylation of ADP yields *adenosine monophosphate* (**AMP**). Finally, dephosphorylation of AMP results in *adenosine.*

Cellular respiration starts with the products of glycolysis being transported into the mitochondria. Glycolysis splits six-carbon glucose into three-carbon *pyruvic acid* molecules and two ATP molecules. This process does not require the presence of molecular oxygen, i.e., it is *anaerobic*. The pyruvic acid is broken down in the *fermentation* process without using molecular oxygen. Then, after this point, there are three alternative possibilities:

- **Aerobic respiration**: The Krebs (or Citric Acid) cycle, which from each glucose molecule produces two ATP molecules, ten carrier molecules, and $CO_2$, followed by the *electron-transport chain*, which produces 34 ATP molecules and $H_2O$ from the carrier molecules (in all, 36 ATP molecules).
- **Lactic acid fermentation**: This takes place in animal cells because of the lack of oxygen, as lactic acid is produced and causes muscle soreness. Although no ATP is produced, the glycolysis continues with the help of a carrier compound generated.
- **Alcoholic fermentation**: This takes place in some plants and one-cell organisms. In this process, the pyruvic acid is transformed into ethyl alcohol while a carrier compound helps the glycolysis to continue.

As we have seen, the *aerobic respiration* (with oxygen) gives 36 ATP molecules from each glucose molecule, while the *anaerobic respiration* (without oxygen) produces two ATP molecules per glucose molecule. Thus, the efficiency of aerobic

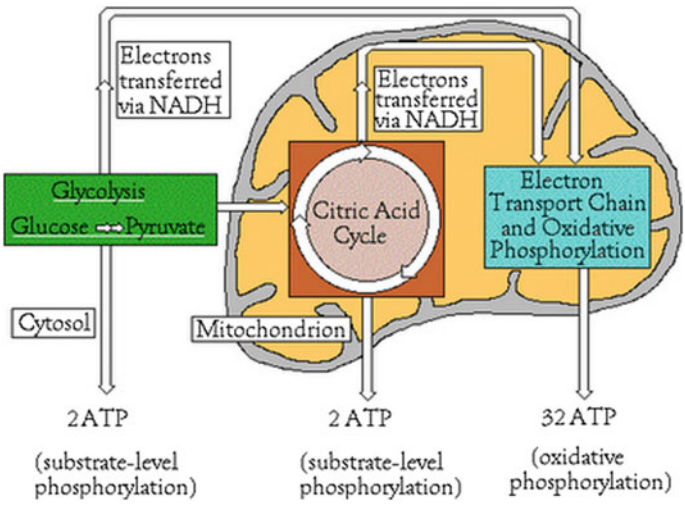



**Fig. 10.7** Pictorial overview of the cellular respiration processes (http://elinow-bioreview2.wikispaces.com/file/view/cellular_respiration.jpg/191370366/cellular_respiration.jpg)

respiration is 18 times higher than that of anaerobic respiration. A pictorial representation of the cellular respiration processes is shown in Fig. 10.7. In prokaryotes, these processes occur in the cytoplasm, while in eukaryotes they take place in the mitochondria.

The human body contains around 0.1 mol of ATP, while the energy used in the body's cell requires the energy contained in about 200–300 mol of ATP each day. Therefore, each ATP molecule is recycled 2000–3000 times every day. The amount of ATP produced, processed, and recycled in the human body every hour is about 1 kg.

### *10.2.3 Metabolism*

*Metabolism*, from the Greek "μεταβάλλω" (metavallo = change) involves the chemical processes via which living cells generate the substances and energy required to maintain life. Actually, metabolism is the ingestion and breakdown of complex compounds, followed by the liberation of energy and the production of waste materials.

Metabolism can be subdivided into these kinds:

**Total Metabolism**: This is the metabolism that concerns the whole organism, i.e., all metabolic chemical processes of the living organism, including transport of compounds between cells whenever it is needed.
**Specific metabolism**: This is the metabolism that concerns a particular compound in the living organism.
**Cell metabolism**: This is the metabolism that refers to the chemical processes that take place within a particular living cell.

Metabolic processes do not include single processes or functions (e.g., protein–protein interactions, protein–nucleic acids interactions, or receptor–ligand interactions)

Metabolism can also be categorized into two principal genres:

**Primary metabolism**: This includes the metabolism (anabolism/catabolism) processes that occur in most (if not all) cells of the organism.
**Secondary metabolism**: This concerns the metabolism of compounds that are not primary compounds of the organism. Secondary (non-primary) *metabolite* is defined as a compound that does not directly function in the processes of growth and development. Examples of secondary metabolites are antibiotics and plant chemical defense compounds such, as alkaloids and steroids.

The metabolic processes of multicellular organisms that take place at the tissue, organ, or total organism level are globally called *organismal metabolism*.

Some examples of organismal metabolism are as follows: photosynthesis, salivary polysaccharide metabolism lignification, starch metabolism, and lipid metabolism.

Moreover, we have two other types of metabolism:

- Catabolism and
- Anabolism

*Catabolism* or destructive metabolism is the breaking down of more complex compounds in the organism into simpler ones, followed by the release of energy. For example, glycolysis is a catabolic process that breaks down carbohydrates (sugars) through a series of reactions into either pyruvic acid or lactic acid, and this releases energy in the body in the form of ATP, as described in Sect. 10.2.2.

*Anabolism* or constructive metabolism is the synthesis in living organisms of more complex substances (e.g., protein for growth and repair of living tissue) from simpler ones using energy released by catabolism.

Most metabolic processes are brought about by the action of *enzymes*. The speed at which the metabolic processes take place in a living organism is called its *metabolic rate*. The metabolic rate of an organism at rest is called its *basal metabolic rate*. The locomotion of birds (flight) has high-energy demands, and so birds have a very high metabolic rate.

While in plant and all organisms that have the pigment chlorophyll, ATP is formed from ADP by trapping solar energy via photosynthesis, in animals, ATP is formed from ADP and inorganic phosphate (with the aid of the ATP synthase enzyme) by using energy produced by catabolism. This ATP is then used to fuel the organism's growth and cell maintenance [9, 15, 16].

## 10.3 Energy Movement in an Ecosystem

### *10.3.1 General Issues*

An *Ecosystem* is the biological community that exists in some geographic area on Earth and embraces the physical and chemical issues affecting its nonliving (abiotic) environment. Ecosystems are the minimal units of ecology. The term *ecology* comes from the Greek word "*οικολογία*" where "*οίκος*" (ecos) = house and "*λόγος*" (logos) = speech. Here "*οίκος*" is the *Earth* (nature) and "*λόγος*" has the meaning of *study*. Actually, ecology is the study (science) of the interaction of living and nonliving entities in nature [17–21].

Examples of ecosystems are a lake, a forest, an estuary, a grassland, etc. In most cases, the boundaries of ecosystems cannot be specified in a clear way, but, for practical scientific reasons, the boundaries are usually specified according to the goals of each study.

The two principal processes that are the subject matter of ecosystems are as follows:

- Energy movement
- Biochemical cycling

In the field of *ecosystem ecology*, the study of both individual organisms and populations is included. Regarding individuals, the aspects of physiology, reproduction, and development are examined, and, for the populations, the aspects of group behavior, population growth, and abundance/extinction are considered.

An ecosystem consists of two kinds of components, namely:

**Abiotic**: These components include sunlight, temperature, water/moisture, precipitation, soil or water, chemistry, and nutrients (organic and inorganic, poisons, etc.).

**Biotic**: These components include primary producers, herbivores, carnivores, omnivores, detritivores, etc.

As we have seen in Sect. 10.2, energy enters the ecosystem as light energy (Sun energy) and is transformed into chemical energy in organic molecules with the action of biochemical pathways (photosynthesis, respiration, and metabolism). This energy is finally converted to heat energy and dissipated in the ecosystem, being unable to be recycled. To live, the ecosystem needs the continuous solar energy input; otherwise, the living organisms would quickly cease to live.

This means that the *Earth is an open thermodynamic system.* But, with regard to the material moving on it, the Earth is a *closed system*, since all chemical elements such as carbon, phosphorus, and minerals entering living systems by any means possible (atmosphere, water, soils, and consumption of other organisms) are not destroyed or lost in ecosystems. The human body is made mostly (about 98%) of six main types of elements: carbon 19.4%, hydrogen 9.3%, nitrogen 5.1%, oxygen 62.8%, and phosphorus and sulfur each about 1%. The chemical elements are *cycled perpetually* between their biotic and abiotic states. Of course, an exception to the material normally enclosed within the Earth system is the case of a meteorite that entered in the past or possibly will arrive in the future.

### *10.3.2 Energy Flow Through Food Chains*

Energy on land flows from plant to animal to animal in bite-sized chunks. In the ocean (marine life), energy moves from phytoplankton to zooplankton (e.g., shrimp and small squid)) to little fish to bigger fish, and so on. Phytoplankton do for ocean animals through photosynthesis what green plants on land do for land animals. They provide the food for all the other marine creatures. Such a path of food consumption is called a **food chain**.

Organisms that consume other organisms are classified in several "**vores**". "**Vore**" means "**eater**". Specifically, we have:

- **Herbivores**: That is, organisms that, to live and act, eat only plants (herbi = plant) or plant-like materials.
- **Carnivores**: That is, organisms that eat only meat (carni = muscle), i.e., animals.

- **Omnivores**: Organisms (as do humans) that eat both plant and animal material (omni = everything) on a regular basis.
- **Detritivores and Decomposers**: Organisms that eat all forms of waste material and dead, once living organisms (detritus = garbage/undissolved matter resulting from the decomposition of parent material).

A different, more complete, and popular classification of organisms on Earth is the following:

**Autotrophs**: Organisms that produce their own food (e.g., carbohydrates). Plants, algae, and photosynthetic bacteria all belong to the class of autotrophs. (The term *autotroph* comes from the Greek "αυτός" (self) and "τροφή" (trophi = food).

**Heterotrophs**: Organisms that get their energy by consuming either organisms or the nonessential parts of other organisms. In general, all non-autotroph organisms are heterotrophs and eat autotrophs or other heterotrophs, or live off their remains (The word "*hetero*" comes from the Greek "έτερος" which means "*somebody else*", not yourself).

The autotrophs of an ecosystem are called *producers.* It is noted that producers do not produce energy, but transport energy into stored carbohydrates. This stored energy is all non-heat energy available to the heterotrophic part of an ecosystem.

The organisms that consume other living organisms to get their energy (i.e., the heterotrophs) are called *consumers*. The material being produced or consumed is carbohydrate (sugar and related compounds)

Consumers are distinguished in the following *trophic levels* (from the lowest to the highest level):

- **Primary consumers or herbivores**: Consumers that eat primary producers. A fish that browses on algae and a cow that browses on grass are primary consumers.
- **Secondary consumers or carnivores**: Consumers that eat primary consumers. A fish that browses on fish that browse on algae is a secondary consumer.
- **Tertiary consumers**: These are consumers that, at least in part, eat secondary consumers. A bird that browses on fish that browse on fish that browses on algae is a tertiary consumer.

It is noted that "a single organism" may easily be functioning on several "trophic levels". Almost all carnivores can function at any trophic level, above the primary consumer. For example, a lion that functions as a secondary consumer could also very easily be a tertiary consumer.

This is a matter of what the lion is eating right now. A human may function one meal at several trophic levels. He/she is a primary consumer for the green salad, secondary consumer for the lamb chop, tertiary for a salmon steak, and so on.

Besides the green (autotrophic) plants that produce all of the energy required for their living, via photosynthesis, there are some plants that cannot do this totally from water, carbon dioxide, and a few other nutrients (like nitrogen). These plants have some green parts, but they need to take some nutrients from autotrophic *host* plants. These plants are called *semi-parasitic plants.* Besides these plants, there also

exist *holoparasitic plants* ("*holon*" comes from the Greek "*όλον*" = *total*) that have no green parts at all and cannot perform photosynthesis. They have to take all of their nutrients from host plants. Other plants that do not use $CO_2$ are the *saprophytes*, which have a fascinating relationship with the trees providing them their *food:* Fungus which is called a *mycorrhizal* fungus connects the myco-heterophytes to the tree and transfers the nutrients from the hosting plant to the parasitic plant. Finally, there are some plants that they do photosynthesis using $CO_2$, but they need some more, which they collect by catching and slowing digesting insects and other little animals that fall into their *fly traps*. These plants are called *insectivorous* plants. Examples of such plants are the cool and creepy "venus fly trap" and the *pitcher plants*.

The energy flow through food chains is not efficient because less energy (exergy) is available at the primary consumer (herbivore) level than at the producer level, still less at the secondary consumer (carnivore) level, and so on. A short discussion of the *efficiency* of energy flow through food chains follows [22–25].

### *10.3.3 Efficiency of Energy Flow Through Food Chains*

A *food chain* is a single pathway of energy and nutrients' flow, and, as we have already indicated, each level of consumption in a food chain is called a *trophic level*. Most food chains are interconnected. These interconnections create the so-called *food webs*. Food chains and food webs ensure the equilibrium of the ecosystem, i.e., the balance and harmony of the Earth's flow of nutrients and energy. Food chains are simplistic representatives of what is actually occurring in an ecosystem. A food chain shows only one pathway of energy and material flow. The majority of consumers feed on multiple species and, in turn, are fed upon by many other species.

Two food web examples are shown in Fig. 10.8: a simple land-food web (a) and the Antarctic food web (b), which is also very simple (simpler than many food webs in other ecosystems of the Earth). In particular, the food chain "phytoplankton → krill → baleen whale" is very short, and the corresponding energy pyramid ensures that as much as possible of the phytoplankton's energy is transferred with very high efficiency. The size of whales is extremely large (more than a hundred tons), and their density is very close to that of the salt water. This allows the buoyancy of the water to support their great bulk.

As we have seen, in a web there are sequential interconnections from one trophic level to the next and several other interconnections that may involve the case of bigger animals (like a whale) eating very small organisms (like plankton).

At each step, from one trophic level to the next, there occurs an energy loss, and so the higher trophic levels have fewer individuals to keep the ecosystem in balance. Figure 10.9 shows the pyramid of population numbers at each trophic level in an *acre of bluegrass* [18].

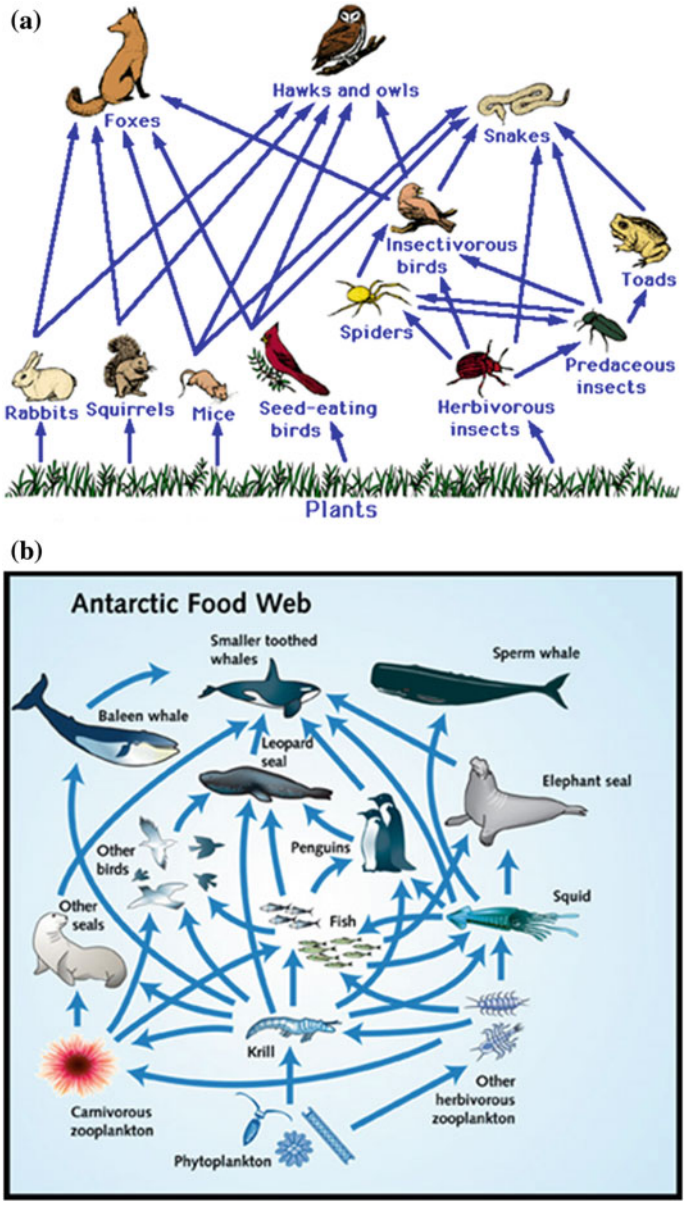



**Fig. 10.8** **a** A simple land food web, and **b** the Antarctic food web (**a** http://www.envirotrain.co.uk/wp%2Dcontent/uploads/2010/08/1.3%2D4.5%2Dland%2Dfoodweb.gif; **b** http://science.jrank.org/kids/article_images/chains_p37.jpg)

The pyramid occurs because the total biomass of each species is limited by its trophic level. This implies that, if the size of the individuals at a given trophic level

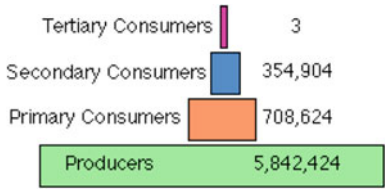



**Fig. 10.9** The pyramid of population numbers in the ecosystem of an acre of bluegrass

is large, their number must be small and vice versa. The size of prey is smaller than that of their predators. The total biomass of predators is smaller since they belong to a higher trophic level. Thus, the number of individuals in the prey population is much higher than that in the predator population. According to Chapin [17], this is a two-way control: *bottom-up* control and *top-down* control.

The *ecological efficiency* is distinguished in [17–19]:

- Consumption efficiency
- Assimilation efficiency
- Production efficiency

The biomass *consumption efficiency* of a trophic level must be lower than that of underlying level. Otherwise, the underlying level would be subject to extinction

The *assimilation efficiency* is specified by the food quality and the physiology of the consumer. Non-assimilated food is excreted as feces. The assimilation efficiency of carnivores is higher (about 80%) than that of herbivores on land (about 5–20%)

The *production efficiency* refers to the amount of energy used for animal production (reproduction and growth) and depends principally on the metabolism of the animal. Besides the population number's pyramid of Fig. 10.6, we also have the *energy pyramid* and the *biomass pyramid*.

**Energy Pyramid**: At each link in a food chain, a large part of the Sun's energy (initially used by a photosynthesizing autotroph) is dissipated as heat to the environment. Thus, the energy conversion efficiency is always less than 100%. This means that the total amount of energy stored in the bodies of a given population depends on the trophic level to which it belongs. An *energy pyramid* example is shown in Fig. 10.10.

**Biomass Pyramid**: The *biomass* B of a given population is defined as

$$B = N \times \bar{W}$$

where N is the number of individuals in the population, and $\bar{W}$ is the average weight of an individual. This definition is based on the fact that all organisms in a population are made of nearly the same organic molecules in similar proportions. The *biomass* (or, as otherwise called, the *standing crop*) is reduced from one level to the next level in the food chains. The biomass pyramid for the river ecosystem of Fig. 10.10 is shown in Fig. 10.11, where the figures represent the dry weight of organic matter per square meter [18].

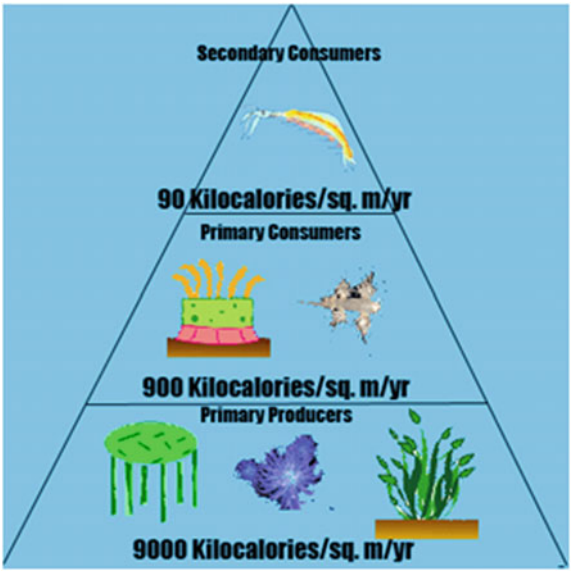



**Fig. 10.10** Energy pyramid in the flow of energy through a water ecosystem (http://www.world-builders.org/worlds/planets_04/spirit/sdimages/waterpyramid.gif)

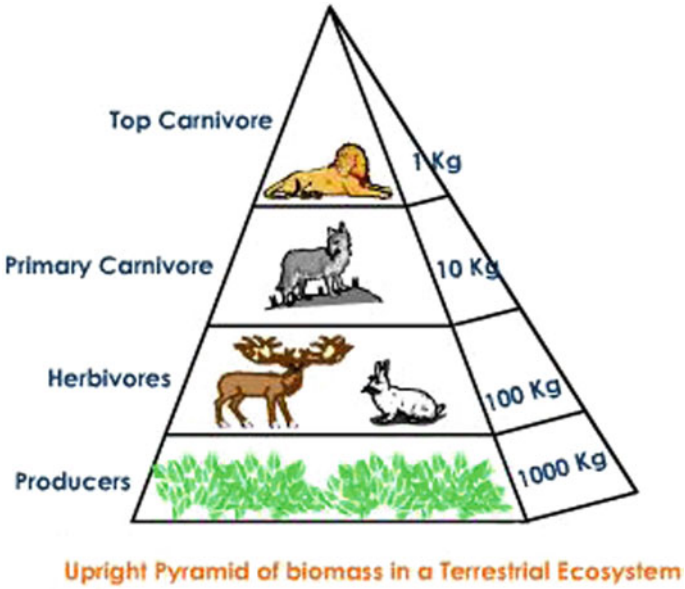



**Fig. 10.11** An example of a biomass pyramid (http://images.tutorvista.com/content/ecosystem/biomass%2Dupright%2Dpyramid.jpeg)

Several ecosystem experiments have shown that the *energy conversion efficiency* $\eta_{ce}$ varies from about 1–85% in various food webs [17–24]. The conversion efficiency is defined as

$\eta_{ce}$ = Net level production at one trophic level/Net production at the next higher level

The net production is only a fraction of gross production, because the organisms have to spend energy to keep alive. The difference between gross and net production is greater for consumers than for the producers. In general, a good working figure for the average conversion efficiency from producers to primary consumers is 10%. For example, in the river ecosystem at Silver Spring, the *conversion efficiency* was measured to be [18]

- 1.7% from producers to primary consumers
- 4.5% from primary to secondary consumers

Using the 10% figure for the conversion efficiency, we see that tertiary consumers contain about $10\% \times 10\% \times 10\% = 0.1\%$ of the net production of the producers.

## 10.4 Energy and Human Society

### *10.4.1 General Issues*

Energy plays a dominant and critical role in human society. No activity of any form can be performed without energy flow or conversion. Energy use in human society starts with a source (exhaustive or non-exhaustive) and is subject to many intermediate process stages (conversion, refinement, etc.) before arriving at an individual entity: a public utility, a public or private institution, a home, a land, air, or sea vehicle, or a service provider, where it is exploited by prime movers, heating/cooling devices, and other equipment. This is why, in the context of society, the term energy is used to mean energy resources. As we have seen in Chap. 3, although the total amount of energy is conserved, every time energy is used or converted into another form, its quality and availability decreases and becomes less useful to human society. As a consequence, after these processes, people in the society think that energy has been used up. The study of energy's role in human social activity and development was by the necessity of primary concern over the centuries of human existence. The progress steps, the transformation, the utilization, and, in general, the role and impact of energy upon society and the environment are extensively described and discussed in the literature of the field, including a variety of websites and other documents. At this point, it is useful to mention a few books [25–30]. In [25], *Vaclav Smith* introduces the term *general energetics* to cover all types of energy and energy flows, i.e.,

- *Planetary energetics* (e.g., solar radiation, geomorphic processes),
- *Bioenergetics* (e.g., photosynthesis),
- *Human energetics* (e.g., metabolism, muscle power, and thermoregulation), and
- *Socioenergetics* (e.g., socioeconomic issues of energy use, energy role in quality of life, etc.).

In [27], an introduction to work, energy, and efficiency is provided, including the study of home-comfort energy devices, heat transfer devices, cars, wind turbines, nuclear energy plants, etc. In [28], the potential limits of technological and social issues are investigated via global case studies, and social, economic, and political solutions for avoiding serious environmental damage in the future, are suggested. In [29], the relations of energy, social change, and economic growth are thoroughly studied. Last, but not least, in [30], the historical path of exhaustible and renewable energy consumption in American society is presented.

### *10.4.2 Evolution of Energy Resources*

Before discussing particular aspects of energy in society, it is useful to have a brief look at the evolution of energy resources used by human societies over time. The development of energy throughout the world's history has passed through two quite asymmetrical periods:

- Renewable (non-exhaustible) energy period and
- Nonrenewable (exhaustible) energy period

These types of energy were presented in Sect. 2.5.
According to [31, 32] the evolution of energy types involves the following:

- Biomass-based energy
- Human and animal muscle energy
- Preindustrial inanimate prime movers
- Fossils fuels, mechanical prime movers, and electricity, and
- Modern energy systems.

*Biomass fuels* have low *power density* ($W/m^2$) and low-*energy density* (J/Kg). Even the richest forest biomass (thick tree trunks) has a power density not exceeding 1 $W/m^2$, whereas smaller trees and leaves have much less power density (Fig. 10.12).

During the wood-biomass period, it was not possible to have megacities ($\sim$10 million people) because this would have required a nearby wooded area of up to 300 times the city size to supply its fuel. The wood era was followed by the *charcoal* period. Charcoal has about 60% higher energy density than wood fuel (18–24 MJ/Kg) (see Fig. 10.13).

Although the charcoal has the benefits of smaller mass to be transported and stored, its very high inefficiency (about 80% of the used wood was wasted for the

**Fig. 10.12** Typical wood biomass (http://www.power-technology.com/projects/western-wood-energy/images/1-logs.jpg)

**Fig. 10.13** Standard charcoal made by burning of suitable wood (http://usercontent2.hubimg.com/7336804_f496.jpg, http://charcoalkiln.com/wp%2Dcontent/uploads/sites/5/2013/03/Coconut%2Dshells%2Dcharcoal.jpg)

charcoaling process itself) was the main reason for the end of the use of coal-derived coke.

*Human and animal muscles* have limited power. Humans can provide 70–100 W, light horses can provide about 500 W, heavier horses about 800–900 W (1 horsepower = 745 W), and even heavier animals can develop, for short times, up to 3 kW, being able to do tasks impossible for men. These increased capabilities obtained by big animals were counterbalanced by the need to feed and take care of them. Unavoidably, the human and animal muscle power was mainly employed in agriculture. Also, the speed of travel and transportation achieved by humans and animals (animate metabolism) was very much limited, and so speedy running and horse riding were used only in the most urgent cases.

The first *preindustrial inanimate prime movers* were simple mechanical devices for the conversion of flowing water and wind (both Sun-based, inexhaustible energy resources) to rotary power and ship propulsion power. Then the *water wheel* was

developed (Fig. 10.14), which was used by medieval societies principally for food processing, wood sawing, and metallurgical processing. In the early eighteenth century, the large water wheels in Europe reached about 4 kW.

The *windmill*, which has existed since the tenth century, has been in extended use in the Mediterranean, Middle East, and some coastal areas of Atlantic Europe. The average power of the more advanced windmills was 5 kW (Fig. 10.15). Their power was added to human/animal muscle power, but it was not sufficient to secure the supply of food and other material comforts most of the populations.

*Fossils fuels* have changed these conditions of human life. Coal has been used in Europe and Asia to a limited extent for centuries, but the large-scale transition from primitive (wood) biomass to mined coal occurred in Europe and America during the

**Fig. 10.14** A water-wheel system (http://www.clker.com/cliparts/d/b/1/b/13275779421548061843rossett%20water%20wheel300.jpg)

**Fig. 10.15** A typical Dutch windmill (http://2.bp.blogspot.com/%2D2j6LdvVdYWs/TjJ01WBEvpI/AAAAAAAAACk/V7jnpUo0cME/s320/windmill2.jpg)

nineteenth century, and in the large Asian countries during the second half of the twentieth century. Coal mines (Fig. 10.16) with multiple seams and rich oil and gas reservoirs can provide power densities in the range 1000–10,000 W/m$^2$, i.e., power densities 10,000–100,000 higher than those of wood biomass.

Unfortunately, high-energy-density fuels (oil and gas) are drawn from limited stocks of exhaustible deposits. Their distribution has passed from the regional level to the international level via transportation in huge tankers.

The time evolution of *prime movers* has the sequence: *steam engines*, *internal combustion engines*, *steam turbines*, and *gas turbines*. These engines have secured higher conversion rates and higher overall production capacities. *James Watt* increased the performance of the steam engine from 0.1% in the 1700s to 5% in the 1800s, with an average power 20 kW (about 24 horses). By the end of the nineteenth century, the larger steam engines were providing the power of 400 horses with 10% efficiency (Fig. 10.17).

**Fig. 10.16** A mine entrance (https://otahkeyah.files.wordpress.com/2012/08/img_0320.jpg)

**Fig. 10.17** A steam engine-powered train (http://farm8.staticflickr.com/7134/7839345256_3b9a142b2a_z.jpg)

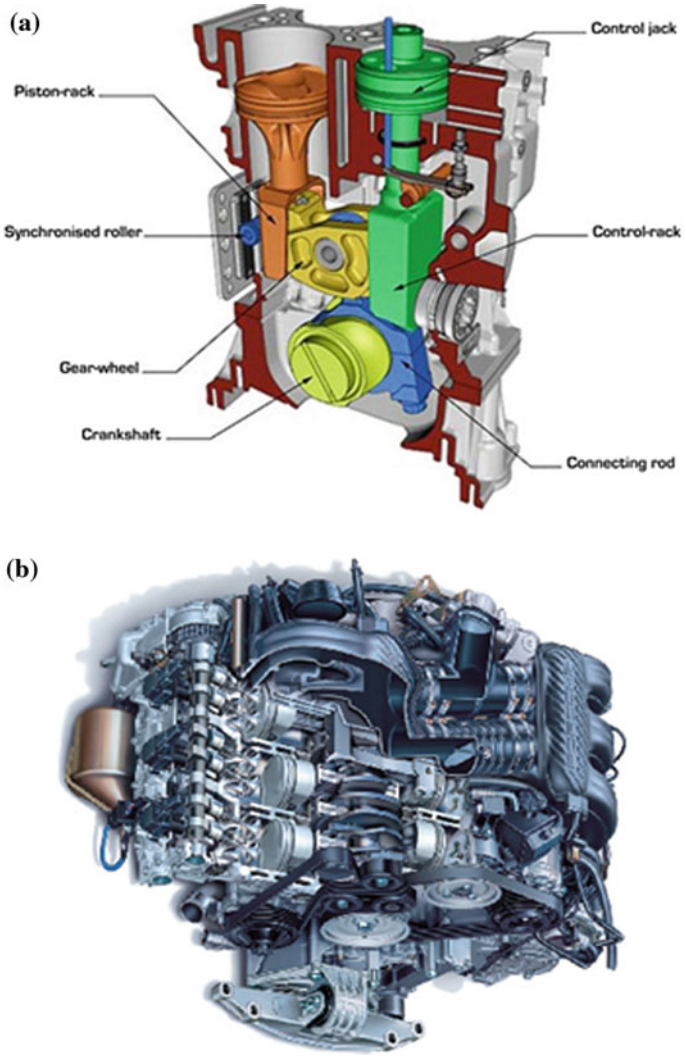



**Fig. 10.18** **a** Cutaway diagram of an internal combustion engine (http://image.motortrend.com/f/technology/mce%2D5%2Dto%2Ddebut%2D220%2Dhp%2D15l%2Dengine%2Dwith%2Dvariable%2Dcompression%2Dratio%2Dat%2Dgeneva/14885447+cr1+re0+ar1/mce%2D5%2Dvcri%2Dengine% 2Dcutaway%2Ddiagram.jpg). **b** Illustration of an advanced, modern combustion engine (http://www.seriouswheels.com/pics%2D2000%2D2003/2003%2DPorsche%2DBoxster%2DEngine%2DCutaway.jpg)

These engines have contributed substantially to the nineteenth-century industrialization and human development (industrial process mechanization, increased production capacity, extended and fast transportations, and the lower prices of basic products making them affordable to the average family).

**Fig. 10.19** The future of internal combustion engines (Steve Dinan-6) (http://www.bimmerfest.com/photos/data/501/medium/Future%2Dof%2Dinternal%2Dcombustion%2Dengine%96Steve%2DDinan%2D6.jpg)

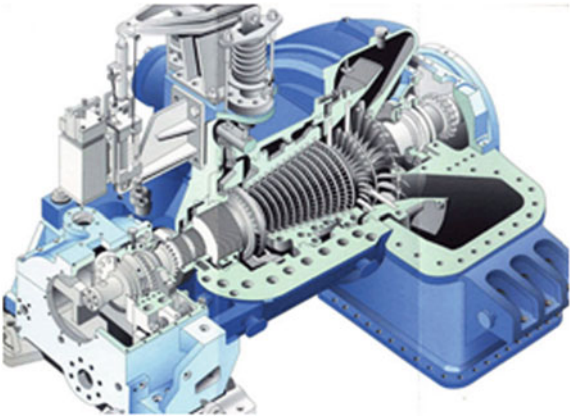


**Fig. 10.20** Steam turbine (http://www.greenesolpower.com/images/steam_turbine_htcprod3.gif)

During the last quarter of the nineteenth century, internal combustion engines and steam turbines started to replace the small and large steam engines, respectively (with efficiencies greater than 20% and Diesel engines reaching 30%) (Figs. 10.18, 10.19 and 10.20).

*Electricity* provided gave a further "boom" to the prime movers' and industrial development, with the *electric motor* reaching efficiencies of over 90% (Fig. 10.21).

The *gas turbine* was first used for aircraft jet propulsion in the 1930s and soon became the standard prime mover for the conversion of mechanical energy into electrical energy (Figs. 10.22 and 10.23).

Then, the nuclear reaction was developed and used both as a destructive weapon (Hiroshima and Nagasaki atomic bombs 1945) and as a power- generator plant (See Sect. 2.5).

Due to the above discoveries and progress, we have in our *modern life* available energy systems with increased efficiencies and large-scale production capacities

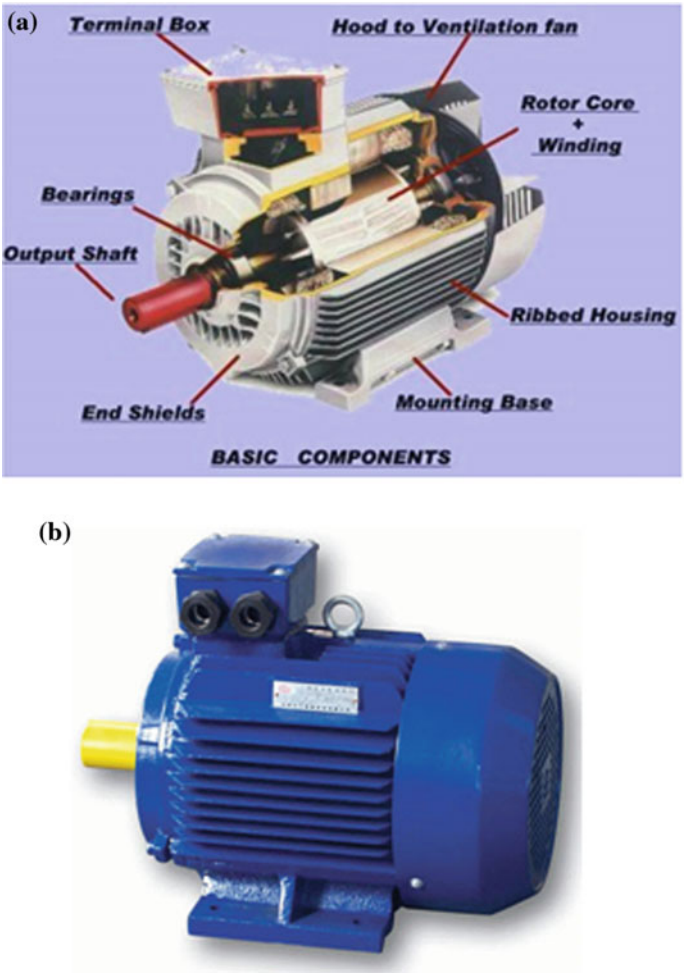



**Fig. 10.21** **a** Components of an electric motor, and **b** an example of an electric motor (**a** http://www.ckit.co.za/secure/conveyor/troughed/electric_motors/electric_motors_comp1.jpg; **b** http://www.pbaindustrial.com/uploads/product/20100428125948_AcDcMotor.gif)

(about 25–30 times more useful commercial energy than in 1900). This resulted in a parallel economic growth and improvement in the quality of life. Unfortunately, these impressive increases in energy availability and social development are still unevenly distributed throughout the world. Discontinuities and nonlinearities in the *gross national product*, *life expectancy*, and *adult literacy* of the countries of the world still exist. As mentioned in Sect. 1.4.2, these three measures of human development are combined in a unique measure called the *human development index* (HDI). According to the United Nations' *Human Development Report* (**HDR**), in 2007/2008 there were 22 countries with low development (HDI < 0.5),

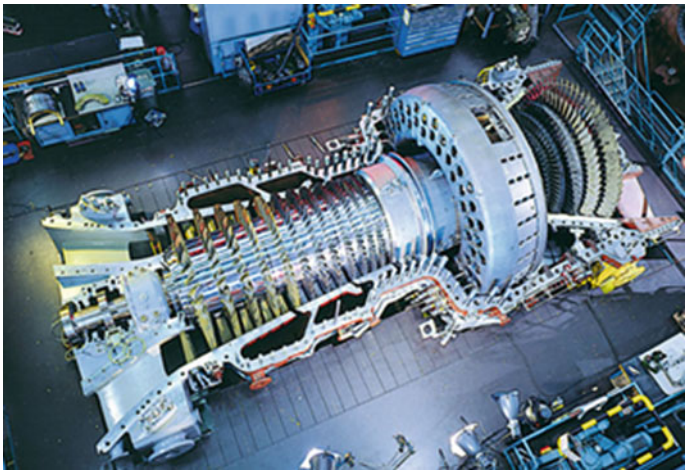


**Fig. 10.22** Gas turbine (http://www.dlr.de/next/Portaldata/69/Resources/images/4_energie/4_2_gasturbinen/Grafik_2_Gasturbine_680x450.jpg)

**Fig. 10.23** A gas turbine-based power station (http://www.oilcare.com/images/yootheme/powerstation.jpg)

located in southern Africa. Countries with high development (HDI $\geq$ 0.8) are those of North America, Western Europe, Oceania, and Eastern Asia (and some of the developing countries that are near HDI = 0.8 and have an ascending trend (www.undp.org/en/humandev/). It should again be emphasized here that energy is one of the most critical components that determine the fortunes of societies, which, however, does not dictate human choices but does secure economic success. On the other hand, the overuse of energy (especially the fossil-based exhaustible energy)

has a severe impact on the environment and Earth's biosphere, which constitute a challenge for modern human societies and their governments.

## 10.5 Energy and Economy

### *10.5.1 General Issues: Thermoeconomics*

Energy is closely connected to the economy. The cost of energy has always been a dominant factor in the performance of the economy of societies. We have seen that the elementary definition of energy is "*the capacity/ability to do work"*. A similar elementary definition of money is "*the ability to make other people work*". Money and its equivalents are the motive power of human activity [33, 34]. Our material world involves the flows of matter and energy, which are shaped by human activities into economic processes. The Earth is endowed by reserves of matter and energy. By the term *resources*, we mean the well-defined transformable parts of these reserves. Part of these resources is transformed into consumable goods that have a *value*, depending on complex social, cultural, conceptual, and political realities. The ownership and control of energy resources is a major factor in national and international politics. At the national level, governments control, in one or the other way, the distribution of the energy resources to the various sections of the society through pricing mechanisms. At the international level, "antagonism" is primarily dictated by the effort of the "great powers" to control the Earth's energy resources and the nations' economies.

*Production of energy* to fulfill sustainment of human needs has been a primary social activity. The types of energy include both exhaustible and inexhaustible forms, and today much effort is devoted to newer forms of energy, such as the production of hydrogen fuel from water. Until now, existing technologies are not adequate to make it a large-scale reality. Other forms of renewable and clean energy that are presently used include biofuels, vegetable oil (biodiesel), wind energy, and solar energy (see Sect. 2.5).

The *Production economy* deals with the production, exchange, and use of goods and the allocation of scarce resources, i.e., resources that can be utilized or exchanged for another scarce resource. The means of exchange for such exchanges is *money* of any form (coins, banknotes, financial "papers", contracts, leases, etc.).

*Price* is the amount of money given in exchange for a unit of a resource.

Resources include land, capital, materials, energy, and their combinations. Human resources are body power and mind power. When somebody uses a resource, another potential user may be deprived of this resource. In an open market, there are inputs (the entities entering the production process) and outputs or products (the utilizable entities that come out of the process) (Fig. 10.24).

Because the input and output entities are *heterogeneous*, they are all expressed with a common entity, i.e., a *monetary value.* This makes it possible to treat

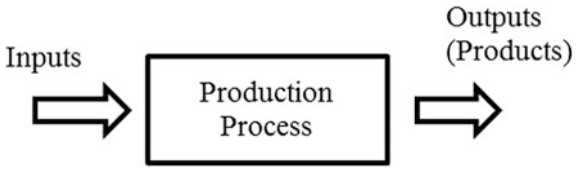



**Fig. 10.24** Systemic representation of a production process

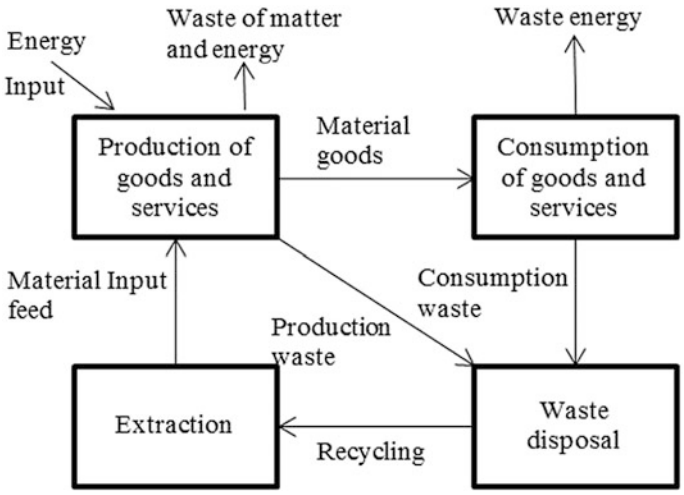



**Fig. 10.25** General equilibrium model, including matter and energy flows

everything (capital, labor, energy, materials, services, etc.) on an equal basis. Buyers and sellers of any item make their economic transactions using a common monetary basis. Under simple assumptions, there exists a price that simultaneously optimizes the utility of all factors for both the sellers and the buyers. In all cases, the regulating factor is the *invested capital*, which is also a condition for the generation of *cycles*. Capital and consumption imply the use of materials and energy, which indirectly is also true for services (even if they are pure mental services). Real materials never vanish truly, because materials return to "nature" as waste. This is illustrated by the *von Neumann equilibrium model*, 1945 (Fig. 10.25) [35].

Thermodynamically, the system is an "*open*" system. Extraction means taking resources from the Earth, and part of the waste returns to the Earth. Energy is conserved. What is wasted is available energy (exergy). The merciless second law of thermodynamics is always valid. Production waste is mainly mass, not available energy. The wasted energy returns to Earth as *heat* and cannot be recovered. But mass can be, at least in principle, recycled. Recycling, however, needs more available energy, which must be obtained from the Earth and also produces new waste and heat which is partly irradiated to space. The above are illustrated in Fig. 10.26 which shows the process of matter and energy flow including the economic cycle.

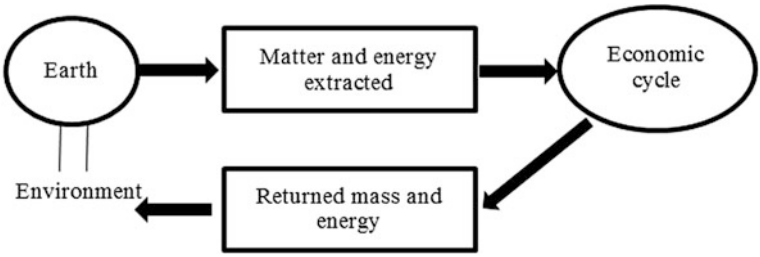



**Fig. 10.26** Generalized energy-matter-economic cycle (the flows are proportional to the amounts of energy and matter processed and transformed into "consumable goods")

The application of the laws of thermodynamics to the economy is called *thermoeconomics*, a term coined by *Myron Tribus* [36–39] and fully developed by *Nicholas Georgescu-Roegen* [40]. In thermoeconomics, economic systems always involve the flow of matter and energy, and the study considers the entire system, using the theory of thermodynamic systems in nonequilibrium (dissipative structures). Specifically, thermoeconomists consider society's economic activity as a dissipative system that grows by consuming available energy in exchange for, and conversions of, resources, goods, and services [37]. More information on this approach is given in Sect. 3.9.3 (economic systems thermodynamics).

### *10.5.2 Sectors of Economy*

From the previous discussion, it follows that the economic activity of a society involves the extraction/production of raw materials and energy, their transformation (possibly through intermediate products) into goods, and the provision of services.

This sequence of stages in the production chain specifies the following standard sectors of modern economic systems [41–45]:

- Primary sector
- Secondary sector
- Tertiary sector, and
- Quaternary sector

**Primary sector**: This sector deals with the extracts or harvest materials from the Earth, the production of raw materials (wood, coal, corn, iron, minerals, etc.), and the transformation of raw materials into basic goods and intermediate (unfinished) goods. Therefore, the primary sector of the economy includes subsistence and commercial agriculture, quarrying, mining, farming, hunting, fishing, and grazing. The accompanying processes of handling and packaging these materials and goods belong also to the primary sector.

**Secondary sector**: This sector involves the manufacturing of finished goods, either directly from raw materials or from unfinished materials. The manufacturing of cars (from steel), clothes (from textiles), and, in general, all manufacturing, processing, and construction activities belong to the secondary sector (e.g., engineering industries, the aerospace industry, shipbuilding, etc.).

**Tertiary sector**: This sector involves the service providing industry. Both services to the general population and to business are included, e.g., transportation, banking, tourism, entertainment, retail, restaurants, and the like. In economics, services are characterized as "*intangible goods*".

**Quaternary sector**: This sector was added to the tertiary sector in order to include separately all *intellectual activities* such as research, education, government, information technology, culture, libraries, etc. Some people classify education and government solely to the tertiary sector, while others add the *quinary sector* to include the high-level and senior-management processes, e.g., top management in nonprofit institutions, media, culture, arts, and government.

The proportion of workers in the primary sector (e.g., coal miners, loggers, fishermen, etc.) is continuously declining, especially in developed and developing countries. Workers in the secondary sectors include dressmakers, builders, potters, carpenters, and so on. Shopkeepers, accountants, dry cleaners, loan officers, real estate agents, bankers, and touristic agents are workers in the tertiary sector. The proportion of workers of developed and developing countries, which belong to the tertiary sector, is increasing. For example, in the USA, tertiary section labor now exceeds 80% of the total labor. The continuum of movement of goods and services from the primary, to secondary, and tertiary/quaternary sector is known in economics as the "*chain of production*". Other classifications of the economic sectors are the following [46–48]:

- *Based on ownership* (public sector, private sector, social economy (volunteer sector)) and
- *Based on product type* (industrial sector, service sector).

Here, a few words on the social or volunteer economy are worthwhile. Social economy is present in almost all economic sectors (primary, secondary, and so on). Social economy enterprises contribute a great deal to the society's health and are continuously growing. Some of them work in competitive markets, and others work very close to the public sector. Most of them are operating on the basis of voluntary participation, membership, and commitment. They are flexible and innovative.

It is obvious that the various sectors of the economy interact and influence each other. For example, the sourcing of raw materials is influenced by the demands occurring in the tertiary sector and the technological advances generated in the secondary sector. These issues should be taken into account by the top management of enterprises and by the government when creating and imposing economic and social policies and regulations.

## 10.6 Management of Energy

Overall, *the management of energy* plays a crucial role in society development by respecting the quality of the environment and saving the consumption of Earth's resources. Thus, management of energy is also called "*management of resources"* or "*energy saving*".

The meaning of energy management as "energy saving" involves the following four steps [49]:

- Monitoring, controlling, and conserving energy in a building, organization, or industry.
- Finding opportunities to save energy and calculating (or at least estimating) the energy amount that could be saved in each opportunity.
- Taking actions that secure that each energy-saving opportunity is followed (e.g., replace or improve the inefficient equipment).
- Evaluation of the metering records in order to see how well the energy-saving actions have been successful.

The importance of energy management (saving) stems from the global requirement to save energy, which of course influences energy prices, emission limits, and legislation for protecting the environment. Energy saving is one of the principal factors that contribute toward "sustainability".

A very practical manual for achieving energy conservation and cost saving in buildings and industry is provided in [50]. This document involves 400 thoroughly treated energy-saving measures, which optimize the use of energy in any kind of building or facility. The operation with respect to the energy use of a variety of equipment types (lighting, insulation, heating, cooling, motors, pumps, fans, etc.) is explained. Comparisons of the properties of energy sources from fossils to solar and wind are provided.

For example, one of these measures deals with "where light fixtures are needed in a predictable variety of patterns, install programmable switches." This measure provides a suitable and accurate method to match lighting to changing requirements. The selection scorecard includes *economics* (savings potential, cost, payback period), *ratings/priorities* (rate of return, new facilities, retrofit, operation, and maintenance), *reliability*, and *ease of retrofit* (or *ease of initiation)* which indicates how easy it is for people involved to perform the measure properly.

A brief description of the four energy-saving steps just cited is as follows:

**Monitoring and collecting data**: The old procedure was to manually read data once per week or month. The modern approach is to fit interval metering devices that measure and record automatically at regular short intervals (e.g., every 15 or 30 min). Clearly, the modern approach provides much more useful data for the process of waste energy saving.

**Finding and quantifying energy-saving opportunities**: The detailed data collected in the previous step are invaluable in the process of finding and locating the energy waste. The simplest and cost-effective energy-saving opportunities

(measures) usually require very small or no capital investment. An example is the proper configuration of heating and cooling equipment.

**Taking actions that target the saving opportunities**: The first prerequisite for this is *energy awareness*, e.g., identify the equipment that needs upgrading or insulation.

**Tracking the progress of energy saving**: The energy actions adopted and enacted must be evaluated professionally. Behavioral styles of work (e.g., getting people to switch off their computers) can lead to considerable saving and should be observed. The money invested for energy saving has to be followed by the levels of energy sought. Methodologically, the energy waste can be found by creating and using the so-called energy profiles, i.e., patterns of energy usage. These profiles can be created and used employing commercially available computer software. They show how much energy is used at particular times of the day and days of the week.

Abundant information on easy energy efficiency improvements and a guide for small businesses' energy efficiency is provided by the official business link to the US government [51]. Practical and easy rules for saving energy in houses can be found in [52]. Additional guidelines for energy saving in personal computers and other office equipment are provided in [53–56].

## 10.7 Demand Management, Economics, and Consumption of Energy

### *10.7.1 Energy Demand Management and Energy Economics*

*Energy demand management* deals with the determination of the proper policies and actions that aim at controlling and regulating the amount and trend of energy use by end users [57–60]. One of the most important problems of energy demand is to derive and apply control actions that reduce *peak demand*, especially during time periods of constrained availability of energy. Energy demand management control is designed to bring the energy supply as close as possible to a desired level of demand. One of the means of control is the "*price paid on the market*", which, however, does not always match the instantaneous cost since additional higher cost sources are needed in "peak periods". In many cases involving electrical energy, the capability or will of consumers to adjust to prices by changing their demand (elasticity of demand) are very low. The demand of energy, like that of any other commodity, can be controlled by policies and actions of market competitors or by government laws (taxation, etc.). In our times, energy-like demand policies are also applied for other resources such as water. The current trend in public and private energy providers is to take measures by which the efficiency of energy consumption and energy conservation is increased. Energy demand management is also called

"*demand-side management*", a term introduced during the periods of energy crisis (1973, 1979) [59].

Tightly connected to energy demand management is the branch of *energy economics* [61–63]. This branch is also related strongly to many other topics in economics, such as resource economics, environmental economics, industrial economics, econometrics, micro- and macroeconomics, etc. Other fields of science and technology that support energy economics, include ecology, climate policy, sustainability, energy markets, etc. Due to the above, *energy economics* is included in many university teaching and research programs within diverse departments. Energy demand management and energy economics finds application in the industrial sector, especially in the *energy industry* which involves all industries that deal with fossil fuel extraction, manufacturing, refining, saving, and distribution of energy (coal, oil, gas, electric nuclear, wind, solar, hydro, etc.). As we have seen in several places of this book, the Industrial Revolution, including, as a primary component, the energy sector, has made possible the development of large-scale production factories with high impact on the society's development.

### *10.7.2 Consumption of Energy*

According to worldwide studies, based on real-life evidence, if the energy consumption per capita is below 9000 KWh (about 1000 lt of oil) per year, life becomes laborious and the quality of life low. Above this limit, the quality of life does not actually depend on the level of energy [64]. Figure 10.27 shows the primary energy consumption in the 12 largest consuming countries in 2006. See also [65]. Table 10.1 shows the actual figures of the per capita energy usage per year by country in 2000—in **toe** (*tons of oil equivalent*). To convert "*kg oil equivalent*" (**kgoe)** into **kWh,** we multiply by the factor 11.628. For example, 8 toe = 8000 kgoe = 8000 kgoe × 11.628 = 93,024 kWh.

Figure 10.28 shows the energy consumption in the US and China for the period 1990–2020 compared to world consumption.

Actually, there are huge differences between developed and underdeveloped countries with similar patterns over several years. Research carried out in Zurich (ETH) showed that the energy usage per capita per year could be reduced in Switzerland to 17,500 kWh without compromising the quality of life, and, similarly, developed countries can reduce their per capita per year energy use between 17,500 and 20,000 kWh without degrading the quality of life. The current energy-usage map in the world shows that such reductions can be achieved with technologies already existing. See http://timeforchange.org.

With regard to global warming and climate change, the developed world must reduce its energy usage and especially fossil fuels. Fossil fuels, when they are burned in cars, heating, airplanes, etc., emit carbon dioxide which is a greenhouse gas and, as such, is the major cause of global warming. Worldwide, about 80% of the total energy used comes from fossil fuels (33% oil, 22% natural gas, 25% coal).

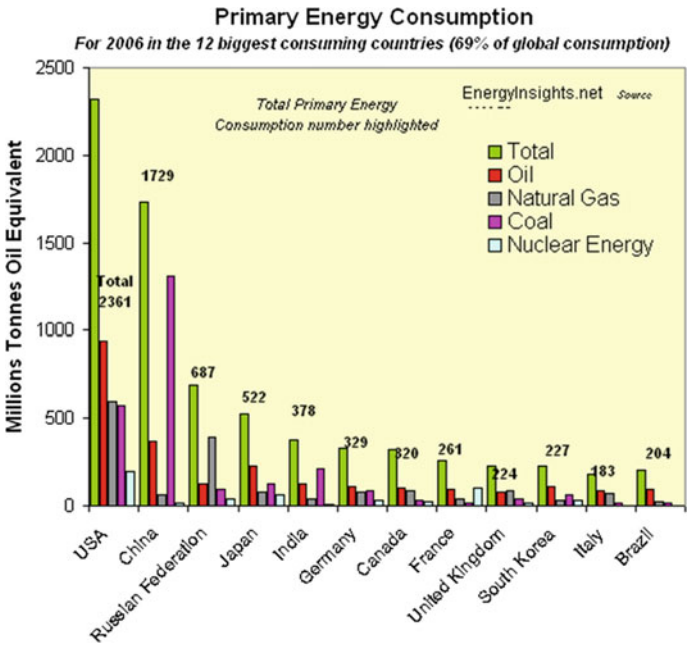



**Fig. 10.27** Primary energy consumption in the 12 largest consuming countries (http://www.energyinsights.net/cgi%2Dscript/csArticles/uploads/134/Primary%20Energy%20Consumption%20%2D%20per%20country%20in%202007%20%2D%20nuclear,%20coal,%20natural%20gas,%20oil%20and%20total.gif)

**Table 10.1** Per capita energy consumption year, by country, in 2000

| Rank | Country | Amount toe per capita | Rank | Country | Amount toe per capita |
|---|---|---|---|---|---|
| #1 | United States | 8.35 | #10 | France | 4.25 |
| #2 | Canada | 8.16 | #11 | Japan | 4.13 |
| #3 | Finland | 6.4 | #11 | Germany | 4.13 |
| #4 | Belgium | 5.78 | #13 | UK | 3.89 |
| #5 | Australia | 5.71 | #14 | Ireland | 3.86 |
| #6 | Sweden | 5.7 | #15 | Switzerland | 3.7 |
| #6 | Norway | 5.7 | #16 | Denmark | 3.64 |
| #8 | New Zealand | 4.86 | #17 | Austria | 3.52 |
| #9 | The Netherlands | 4.76 | #18 | Italy | 2.97 |
| | | | | **Total:** | **89.51** |

Weighted average: 50 toe per capita

Nuclear fuels provide about 5.5% and renewable sources about 16.5% (5.5% hydroelectric and 11.0% noncommercial biomass, i.e., wood, hay, fodder, and other fuels, in rural economies). The total energy use worldwide by source is shown in

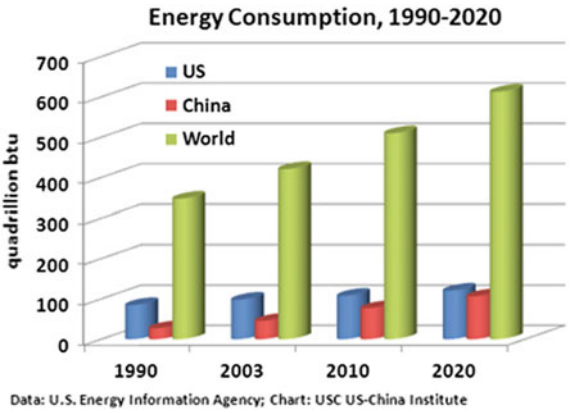



**Fig. 10.28** World, US, and China energy consumption compared (1990–2020) (http://www.environmentabout.com/wp%2Dcontent/uploads/2010/06/World%2Denergy%2Dconsumption1.jpg)

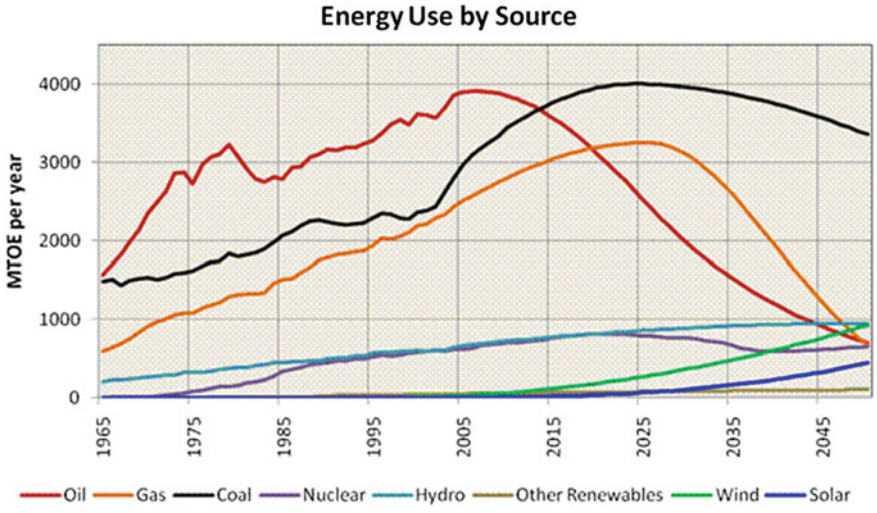



**Fig. 10.29** World energy use by source (http://www.paulchefurka.ca/WEAP2/image030.gif)

Fig. 10.29. The share of fossil fuels by percentage of the total energy used in the various countries of the world is presented in [64].

The predictions of EIA are that, in the world market, energy use derived from fossil fuels will be worse in the future: about 18% more in 2015 and 40% more in 2030 (absolute values) compared to the 2006 figure [66]. However, the average annual growth rate of liquid fuels (0.9%) is much smaller than the growth rate of

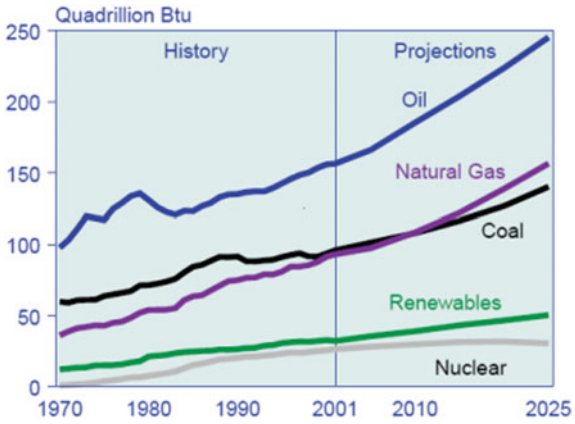



**Fig. 10.30** World energy use by fuel type (1970–2025) [1 kWh of electricity = 3413 British thermal units (btu)] (http://www.timeforchange.org/prediction%2Dof%2Denergy%2Dconsumption)

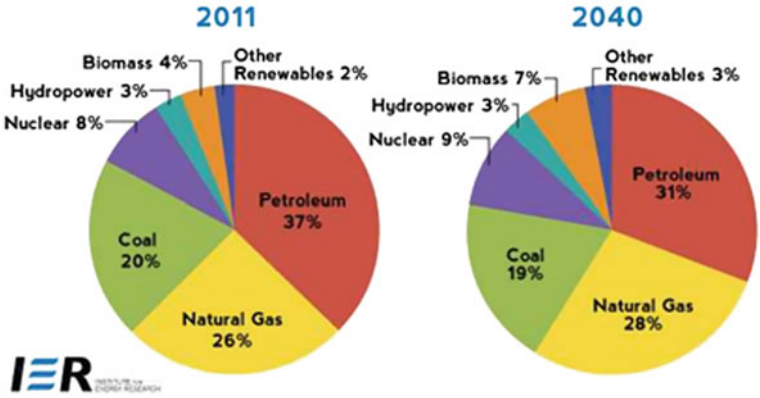



**Fig. 10.31** Percentage share of fossil fuels in 2011 and projected for 2040 (http://www.mailmagazine24.com/politics/12%2D2012/fossil%2Dfuels%2Dstill%2Dking%2Din%2Deias%2Dannual%2Denergy%2Doutlook%2D2013a.jpg)

renewable sources (3.0%). This amounts to a decline of the relative energy consumption based on liquids from 36% in 2006 to 32% in 2030 [66]. See also

http://www.eia.gov/cfapps/ipdbproject/iedindex3.cfm%3Ftid%3D44&pid%3D45&aid%3D2&cid%3Dregions%26syid%3D2005&eyid%3D2009%26unit%3DQBTU

Figure 10.30 shows world energy use by fuel type for the period 1970–2025 based on predictions made in 2001.

Figure 10.31 shows the world fuel energy share in 2011 and the prediction for 2040.

In [66], it is predicted that the energy consumption in OECD countries will increase slowly over the projection period up to 2030, about 0.6% per year (average), while the corresponding average increase in non-OECD countries is 2.3%. Overall, by 2030, OECD countries are predicted to show an increase of about 15%, and non-OECD countries an increase about 55% compared to the 2006 consumption figure. In 2006, 51% of the world energy consumption was in OECD economies. This will fall to 40% in 2030. This is a good sign of the allocation of more energy to poorer economies [66]. Further statistical data on the world's energy, energy balance, fuel reserves, and world's resources are provided in [67–76].

Net electricity generation worldwide in 2003 was 18.0 trillion kWh and in 2006 31.8 trillion kWh, showing a 77% increase. OECD countries have the strongest growth in electricity generation (an increase of about 3.5% per year) which raises the living standards of their people (more home appliances, commercial services, schooling, etc.). The share of renewable energy sources in the world's electricity generation shows the fastest growth (about 2.9% per year from 2006 to 2030), with the dominant components being hydroelectric and wind power. These are good signs, but, even so, the high increase of fossil fuel consumption should be moderated even more, as soon as possible.

The world reserves of oil, coal, and natural gas are reported in Fig. 10.32.

From 1990 to 2000, a total of 42 billion barrels of new oil reserves were discovered. In this period, the world oil consumption was 250 billion barrels (1 barrel of oil contains 5.8 million btu). According to the US Department of Energy 1996 Annual Energy Review [77], if the rate of fossil fuel energy use continues the same as in 1995, then the fossil fuel energy resources (proven as of January 1, 1996) will be exhausted in the year 2111. If the rate continues increasing linearly, with the average rate of the period 1973–1995, then the above oil reserves will be used up by 2074. Although these are crude estimates (since the actual fossil fuel energy use in the future will be determined by many unforeseen factors), it is certain that the Earth's fossil fuel reserves are limited, and, if we keep using them at a rate near to what we do presently, we will run out "fairly soon". In the meantime, human society has to switch to some other source(s) of energy.

According to [78], at the end of 2007, the world's conventional oil reserves in fields that have already been discovered, but not yet developed, amounted to about 257 billion barrels (IEA analysis based on IHS data). OPEC countries' share is a little more than a half of this amount (about 133 million barrels). About 62% of OPEC undeveloped conventional reserves are located onshore. In non-OPEC countries, onshore reserves constitute about 38% of the total reserves. More than half of the world's undeveloped reserves are located in the Middle East and Russia, with 40% of the world's total in the Middle East (about 99 billion barrels).

The 257 billion barrels of yet-to-be developed conventional oil reserves are distributed in 1874 fields (971 onshore and 903 offshore). The average field size in OPEC countries is twice that of fields in non-OPEC countries as shown in Table 10.2, which is based on IEA analysis using IHS data. (See www.iea.org/textbase/nppdf/free/2008/weo2008.pdf).

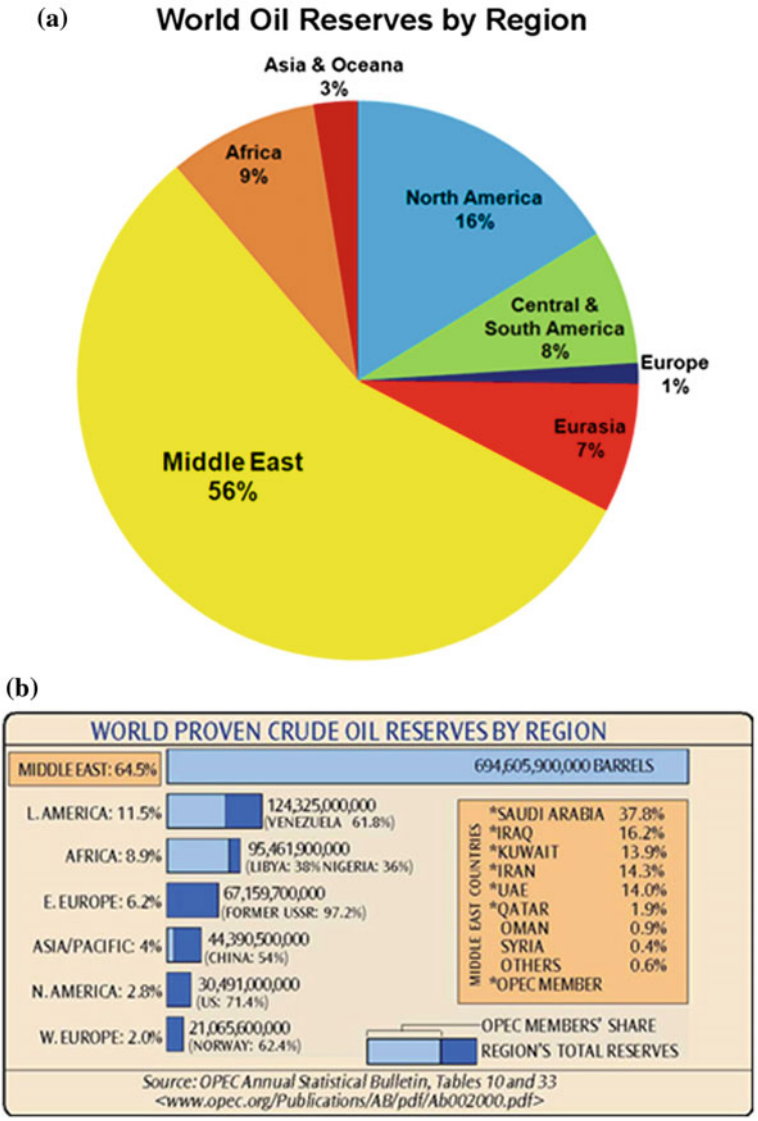



**Fig. 10.32** World proven reserves of oil, coal, and natural gas (oil by region, coal and gas for the respective top 20 and 15 countries). **a** US Energy Information Administration (Oil and Gas Journal, 2007) https://newsocialistproject.files.wordpress.com/2010/11/world_oil_reserves_by_region.png%3D. **b** http://www.theglobaleducationproject.org/earth/images/oilgraph2.jpg. **c** http://www.energyinsights.net/cgi%2Dscript/csArticles/uploads/58/Coal%20Reserves%20Per%20Countries%20%20%2D%20biggest%2020%20countries%20by%20reserves.gif. **d** http://www.arcticgas.gov/sites/default/files/images/world%2Dproven%2Dreserves.png

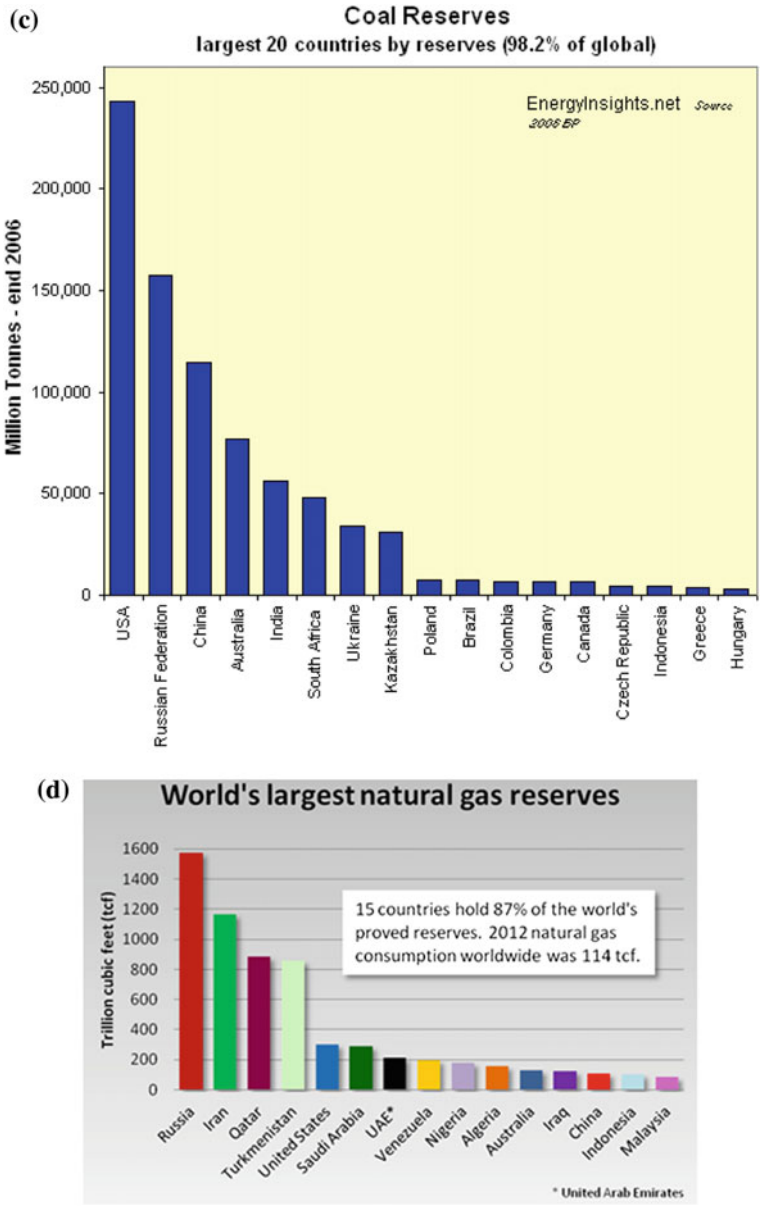



**Fig. 10.32** (continued)

The oil fields yet-to-be-found are anticipated to reach a conventional oil production of 19 million barrels in 2030, based on the projected discovery of 114 billion barrels of reserves worldwide over the projection period. The bulk of

production from yet-to-be found offshore fields comes from non-OPEC countries (estimated to 7.9 mb/d out of a total of 10.7 mb/d in 2030). Non-OPEC offshore yet-to-be-found fields are about equally divided between Russia and other Eurasian countries, Africa, and OECD North America.

Figure 10.33 shows the world map of proven and undiscovered oil reserves.

**Table 10.2** Average size of yet-to-be developed oil fields by region in millions of barrels (end of 2007)

| Region | Onshore | Offshore | Overall average |
|---|---|---|---|
| OECD North America | 28 | 102 | 62 |
| OECD Europe | 64 | 52 | 53 |
| OECD Pacific | 33 | 48 | 47 |
| Russia | 160 | 570 | 187 |
| Other Europe/Eurasia | 116 | 854 (228[a]) | 310 (142[a]) |
| Asia | 58 | 60 | 59 |
| Middle East | 324 | 368 | 335 |
| Africa | 74 | 104 | 91 |
| Latin America | 36 | 232 | 152 |
| OPEC | 213 | 182 | 201 |
| Non-OPEC | 84 | 119 | 103 |

[a]Excluding the Kashagan field in Kazakhstan

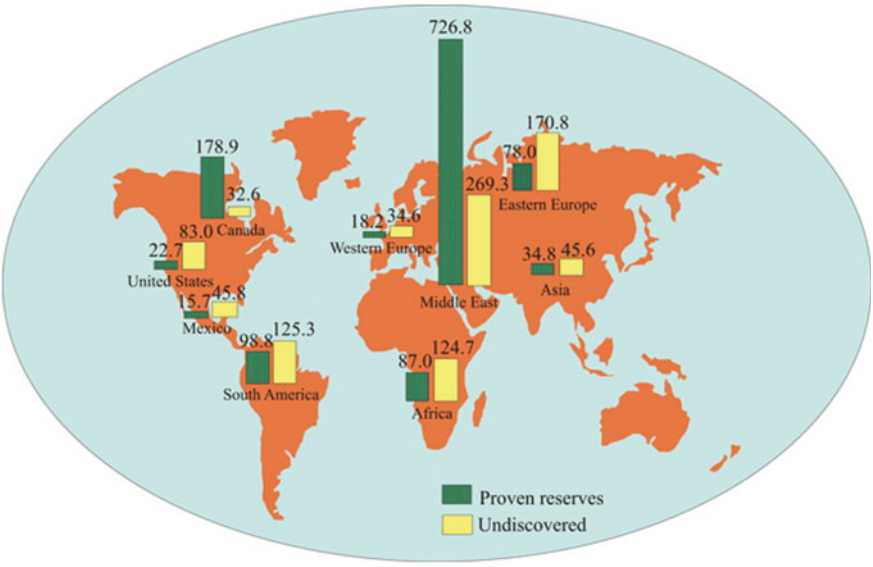



**Fig. 10.33** Distribution of proven and undiscovered oil reserves in the world (https://prienceshrestha.files.wordpress.com/2012/01/fig%2D4%2Doil%2Dreserves.gif)

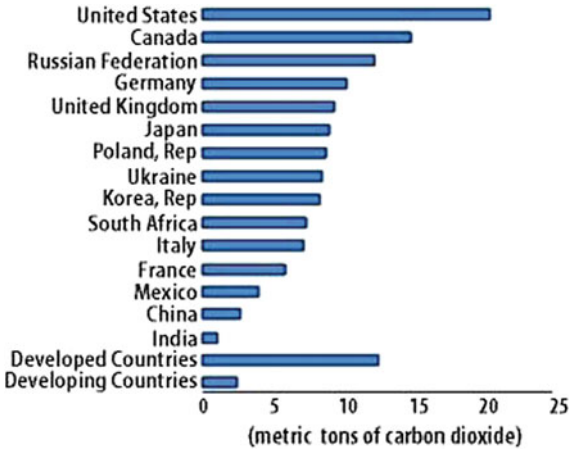



**Fig. 10.34** The cumulative emissions of $CO_2$ for the 15 top countries [79]

Figure 10.34 shows the cumulative $CO_2$ emission in the period 1900–1999 for the top 15 countries out of a total of 193 countries [79]. In the USA, about 42% of $CO_2$ and other toxic emission are due to "400-plus coal-fired power plants" (Toxic Release Inventory 1, 2002).

Finally, Fig. 10.35 shows the distribution of carbon dioxide emissions in OECD and non-OECD countries for the period 1991–2005, and the predicted emissions for the periods 2006–2020 and 2021–2035.

Summarizing, the nonrenewable energy reserves of the Earth are anticipated to be exhausted "fairly soon", and the use of fossil fuels is the major cause of *greenhouse gas* **(GHG)** emissions ($CO_2$ and other substances) and global warming (fully documented in the relevant literature). This indicates that humankind and the planet have a serious *sustainability* problem.

These phenomena can be mitigated if at least the following actions are taken "now" [73–77, 79–81]:

- Reduce drastically the current usage of fossil fuels.
- Develop and use renewable and alternative energy sources as much as possible.
- Change our present lifestyle and use energy more efficiently as discussed in Sect. 10.6.

Of course, energy use and human quality of life improvement are strongly related to the population growth of the Earth's nations. It is a statistically measured fact that nations with high standards of living have weak or no demographic growth. On the contrary, nations with low standards of living (underdeveloped countries) show high population growth rates, sometimes reaching 100% increase in 25–30 years. As we mentioned at the beginning of this section, a minimum of 9000 to 10,000 kWh energy use is necessary for an acceptable level of quality of

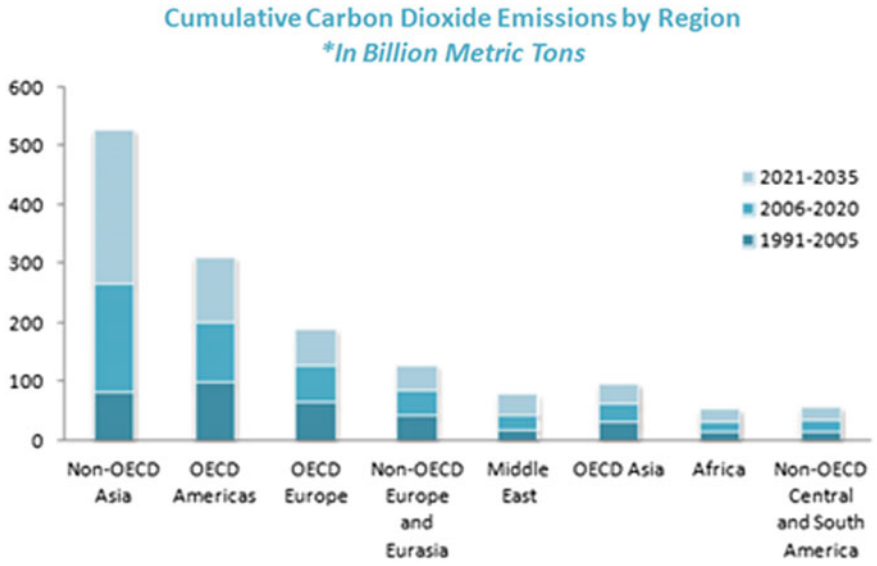



**Fig. 10.35** $CO_2$ emissions for OECD and non-OECD countries (1991–2035). (Energy Information Administration) (http://static.reportlinker.com/public/images/clp/energy/enviroregion.PNG)

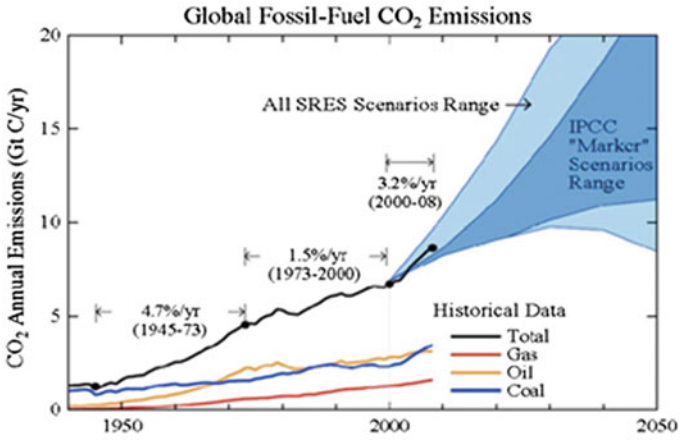



**Fig. 10.36** The carbon dioxide emissions are severely increasing after the Kyoto Protocol, first adopted in 1997 and entered into force in 2005 [78, 87] (https://bravenewclimate.files.wordpress.com/2011/07/fossil_fuel_emissions.jpg)

life. All the above issues indicate the real need to develop and apply national and international energy and environmental policies with long-term goals, as has been recognized by all international bodies and institutes (e.g., [82–86]).

In [78], a world energy scenario, up to 2030, is worked out on the basis of data available at the end of 2007. This seems to be the most recent outlook of world energy which is more accurate than other reports previously published. An updated study of world's climate change, by sector and major regional group, is also provided taking into account the interaction of energy and climate. Figure 10.36 shows the exponential trend of present $CO_2$ emissions on Earth and reveals the necessity to take fast and strong actions that will correct the situation and secure the Earth's overall sustainability.

## 10.8 Concluding Remarks

This chapter has presented the fundamental issues that characterize the connection of energy to life and human society. We have seen that the Sun's energy is stored in ATP molecules (as chemical energy) through the process of photosynthesis. This energy is released when ATP is converted to ADP through the process of dephosphorylation. Ecosystems consist of *abiotic and biotic* entities. To live, an ecosystem needs continuous solar energy; otherwise, it ceases to survive. This means that the ecosystem is an open thermodynamic system with regard to energy, but is a closed system with respect to the material. In an ecosystem, the energy flows through the food chains and food webs, always with a loss of "quality" at each step from one trophic level the next.

Energy also flows, is converted, and is used in human society. These processes are the subject of *general energetics*. The energy types used by human societies as they evolved are biomass, human/animal muscle energy, preindustrial inanimate prime movers, fossil fuels, mechanical prime movers, electricity, and modern energy systems. Energy influences the economy of a society having a "*value*" and a "*price*" that depends on this type and the society's *geopolitical factors*. The management of energy to increase the efficiency of its use (which results in energy saving) is today one of the main concerns of human society. Also in our time, the management of energy demand so as to minimize the *peak demand* is a necessity. Finally, the use (consumption) of available energy in the various parts of the world needs to be smoothened and balanced to reduce as much as possible the current differences, while securing the Earth's and human society's sustainability. In conclusion, all the above aspects show the important and critical role that energy plays both for life and society, providing the fuel required for their existence, activity, and sustainability. Some further references on energy in life and society are [88–106]. Reference [107] presents a review of climate and atmospheric history of the Earth.

## References


1. W.N. Marchuk, *A Life Science Lexicon* (W.C. Brown Publishers, Dubuque, IA, 1992)
2. J. de Rosnay, *The Macroscope: A New World Scientific System*, (Translation from French by R. Edwards) (Harper & Row Public, NY, 1979). http://pespmcl.vub.ac.be/macroscope/default.html
3. N.K. Wessels, J.L. Hopson, *Biology*, 3th edn. (Random House, New York, 1994)
4. N.A. Campbell, G.M. Lawrence, J.B. Reece, *Biology*, 5th edn. (Benjamin/Cumming Publ., Menlo Park, CA, 1999)
5. W.K. Purves, D. Sadava, G.H. Orians, C. Heller, *Life: The Science of Biology*, 7th edn. (Sinauer Associates and W.H., Freeman, New York, 2004)
6. D.O. Hall, K.K. Rao, *Photosynthesis*, 5th edn. (Cambridge Univ. Press, Cambridge, 1994)
7. Photosynthesis: History of Research. http://science.jrank.org/5196/Photosynthesis.html, http://www.indepthinfo.com/biology/photosynthesis.shtml
8. D. Watson, *Photosynthesis and Energy in nature*. http://www.ftexploring.com/me/photosyn1.html
9. ATP and Biological Energy. http://www.emc.maricopa.edu/faculty/farabee/biobk/BioBookntp.html
10. Metabolism: Definition http://www.thefreedictionary.com/metabolism
11. L. Taiz, E. Zeiger, *Surface Protection and Secondary Defense Compounds: Plant Physiology* (Benjamin/Cummings Publ., 1991)
12. D. Watson, *Food Webs—Food Chains—Energy Pyramids*. http://www.ftexploring.com/links/foodchains.html
13. A. Larocque, *Overview of Photosynthesis: The Conversion of Light Energy into Chemical Energy by Plants*. http://biology.suite101.com/article.cfm_overview_of_photosynthesis
14. A. Satyanarayana, *Using Photo-synthesis to Great Renewable Energy*. http://www.brightlub.com
15. C.R. Nave, *Adenosine Triphosphate*, *Georgia State University*. http://hyperphysics.phy-astr.gsu.edu/hbase/biology/atp.html
16. Adenosine Triphosphate@Everything2.com. http://everything2.com/index.pl%3Fnode_id%3D167956
17. S.F. Chapin, P.A. Matson, H.A. Moonly, *Principles of Terrestrial Ecosystem Ecology* (Springer, New York, 2002)
18. E.P. Odum, *Fundamentals of Ecology* (W.B. Saunders Co., PA, 1959)
19. R.L. Lindeman, *The Trophic-Dynamic Aspects of Ecology*, vol. 23 (1942), pp 399–418
20. M.H. Stevens, *A Primer of Ecology* (Springer, Berlin, 2009)
21. W. Whitman, *The Concept of the Ecosystem* (2008). http://www.globalchange.umich.edu/globalchange1/current/lectures/
22. B. Kalman, J. Langille, *What are Food Chains and Webs* (Crabtree, Publ. Co., New York, 1998)
23. Food Chains. http://users.rcn.com/jkimball.ma.ultranet/Biology/Pages/F/FoodChains.html
24. D.E. Watson, *Energy Pyramids and Food Chains*. http://www.ftexploring.com/me/pyramid.html
25. V. Smil, *Energy in Nature and Society: General Energetics of Complex Systems* (The MIT Press, Cambridge, MA, 2008)
26. B. Stone Barger, *Energy and Society* (Hawkhill Associates, Madison, WI, 1990)
27. H.H., Schobert, *Energy and* Society: *A Introduction* (Taylor and Francis, London, 2002)
28. D. Elliott, *Energy Society and Environment: Technology for a Sustainable Future* (Routledge Taylor and Francis, London, 2003)
29. W.F. Cottrell, *Energy and Society: The Relation between Energy, Social Change, and Economic Development* (Greenwood Press, Santa Barbara, CA, 1970)
30. E.W. Miller, R.M. Miller, *Energy and American Society: A Reference Handbook* (ABC/CL10, Santa Barbara, CA, 1993)

31. V. Smil, *Energy* (Berkshire Publishing, Berkshire Encyclopedia of World History, 2007), pp. 646–654
32. V. Smil, *Energy in World History* (Westview, Boulder, CO, 1994)
33. G. Garvey, *Energy, Ecology, Economy* (MacMillan, Green with CONN, 1974)
34. B. Cimbleris, *Economy and Thermodynamics*. http://ecen.com/eeeq/ecoterme.htm
35. J. von Neumann, *Rev. Economic Studies*, vol 13, no 1 (1945)
36. P. Corning, *Thermoeconomies: Beyond the Second Law*. http://www.complexsystems.org/abstracts/thermoec.html/
37. P. Burley, J. Foster, *Economics and Thermodynamics: New Perspectives on Economic Analysis* (Kluwer, Boston/Dordrecht, 1994)
38. Y. El-Sayed, *The Thermoeconomics of Energy Conversions* (Pergamon, London, 2003)
39. S. Sieniutycz, P. Salamon, *Finite-Time Thermodynamics and Thermoeconomics* (Taylor and Francis, London, 1990)
40. N. Georgescu-Roegen, *The Entropy Law and the Economic Process* (Harvard University Press, Cambridge, MA, 1971)
41. About.com: Geography-Sectors of Economy. http://geography.about.com/od/urbaneconomicgeography/a/sectorseconomy.htm
42. Wikipedia: Economic Sector. http://en.wikipedia.org/wiki/List_of_recognized_economic_sectors
43. What are Sectors of Economy? http://www.wisegeek.com/what%2Dare%2Dsectors%2Dof%2Deconomy.htm
44. Primary Sectors in Economic Development. http://links.jstor.org
45. http://www.twnside.org.sg/title2/gtrends6.htm
46. EU Social Economy Enterprises. http://caledonia.org/uk/eu%2Dsee.htm
47. M. Graham, J. Woo (eds.), *Fueling Economic Growth: The Role of Public–Private Sector Research in Development*, Intern. Development Research Center (Ottawa) (Practical Action Publ. Ltd, Rugby, UK, 2009)
48. J.E. Stiglitz, *Economics of the Public Sector*, 3rd edn (W.W., Norton, New York, 2000)
49. Energy Management. http://www.energylens.com/articles
50. D.R., Wulfinghoff, *Energy Efficiency Manual*, Energy Institute Press. http://www.energybooks.com
51. Easy Energy Efficiency Improvements, Small Business Guide to Energy Efficiency http://www.business.gov/expand/green%2Dbusiness/energy-efficiency
52. EIA Energy Efficiency. http://www.eia.gov/emeu/efficiency/energy_savings.htm
53. Energy Saving: Information Technology Services, Energy Saving for Personal Computers. http://www.colorado.edu/its/docs/energy.html
54. http://www.energystar.gov/energy.html
55. http://www.energy-solution.com/off-equip/technical.html
56. Twenty Five Simple Ways to Save, Consumer Reports. http://www.consumerreports.org/cro/home%2Dgarden/resource%2Dcenter/saving%2Don%2Denergy%2Dcosts/25%2Dsimple%2Dways%2Dto%2Dsave
57. J.L. Sweeney, The response to energy demand to higher press: what have we learned? *Am. Econ. Rev.* **74**(2), 31–37 (1984)
58. L.D. Taylor, The demand for electricity: a survey. *The Bell J. Econ.* **6**, 74–110 (1975)
59. D.S. Loughran, S.D. David, J. Kulick, Demand-side management and energy efficiency in the U.S. *The Energy J.* **25**(1) (2004)
60. Energy Demand Management, Wikipedia. http://en.wikipedia.org/wiki/Energy_demand_managment
61. A.N. Kneese, J.L. Sweeney (eds.), *Handbook of Natural Resource and Energy Economics*, vol III (Elsevier, Amsterdam, 1993)
62. J.M. Griffin, H.B. Steele, *Energy Economics and Policy* (Academic Press, N.Y., 1986)
63. J.L. Sweeney, Economics of Energy. http://www.stanford.edu/%7Ejsweeney/paper/EnergyEconomics.PDF

64. Global Warming Solutions-Sensible Energy Consumption, Time for Change and Prediction of Energy Consumption. http://timeforchange.org
65. Usage per Person (Most Recent) by Country, Nationmaster. http://www.nationmaster.com/pie.ene_usa_per_per_energy_usage_per_person
66. Energy Information Administration, International Energy Outlook 2009 & Projections, *EIA: World Energy Projections Plus* (2009). http://www.eia.doe.gov/iea
67. IEA, Energy Balance of OECD Countries 1999–2000, IEA, Paris (2001)
68. World Energy: The Good, Bad and BTUs. http://www.ecoword.com. (World Energy Consumption)
69. Fossil Fuel Reserves-to-Production (R/P) Ratios at End 2009. http://www.bp.com
70. World of Resilience: Growing the Wealth of the Poor, *World Resources Institute*, Washington, DC (2008). http://earthtrends.wri.org
71. Primary Energy Consumption per Capita 2009, *Statistical Review 2010*. http://www.bp.com
72. Energy Statistics Newsletter, *UN Statistics Division*, Issue No. 2, March 2006
73. G.P. Beretta, World energy consumption and resources: an outlook for the rest of the century. Int. J. Environ. Technol. Manage. **7**(1–2), 99–112 (2007)
74. R.C. Duncan, World energy production, population growth, and the road to the Olduvai Gorge. Popul. Environ. **22**(5) (2001)
75. A. Guïbler, N. Nakienovi, Decarbonizing: doing move with lees. Technol. Forecast. Soc. Change **51**(1), 97–110 (1996)
76. R. Shinnar, The hydrogen economy, fuel cells and electrical cars. Technol. Soc. **25**, 455 (2003)
77. CPAST: Overview of Fossil Fuel Energy Resources. http://www.cpast.org/Articles/fetch.adp%3Ftopicnum%3D14
78. http://unfccc.int/kyoto_protocol/status_of_ratification/items/2613.php
79. Time for Change, *$CO_2$ Emissions by Country*. http://timeforchange.org/CO2%2Demissions%2Dby%2Dcountry
80. S.G. Tzafestas, *Human and Nature Minding Automation: An Overview of Concepts, Methods, Tools and Applications* (Springer, Dordrecht/Berlin, 2010)
81. World Energy Supply, *Earth: A Graphic Look at the State of the World*. http://www.theglobaleducationproject.org/earth/energy%2Dsupply.php
82. World Watch Institute, Making Better Energy Choices. http://www.worldwatch.org/node/808
83. P.L. Bishop, *Pollution Prevention: Fundamentals and Practice* (McGraw-Hill, Boston, MA, 2000)
84. United Nations (UN). http://www.un
85. Organization for Economic Cooperation and Development (OECD). http://www.oecd.org
86. ETC-LUCI. http://etc-luci.eionet.europa.eu
87. http://en.wikipedia.org/wiki/Kyoto_Protocol
88. P. Grobstein, *The Essential Link between Life and the Second Law of Thermodynamics* (Serendip, Bryn Mawr, PA, 2000). http://serendip.brynmawr.edu/biology/links.html
89. V. Smil, Moore's Curse and the Great Energy Delusion. Am. Mag., 1–6 (Nov, 2008). http://www.american.com/archive/2008/november%2Ddecember%2Dmagagine/
90. Energy and Society-Wikipedia. http://en.wikipedia.org/wiki/Energy_and_Society
91. Energy Management: Energy Monitoring and Targeting. http://www.energylens.com/articles
92. B. Cimbleris, *Economy and Thermodynamics*. http://ecen.com/eee9/ecoterme.htm
93. Time for Change, Prediction of Energy Consumption World-Wide. http://timeforchange.org/prediction%2Dof%2Denergyconsumption
94. Time for Change, Global Warming Solutions-Sensible Energy Consumption. http://timeforchange.org/global%2Dwarming%2Dsolutions%2Denergyconsuption
95. I. Dincer, M.A. Rosen, *Exergy: Energy, Environment and Sustainable Development* (Elsevier, Amsterdam, 2007)
96. D.R. Wulfinghoff, *Energy Efficiency Manual* (Energy Institute Press, Wheaton, Maryland, 2003). http://www.energybooks.com/

97. Business.Gov, Easy Energy Efficiency Improvements. http://www.business.gov
98. Consumer Reports.Org, Save Energy, Save Money, Cars, 25 Simple Ways to Save, *Consum. Rep. Mag*. (Oct, 2008). http://www.consumerreports.org/
99. Renewable Energy and Energy Efficiency: Economic Drivers for the 21st Century, American Solar Energy Society, R. Bezdek, Management Information Services, Inc. (2007). http://www.ases.org
100. Poverty, Energy and Society, The Baker Energy Forum, Rice University. http://www.rice.edu/energy/research/poverty%26energy/index.html
101. L. Hughes, *The four 'R's of energy security. Energy Policy* (2009). doi:10.1016/j.enpol.2009.02.038
102. EIA (Energy Information Administration) Energy Efficiency. http://www.eia.doe.gov/emeu/efficiency/energy_savings.html
103. Saving Energy. http://www.energyquest.ca.gov/saving_energy/index.html
104. J. Masserau, *Petroleum Economics*, 4th edn. (Edition Technip, Paris, 1990)
105. W.S. Peirce, *Economics of the Energy, Industries*, 2nd edn. (Praeger Publishers, Westport, CT, 1996)
106. M.S. Raymond, W.L. Leffler, Oil and Gas Production in Nontechnical Language, *PennWell* Corp., Tulsa, Oklahoma (2005)
107. J.R. Petit, J. Jouzel et al., Climate and atmospheric history of the past 420,000 Years from the Vostok Ice Core, Antarctica. Nature **399**, 429–436 (1999). http://www.glu.uga.edu/railsback/1122/PettitetalEKG900.jpeg

# Chapter 11
# Information in Life and Society

*The only good is knowledge and the only evil is ignorance.*
Socrates
*The possession of knowledge does not kill the sense of wonder and mystery. There is always more mystery.*
Anais Nin

**Abstract** Information is present in all natural, living, and technological systems, and is recognized as the third basic universal quantity after energy and matter. For this reason, information manifestations in both natural and man-made systems have attracted the interest of humans through the historical evolution of the humankind. On the life and biological side of information there are two axes of study, namely: (i) the study of the underlying natural/biological mechanisms of storing, processing, and transmission of information from cells to entire organisms, and (ii) the use of biological mechanisms of computation in the design and implementation of new types of man-made computational systems. On the technological side, information and communication technology (ICT) is increasingly entering to the "heart" of large-scale competitive policies, due to its capacity as a key player in the ongoing human growth, development, and modernization. This chapter is concerned with the role and application of information to life and society. Regarding the life side the issues of the substantive role and the transmission sense of information in biology, the natural information principles, and biocomputation, are discussed. On the society side, the application of IT to office automation, power generation and distribution, computer-integrated manufacturing, robotics, business and electronic commerce, education, medicine, and transportation, is investigated. This chapter ends with a look at the issues of social networking, and ethics of IT (infoethics).

**Keywords** Information · Life · Society · Intentional/functional information
Information society · Information revolution · Information technology (IT)
Information store principle · Borrowing/reorganizing principle · Biocomputation
IT in office automation · IT in business · IT in medicine · IT in power systems
IT in manufacturing · IT in transportation systems · Intelligent transportation system (ITS) · IT in education · CIM managerial functions · CIM financial functions



S. G. Tzafestas, *Energy, Information, Feedback, Adaptation, and Self-organization*,
Intelligent Systems, Control and Automation: Science and Engineering 90,
https://doi.org/10.1007/978-3-319-66999-1_11

## 11.1 Introduction

Energy and matter are recognized as the two fundamental universal quantities. Information has now become the third basic universal quantity. It is one of the endogenous features of life and society and is present in all natural and man-made systems, e.g., living cells, ecosystems, atmosphere, data processing systems, communications, control systems, human languages, social systems, information systems, industrial systems, economic systems, etc.

The progress of computers obeys *Moore's law* according to which computers steadily achieve improved performance, need less power, and fall in price. The same is true for everyday communications and data exchange between computers that are far physically far apart. Throughout the world, *information and communication technology* (**ICT**) has empowered people with enormous access to information and knowledge, with highly beneficial consequences for education, medical care, markets, business, and social interactions. Moreover, the ICT's extraordinary features enable it to play a dominant and critical role in the social and economic growth of all nations into which it has been introduced.

Particular technologies and ICT processes of great importance include, among others

- Mobile telephony and its role in networked readiness
- The shift from mobility to ubiquity, thanks to universal Internet connectivity.
- Medical informatics and telemedicine
- Competitive market operations that balance investment incentives and efficient service.

According to *Leonard Waverman* (London Business School) and *Kalyon Dasgupta* (LECG), the development of an *information society* passes through the following four stages, which are called "*waves*"[1]:

- Simple access
- Universal service
- Usage
- Provision of complementary skills and assets.

They point out that so far research and development efforts were concentrated in the first two waves and that additional research in the last two waves is needed for understanding the role of *usage* and complementary capital to control the profits from ICT in a developing country context.

This chapter discusses a number of issues concerning the role of information in life and society. In respect to life, these issues include the substantive role and the transmission sense of information in biology, the natural information principles, and biocomputation. In respect to society, we discuss IT applications in office automation, power generation and distribution, computer-integrated manufacturing

[1]S. Dutta and I. Mia (Insead, World Economic Forum: Mobility in a Networked).

(CIM), business and electronic commerce, education, medicine, and transportation. Finally, the issues of social networking and ethics of IT are addressed.

## 11.2 Information and Life

### *11.2.1 General Issues*

As we discussed in Chap. 1, the instructions for living beings are contained in the cell *genome,* which consists of DNA molecules made up of sequences of nucleotides. Particular parts of DNA, the *genes*, contain the information the cell uses to make proteins. The formation of cells signaled the initiation of biological evolution. The relationship and interplay of information and life systems has been thoroughly studied by a large number of researchers and research groups. This research is still continuing. Some of the most critical ones are documented in the Ref. [1–32].

In [7], Bajic and Wee provide a comprehensive treatment of information processing as applied to living organisms with the ultimate goal to develop the tools for better understanding of "*artificial*" and "*natural*" processing of biological information. Biocomputing topics at the molecular, cellular, systemic, and evolutionary levels are considered along with integrated prognostic profiles, analysis of DNA sequences, and dynamic biological databases and web sources.

Collier [8] reviews some of the ways in which information concepts have been used in biology. He distinguishes *instrumental* from *substantive* use of information in biological studies. He claims that "substantive use of information employs information in an explanatory way, in addition to any representational instruments," and he focuses on the association of information with heredity which was previously studied by Weissman (1904) and Lorenz [10]. Today, all well-known biologists do not object the use of information concepts in biology in connection with genetics. This point of view has received increasing attention during the last two or three decades.

Bergstrom and Rosvall [11] propose the *transmission sense* of information in genetics and evolutionary biology, which is an alternative to the *causal sense* based on Shannon's theory and the *semantic sense* which is deliberately omitted from Shannon's theory. In the transmission sense, "an object X conveys information if the function X is to reduce, by virtue of its sequence properties, uncertainty on the part of an agent who observes X." The transmission sense captures much of what biologists mean when talking about information in genes and also reinserts Shannon's theory into biology, in a well-justified way.

The thesis that biological evolution via natural selection and the development of human knowledge have a common underlying foundation was first proposed by Campell (1960) [12], and further developed by Popper (1979) [13]. More recently, work in this area has revealed the relations between behavioral, cultural and

biological evolution [14]. However, although the above work implicitly recognizes that natural information processing systems (e.g., biological evolution, human cognition) organize the information that drives the activities of living entities, it does not provide an explicit analysis. This gap was filled in by Sweller (2006) [15], who suggested that "both human cognition (when dealing with certain knowledge examples) and evolution by natural selection provide instances of natural information processing systems." He thoroughly explained how such systems can be specified by some basic principles that detail the system's mechanisms.

In the following, we briefly review the fundamental works described in [8, 11, 15], and discuss the so-called *biocomputation* field along the lines of [4, 32].

### *11.2.2 Substantive Role of Information in Biology*

The *substantive* role of information was first assigned by Maynard Smith and Szathmary [16] using information explanatorily. Although Shannon's statistical approach to information, which deals with averages, has been extensively applied to biological systems, the *combinatorial* and *algorithmic* approaches that deal with individuals are preferable. These two later approaches have been used to develop the *minimum description length* (MDL) and *minimum message length* (MML) methods for measuring the information in DNA and other biological entities [17, 18].

Collier [19, 20] provides a nested classification of the ways the substantive role of information has been used in the sciences. This hierarchical classification can be represented as shown in Fig. 11.1, in which each level inherits the logical and ontological commitments of the containing views, but adds further restrictions, as explained below.

**IT from bit** This view of information is objective and the most liberal and is based on Leibnitz's prospect that the world has a logical structure of perceptions derived from the ability to discriminate.

**Negentropy** This is a restriction of the "it from bit" view and is based on Schrödinger's and Szillard's prospects. From the "its", only the ones that can do

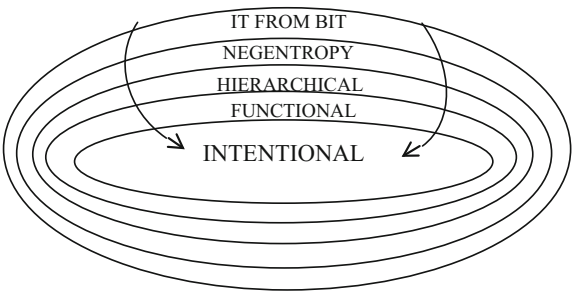



**Fig. 11.1** Nested inheritance of the major types of information

work (e.g., sorting things, using energy) are regarded as information. The others are disorder.

**Hierarchical entropy** This is a restriction of the negentropy view to particular levels of a physical hierarchy, indicating that not all negentropy is expressed at each level, and that the available "its" are level relative. To distinguish this expressed information from other forms of information, Collier has called it *enformation.*

**Functional information** This is the expressed information that is *functional* (i.e., it has both *syntax* and *semantics*) and comes from the outside. This means that it is not required that the information is information for the system itself.

**Intentional information** This lies within the scope of meaningful information, and is also known as *cognitive content*. The next level of restriction is *social information*, which is connected to various forms of biological information.

Next, Collier discussed the role of information in biology as a tool for calculating or estimating the information content in biological structures, from macromolecules to entire organisms and ecosystems. Also, communication systems theory can be used to analyze several biological channels (e.g., sensory processes, molecular communication, neural communication, intraspecies and interspecies communication, and ecological systems) in terms of their capacity, connectivity, order, and organization. DNA codes exist for proteins, regulation, and several phenotypic features from chemical networks in the body to social phenomena. Of course, none of these codes are straightforward, not even the mapping of DNA onto proteins. According to Maynard Smith and Szathmary, the major transitions in evolution are [16] as follows:

- Replicating molecules → Population of molecules in compartments
- Independent replicators → Chromosomes (linked replicators)
- RNA as gene and enzyme → DNA + protein (genetic code)
- Prokaryotes → Eukaryotes
- Asexual clones → Sexual populations
- Protists → Animals, plants, fungi (cell differentiation)
- Solitary individuals → Colonies (nonreproductive castes)
- Primate societies → Human societies.

The first four transitions are common to all transitions, which means that there are some general underlying principles. Collier argues that the third transition enhances significantly the role for *substantive information* by decoupling (separating) the roles of energy and information budgets in prebiotic and living systems, thus allowing the use of semantic information in biological systems. This decoupling of information and energy budgets permits self-organization within the information system itself [19, 20] and implies the existence of information channels, particularly channels from DNA to phenotypic traits. The DNA code ensures that no ambiguity exists in the regularities underlying information channels involved in gene expression and increases the fidelity of reproduction even in nonsexual organisms. The advantage of the information approach based on

substantive roles of information is that it can yield explanations, even across nonreducible hierarchical levels [20]. Collier, in [8], explained that he named the hereditary processes after Maynard Smith's transition—three *information processes* because they involve storage and transmission. He also indicates that much of what he said about genetic information applies *mutatis mutandis* to other forms of biological information, such as molecular communication, nervous system communication, immune systems, hormones, and behavioral transmission between organisms.

### *11.2.3 The Transmission Sense of Information in Biology*

In some areas of biology, the usability and usefulness of the information concept, e.g., transmission information signals by nerves, goes unchallenged. But when geneticists and evolutionary biologists use the language of information, some of them and many philosophers worry about whether this language is anything more than a "facile difference" [21]. Bergstrom and Rosvall [11] explained and showed that philosophers are wrong in their view that Shannon's theory is only useful for developing a shallow notion of correlation, the so-called "*causal sense*" of information or that in genetics and evolutionary biology, the information is used in a pure "*semantic view*" not addressed by Shannon's information theory. Bergstrom and Rosvall base their reasoning on the introduction of a new approach, which they call "*the transmission sense of information*". Their viewpoint is that, by focusing on the decision problem of how information is to be packaged for transport (an issue faced by communication engineers), many problems, encountered when applying the information concept to biology, are overcome. Adopting this point of view, several important features of the way information is encoded, stored, and transmitted as genetic sequences, are revealed and highlighted.

To make this clear, the three views of information in evolutionary biology and genetics are outlined. These views are [11]

- Causal view
- Semantic view
- Transmission view.

**Causal view** In this view (sense), the key statistical measure of the mutual information between two random variables X and Y is employed, i.e.,

$$I(X, Y) = H(X) - H(X|Y)$$

This measures how much we learn about the value of X by knowing Y. Information is conveyed (i.e., Y caries information about X) whenever Y is correlated with X. In this sense of information, *genes* carry information about *phenotypes*, but no particular details are provided, i.e., the information concept is used as shorthand of correlations. This is a shallow use of the information concept in

biology, in contrast to what communication engineers do in their field. Suppose that by "**G**" ("genotype") *has information about* "**P**" ("phenotype"), we mean only the mutual information I(G;P). Then, because of the symmetry, I(G;P) = I(P;G), the amount of information the genotype G provides about the phenotype P is always exactly equal to the information provided about the genotype by knowing the phenotype, i.e.,

$$\mathrm{I}(\mathrm{G};\mathrm{P})\ \text{Causal Information}\ \mathrm{I}(\mathrm{P};\mathrm{G})$$

This view leads to strange conclusions. For example, assume that the genotype to phenotype map is *degenerate*, i.e., there are nk possible genotypes but only n possible phenotypes. Using the above causal (mutual information) sense, we cannot measure the fact that, after making our observations, there are k possibilities (from G to P), but only one possibility in the reverse direction (from P to G). This is because only a few tools of Shannon's theory were used and the "*raison d'etre*" (reason for existence), i.e., the underlying decision of how to package information for transport, is neglected. In addition, the causal view of information cannot capture and reveal the intentional and representational nature of genes and other biological entities [22, 23].

**Semantic view** In this view, "*genes semantically specify their normal products.*" Biologists speak about genes as informational molecules, not because genes are correlated with other objects (e.g., amino acid sequence or phenotype) but because they represent other things. This is the *semantic sense* of information, which cannot be captured by Shannon information theory that does not deal with semantic aspects. Thus, the question is "what genes are supposed to represent?" Maynard Smith [24] suggests, as an obvious candidate, the phenotypes. Genes have a representational message about phenotype, but environment has not. However, as Godfrey Smith [22] and Griffiths [25, 26] argue, the parity thesis challenges the semantic view of information, which does cover all issues of information in biology.

**Transmission view** This view [11] is based on Shannon's *data compression* theory which has led to the practical theory of *coding*. Information transmission takes place when the information is properly packaged, and then it is sent from one place to the other (space dimension) or across the time dimension (from one time to a later time) via storage and retrieval (e.g., via a CD). In the causal view, we have simply a correlation, and in the semantic view, we have a translation. There is neither a space dimension nor a time dimension in which transmission of information takes place. In biology, genetic transmission occurs along the time dimension from parent to offspring to grand offspring as shown in Fig. 11.2.

This transmission of genetic information includes all aspects of life, e.g., metabolizing sugars, create cell walls, etc. Taking the transmission of information view, according to which "an object X conveys information if the function of X is to reduce, by virtue of its sequence properties, uncertainty on the part on an agent who observes X," we can see information as it flows through the process of

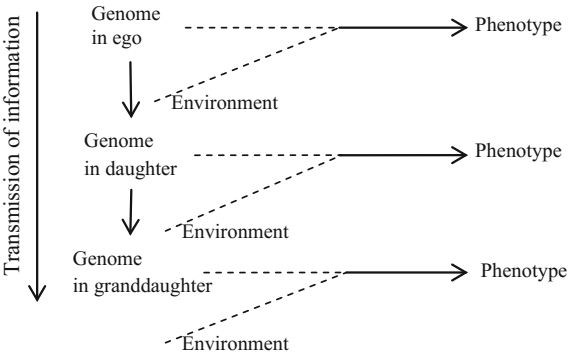



**Fig. 11.2** Transmission of information in biology from parent to offspring

intergenerational genetic transmission. We can also think about natural selection, the evolutionary process, and how information gets into the genome in the first place. The transmission sense of information is not restricted to the fact that *genes* have some special property not possessed by any other biological entity, but goes to the identification of those biological components that have information storage (memory), transmission and retrieval as the primary functions. DNA has actually the properties of storage and transmission. The variety of information sent by the parent to each offspring and the action of natural selection on the phenotype facilitate the accumulation of information on the genome.

Bergström and Rosvall conclude that the transmission sense of information justifies "information theory" as more than a shallow metaphor. Correlations, mutual information symmetry, the parity thesis, and coding, all come into focus via the communication systems approach.

## 11.3 Natural Information Processing Principles

Human knowledge is distinguished in biologically *primary* and biologically *secondary* knowledge. Primary knowledge includes all types of information that we acquire and use naturally because of our evolution (e.g., learning to listen and speak a native language, learning to interact socially, and learning to use general problem-solving strategies).

Secondary knowledge refers to classes of knowledge that are not the result of evolution because they have only become culturally important fairly recently. Learning to read and write belong to secondary knowledge because reading and writing is quite different from the moment in which the skills of listening and speaking are acquired.

Primary knowledge is acquired easily, rapidly, and unconsciously, i.e., biologically and automatically. On the contrary, for most individuals, secondary knowledge will not be acquired without purposeful assistance by social mechanisms.

According to John Sweller, natural information processing systems, which direct the activities of living organisms, can be studied through the following five basic principles that apply to both *human cognition* and *evolution* by *natural selection* [15]:

- The information store principle
- The borrowing and reorganizing principle
- The randomness as genesis principle
- The narrow limits of change principle
- The environmental organizing and linking principle.

A brief description of these principles follows.

### *11.3.1 The Information Store Principle*

The natural information store has an enormous size which is needed to store the vast variety of conditions faced in the very complex environment in which the humans live. The contents of long-term memory provide the storage for the information for human cognition, which means that the human cognitive activity is directly determined by long-term memory. This activity includes, for example, what we see, hear, and think. Primary knowledge, on the other hand, can be used to obtain and memorize large amounts of *biologically secondary knowledge* (**BSK**). In the same way, evolution by natural selection relies equally on the information storage principle. For example, in genetics, the production of protein is determined by organized information which is stored in the genome. In cognition, if no change occurs in long-term memory, no learning takes place. Similarly, if there is no change in a species' genome, no evolution takes place. Evolution means genomic change. A genome is a large information storage that lies at the heart of natural information systems.

### *11.3.2 The Borrowing and Reorganizing Principle*

This principle provides the mechanism for acquiring the large amounts of information to be stored in long-term memory. Nearly all of the semantic information that an individual has in his/her long-term memory had been borrowed from the long-term memory of other individuals. Human imitation of the way other people read and write is the basic mechanism that involves the function of transferring information from the long-term memory of one person to that of another person. Books or electronic storage are actually intermediate in this information transfer process from one person to another person. Previous information is combined with new information by means of a *reorganization* process, usually implemented

through *schemas* [27]. A schema can be used for classifying multiple information elements according to their use. The importance of the information borrowing and reorganizing principle has been verified by a series of "worked examples" experiments. Novices who were presented worked (solved) examples to study, instead of the problems under solution themselves, were better able to solve subsequent new problems [28]. Worked examples reduce or eliminate random problem- solving parameters.

Examples of genetic processes that use the borrowing and reorganizing principle are [15] as follows:

- Splicing of RNA coding in alternative ways. This rearranges the order of the coding sections and thereby coding for a different protein. As a result of splicing, one gene can generate more than one protein.
- Viruses can reproduce only inside the hosting cells borrowing and using their genetic machinery.
- Rearranging DNA stretches when active sections of the DNA move from one location to another location inside the genome, thus altering the output of various genes.

In conclusion, both human cognition and biological evolution are structured to reorganize the borrowed information at the time it is borrowed, test the effectiveness of the organization, and accept or reject it depending on the result of the test.

### *11.3.3 Randomness as a Genesis Principle*

This principle refers to the ability of humans and other natural systems to create new information through random search. The method mostly used by humans for creating new information (i.e., for solving problems) is the *means-ends strategy* [29] (see Sect. 5.3.1.6). In this strategy, a *problem solver* considers repeatedly the current state of the problem, compares it with the goal state, and finds an operator that reduces the difference. Successive problem-solving solutions can generate new knowledge which can be stored in the long-term memory. Of course, to do this, the problem solver must either have knowledge in the long-term memory that suggests which problem-solving operators may reduce the differences, or have access to the relevant knowledge of someone else's long-term memory.

Evolution by natural selection uses a similar process for producing new information. Because mutations are random (possibly nonadaptive) and may lead to dead ends, the random generation (search) must be followed by effectiveness tests.

All living organisms face the problems of survival and reproduction in a given environment. If mutation leads to increasing offspring (i.e., it is successful), it is inherited; otherwise, if it is not successful, it is rejected. This process of mutation follows the genesis by randomness principle and is the basis for the genesis of all biological variation.

### 11.3.4 The Narrow Limits of Change Principle

This principle refers to the necessity of reducing the extent of random generation because all real systems have time limitations for creating the new knowledge (i.e., for solving the problems). A time-constrained system cannot operate if it has to handle more than a few elements needing to be randomly combined. Actually, the probability of finding appropriate combinations for more than a few elements via random generation is very small. Although the borrowing and reorganizing principle employs already-organized knowledge (thus reducing the combinatorial explosion effect), it is not sufficient because learning always involves a random component. It is the "narrow-limits-of-change principle" that reduces the extent of the untested information's entrance into long-term memory.

The "narrow-limits-of-change principle" holds also in biological evolution, where rapid random changes to a genome are, similarly, not successfully operational. Only small genetic changes over many generations (tested for effectiveness) can become well established. Because of this, mutation is rare and ensures a very small rate during human DNA replication. Thus, the low mutation rate limits the rate of evolutionary change in the same way as a limited working memory limits the changes in long-term memory. Here, the interactions between the environment and the sequence of bases in the DNA are not controlled by the DNA-based system, but by the *epigenetic system* [30].

### 11.3.5 The Environmental Organizing and Linking Principle

This principle refers to the fact that knowledge in long-term memory provides a central executive that indicates what must be attended, how information should be processed, and what actions should be taken, and also organizes the environmental information that has to be processed. Specifically, working memory, in the form of long-term memory, allows large quantities of previously organized information from long-term memory to be suitably employed in order to organize human interactions with the environment [31]. This means that the working memory (or long-term working memory) provides an unlimited link between the long-term memory and the environment.

The same principle also applies to evolution by natural selection. This is accomplished through the interaction of *the epigenetic system* with the *genetic system*, which is manifested in two ways as follows:

- As a response system responding to changes outside the DNA strand
- As a transmission system that directs changes in the interpretation of the DNA strand.

These two ways are analogous to the ways the working memory operates, i.e.,

- Handle environment information and evaluate it for long-term memory
- Use the information available in the long-term memory to drive the interactions with the environment.

In conclusion, both the *working memory* and the *epigenetic system* act as an intermediate between the information storage and the environment.

## 11.4 Biocomputation

*Biological computation* or, briefly, *biocomputation* is the term coined to catch all the research work performed at the interface of biology and computation. However, this term has been used in so many diverse ways that confusion is in the air. Hickman, Darema, and Adrion [32] provided a classification of four areas of biocomputation which can contribute toward a convenient organization of biocomputation. These areas are as follows:

- **Biomolecular computation:** This area includes the exploitation of biological macromolecules for implementing standard computation methods. Examples are DNA computing and storage media using the bacteria rhodopsin.
- **Computational biology:** This area includes the efforts to solve biological problems via computational techniques and tools to model biology and handle the complex mathematical expressions of biological phenomena. Examples are calculation from first principles (e.g., Ab Initio), Monte Carlo methods, and other simulation methods used to study protein folding and protein–protein interactions.
- **Bioinformatics:** This area involves data mining, data modeling, data management, and other methods to handle biological databases (e.g., genome databases). Examples are data mining for determining sequence homology information, in silico *models* as a predictive method.
- **Biological computation:** This area is concerned with how biology performs information technology operations from the subcellular up to the population level. Examples include hybrid systems to reverse engineer biology, and techniques to study biological systems at the multicellular level and beyond.

Kari [4] is particularly concerned with *DNA computing* and introduces the term *biological mathematics* (biomathematics) on the basis of the observation that the following two processes, one biological and one mathematical, are analogous:

- The extremely complex structures of living beings are the outcome of applying simple operations (e.g., copying, splicing, etc.) to initial information encoded in a DNA sequence.

- The application of a computable function "*f*" to an argument "*w*", i.e., the computation of *f*(*w*) can be found via the application of elementary (simple) functions to *w*.

According to Adleman [33], DNA strings can be used to encode information while enzymes can be employed to simulate simple computations. A single strand of DNA can be linked to a string formed by the combination of four different symbols A, G, C, and T, corresponding to the four bases, adenine, guanine, cytosine, and thymine. From a mathematical point of view, this means that we have available a four-letter alphabet $\Sigma = \{A,G,C,T\}$ to encode information, which is much richer than the two-digit coding used in digital computers.

Because the interactions between DNA and enzymes involve a very high amount of parallelism, DNA and enzymes (i.e., DNA computers) will probably soon outperform the fastest supercomputers. Standard computers are limited by the number of processors built into the hardware. On the other hand, a test tube filled with DNA fragments and enzymes has all the molecules exposed to each other at once. Due to this, DNA has the capability of rapidly searching, sorting, and manipulating massive amounts of data in parallel. Adleman devised his DNA computer to perform the same basic algorithm as standard silicon chip-based computers. The problem of current DNA models is that DNA is relatively unstable compared to a hardware processor.

In [4], the issue whether DNA computing can be used to compute everything is investigated. Specifically, the following questions are addressed:

- Is the DNA computation model computationally complete? In other words, can the computation required of any Turing machine be performed by DNA manipulation?
- Does there exist a universal DNA system, i.e., a system that, given the encoding of a computable function as an input, can simulate the action of that function for any argument? In other words: Is it possible to design, at least in principle, a programmable DNA computer?

The opinions on the practical importance of these questions vary considerably, especially because the issue is not to try to fit the DNA model into the Procrustean bed of classical computation, but rather to try to completely rethink the concept of computation. These and other issues of biomolecular (DNA) computation are extensively discussed in [4], where mathematical models of splicing, etc., are also provided, However, independently of the above discussion, Adleman's DNA computing is expanding rapidly, many NP-complete problems are being satisfactorily tackled, and several DNA computers are being developed (e.g., [34, 35]).

## 11.5 Information and Society: Introduction

As we have already mentioned, human is society presently in the middle of the *information revolution* which includes the advancements in the movement of data, telecommunications, wireless communications, and the Internet. Present society is justifiably called the *information society*. The members of this society are very often called *digital citizens*, to reflect the fact that the various activities of society are technologically implemented in a digital {0, 1} way.

A representative but not exhaustive set of activities and processes in modern society that interact with information technology (IT) is the following:

- Culture
- Education
- Production
- Business
- Commerce
- Transportation
- Medicine
- Tourism
- Politics.

In any information-based system, the following players are involved:

- People: individuals and society
- Processes and procedures
- Computer hardware
- Computer software
- Communication systems
- Other technology and equipment.

Today, there is a shift from an economy based on material goods to one based on knowledge, which is called the *information economy* [36]. According to OECD, a society can be characterized as an *information society* if more than 50% of its *gross national product* (GNP) is produced and more than 50% of the employees are involved in information processing activities in their jobs. Actually, a postindustrial society is based on services. According to Daniel Bell, the informational character of a society is specified by the number of employees producing services, i.e., nontangible goods [37]. A society in which the majority of workers are not working in the production of tangible goods is characterized as a *post-industrial society*.

The information (or knowledge) society has available and processes large quantities of data and knowledge. It has adapted its lifestyle to this fact and, consequently, now depends on data (information and knowledge). Thus, an information society becomes more and more demanding of information technology products and wants the production cycles of goods and services to be shortened and the costs to be lowered. People outside the information society do not have all these demands. But information technology should also be beneficial to them, in order for them to remain willing to contribute towards its development.

To participate effectively in the production and economic activities of an information society, employees should know or be trained in information-knowledge-oriented applications. Here is where much attention and care should be given. According to Thomas Davenport, a pioneer thinker in business process reengineering and knowledge management, this issue is not properly faced. Specifically, in an interview (2005) [38], he revealed that: "Even when people are trained on knowledge-oriented applications, such as Excel, PowerPoint, CAD or CRM, the training focuses on how the software package works, not on how it fits into the context of the job. The vast majority of organizations that implemented CRM didn't really help their salespeople figure out how to use the system effectively to help them sell better."

Davenport has pointed out a recurring situation in most professional fields where software systems and tools are used for the purpose of improving professional activity. This happens in many fields including management, industry, and education. Davenport said: "Most organizations have no training or education on how to use these tools effectively in their work."

It is obvious that, to do such training, someone must know how an operation of software is meaningfully connected to a specific task. Even if the "end user" is exempted from knowing this, someone along the chain (system developer, users' trainer, etc.) must understand this concept and have a clear and explicit definition of which and how a program can fit into the context of jobs of the workers in that field. The users of information technology must become technology-enhanced knowledge workers, which can only occur if such a feature is included in their education and training.

Some books on information society and the information revolution are [39–45]. In [39], a comparative perspective on information societies and a scenario for the information society of tomorrow are provided, including the issue of privacy in an information society. In [40], the question what macroeconomic data do and do not show about the impact of information technology on service sector productivity is considered, and the ways in which different service companies have implemented information theory are assessed. In [43], a critical perspective on information politics in the information age, i.e., the way social and technical choices about information and communication technologies (**ICTs**) influence access to information, people, services, and technologies themselves is given. The author coined the term "*the shaping of tele-access*", and showed how this concept challenges prevailing theoretical perspectives on the information and communication revolution.

In [45], the works and opinions of eight key *thinkers* about information society are presented. These theoretical thinkers share the view that *information and communication technologies* have produced a revolution, a remaking of the world, and that there are sufficient analytical tools available to understand the current status of the information society, which is no longer organized on the basis of material goods, but on intangible goods. A hint on the point of view of each of them (as presented by the contributors of the respective chapter of [45]) is as follows:

- **Walter Benjamin** (presented by Marianne Franklin): Benjamin's thesis is that "During long periods of history, the mode of human sense perception changes with humanity entire mode of existence … All manner of financial transactions, manufacturing process, and service industries would be hard put to function these days without information and communication technologies (ICTS)."
- **Murray Edelman** (presented by Andrew Chadwick): Edelman states that: "For most men of the time, politics is a series of the mind, placed there by television news, newspapers, magazines, and discussions… Because politics does visibly confer wealth, take life, imprison and free people, and represents a history with strong emotional and ideological associations, its processes become easy objects upon which to displace private emotions, especially strong anxieties and hopes."
- **Jacques Ellul** (presented by Karim H. Karim): His views on propaganda and myth have significant potential for understanding the ideological promotion of information society.
- **Harold Innis** (presented by Edward Comor): His concerns lay in the thought processes through which people of different civilizations define their vision of reality … His focus is less on the individual than on the character of the society that produces individuals and either releases or suppresses their creative potential (Cox 1995)
- **Lewis Mumford** His thesis, as summarized by Christopher May, is that the "the inventors of computers are the pyramid builders of our own age: psychologically inflated by a similar myth of unqualified boasting through their science of their increasing omnipotence, if not omniscience, moved by obsessions and compulsions not less irrational than those of earlier absolute systems: particularly the notion that the system itself must be expanded, at whatever eventual cost to life."
- **Karl Polanyi** (presented by Kenneth S. Rogerson): His work focused on the Industrial Revolution. He explained thoroughly how economic relations affect individuals and societal groups. He considered himself an economic anthropologist.
- **Elemer Eric Schattschneider** (presented by Robin Brown): His work shows how informationalization changes political life, and provides a clear way of understanding how politics is changed in a more transparent (information) society.
- **Raymond Williams** (presented by Des Freedman): His books on "Culture and Society" (1958) and "The Long Revolution" (1961) opened up "an anti-elitistic approach to culture that emphasised the expressive contributions made by those traditionally written out of cultural history: the poor and the exploited." He challenged the thesis that culture was an elite enjoyment referring only to fine arts and insisted instead that culture emerges out of the soil of everyday life. Hence, the study of culture required anthropological as much aesthetic … sensitivity to the history, traditions, and daily practices of working people.

The impact of information and communication technology (**ICT**) on our society is multifaceted and touches all aspects of our everyday life. Thus, here we will focus our discussion on a few, but very important, applications of ICT, namely,

- Office automation
- Power generation and distribution
- Computer-integrated manufacturing
- Business
- Education
- Medicine
- Transportation.

## 11.6 Information Technology in Office Automation

According to "webpedia" [46], the term *office automation* refers to the use of computer systems to execute a variety of office operations, such as word processing, accounting, and e-mail, and it almost always implies a network of computers with a variety of available programs. In other words, *office automation systems* (**OAS**) are information systems that are capable to handle, process, transfer, and distribute all the data/information involved in the operation of an institute, company, or enterprise. Office automation aims at providing ICT elements that make it possible to facilitate, simplify, improve, and automate the organization of the activities and processes of a company or a group of people, such as administrative data management, synchronization of meetings, cooperation between various departments of the company, etc. To achieve its goal, office automation involves and uses the following:

**Functions**

- Acquisition
- Registration
- Storage
- Searching
- Scheduling.

**Components**

- Computer systems (hardware/software)
- Communication network(s)
- Procedures
- Human decisions and actions.

**Technologies**

- Data processing
- Word and text processing
- Image processing
- Voice processing
- Communications processing.

Some currently available suites are Apple Works, IBM/Lotus Smart Suite, Microsoft Office, Corel Word Perfect, Sun Start Office, and Open Office (freeware).

Figure 11.3 shows the office workflow system architecture of an office-automation example.

The range of the Internet uses in office automation-related applications includes the following [47]:

(i) Government
(ii) Health care
(iii) Education
(iv) Enterprise
(v) Business
(vi) Stock market
(vii) Online operations
(viii) E-commerce
(ix) Databases
(x) E-mail.

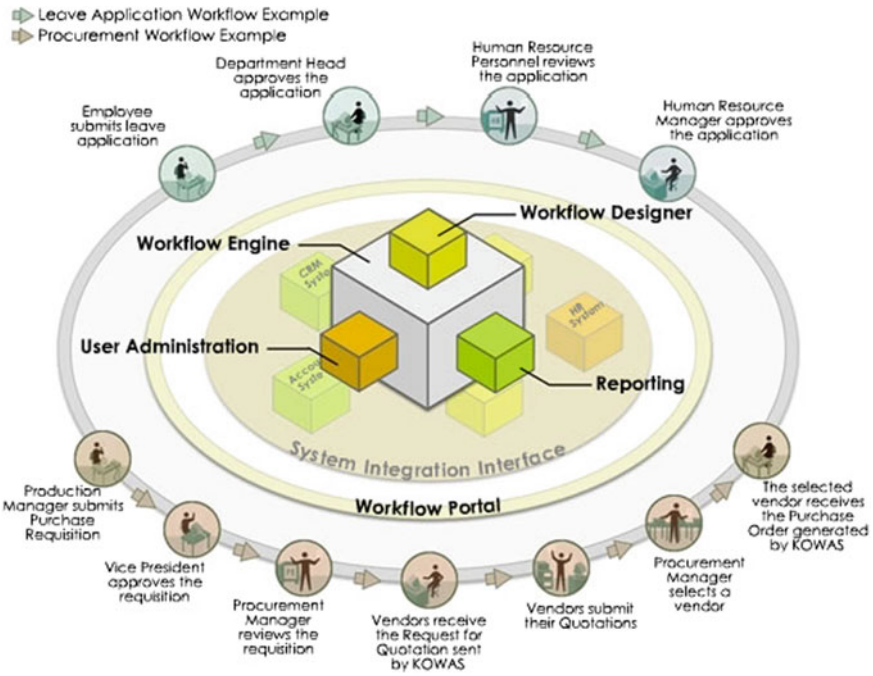



**Fig. 11.3** An application example of office automation (leave application and procurement workflow). [http://www.kbquest.com/user_uploaded/image/kowas_architecture.jpg (The reader is informed that web figures and references were collected at the time of writing the book. Since some of them may no longer be valid due to change or removal by their creators, they may no longer be available.)]

## 11.7 Computer-Based Power Generation and Distribution

*Electric power utilities* involve two principal functions

- Power generation
- Power distribution.

**Power generation:** Today's electric power systems are very large and complex and can no more be controlled and manually operated by analog control loops. This type of control is costly, unreliable, and sometimes hazardous. For this reason, researchers, developers, and companies have designed, implemented, tested, and put on the market high-level and high-precision computer-based control and monitoring systems, which are flexible and adaptable to various situations and power levels. These systems provide fault-free or at least fault-tolerant operational capabilities, properties that are critical, given the huge investments needed for modern boilers and turbine generators. Figure 11.4 gives, in block diagram form, the components of a computer-based electric power generating system.

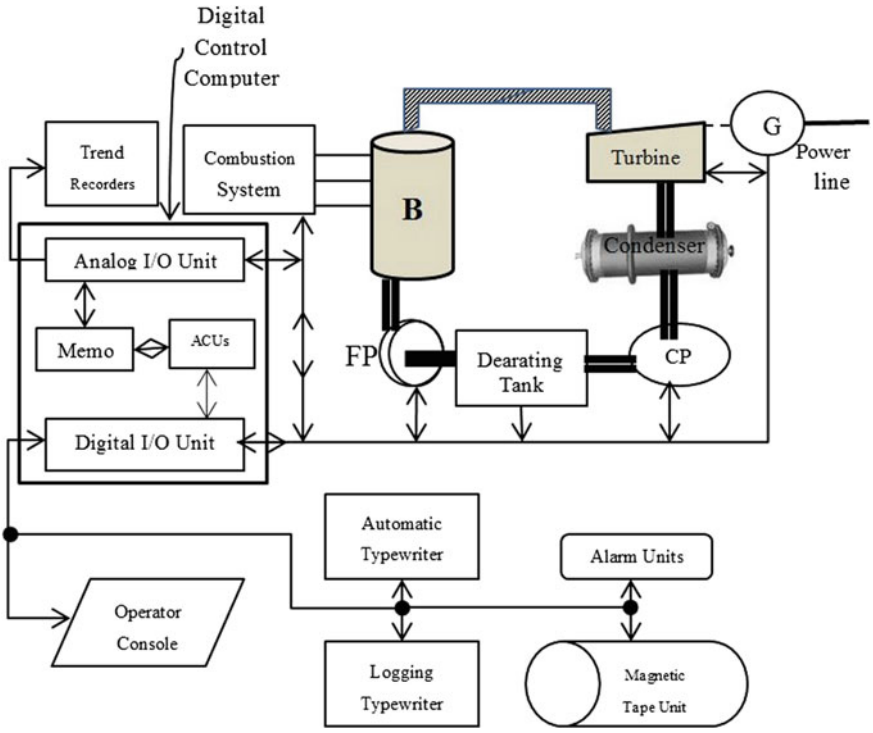



**Fig. 11.4** General architecture of a digital controller/computer (ACU = Arithmetic and control unit, B = boiler, G = generator, FP = Feed pump, CP = condensate pump) for a steam-generating power system

The control computer should be able to receive all data from the steam-generating plant, carry out accurate computations, perform the suitable logical operations, and provide control signals that guarantee the smooth, safe, and economic operation of the entire system.

**Power distribution** An electric power distribution system is a necessary part of electric power utilities for delivering electricity to consumers. Reliable power distribution and delivery is a very important requirement for the economic growth and development of a country. A modern electric power utility should be able to perform on all the days of the year and the hours of a day with a steady and uninterrupted power supply, even during the high-demand hours, so as to satisfy its consumers. Therefore, due to the large size of the distribution networks, the system must have, in addition to the controller of power generation, a suitable computer-based monitoring and distribution controller. According to IEEE, a system of this type is defined as "a system that enables an electric company to remotely monitor, coordinate, and operate distribution components in a real-time mode from remote location" [48].

The control decisions are made at the so-called "*distribution control center* (**DCC**)" which involves several application software programs that cooperate to achieve the task (see Fig. 11.5) [49, 50].

The power distribution system needs, in addition to the DCC several other subsystems, as shown in Fig. 11.6, namely,

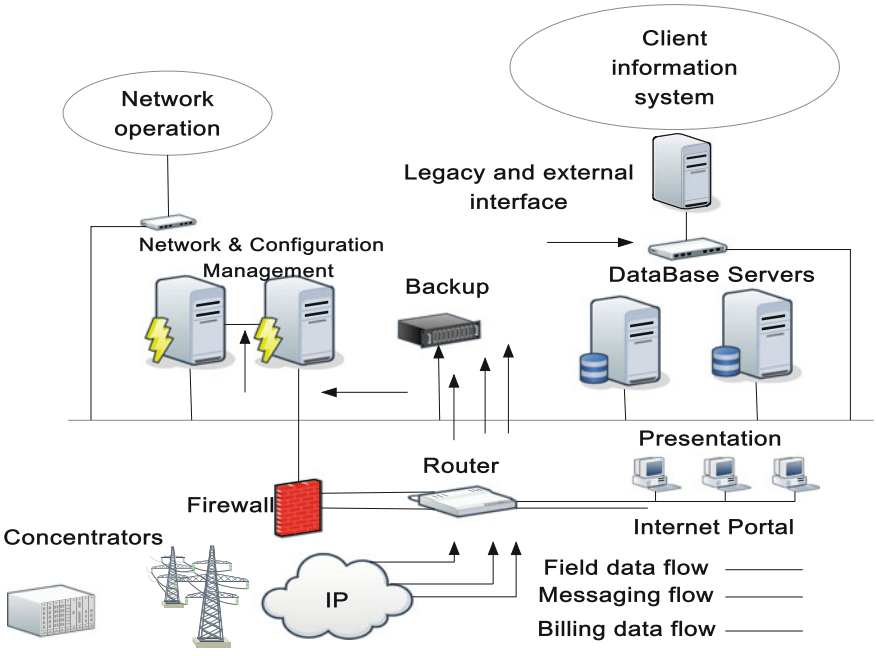



**Fig. 11.5** Architecture of a digital control power distribution center

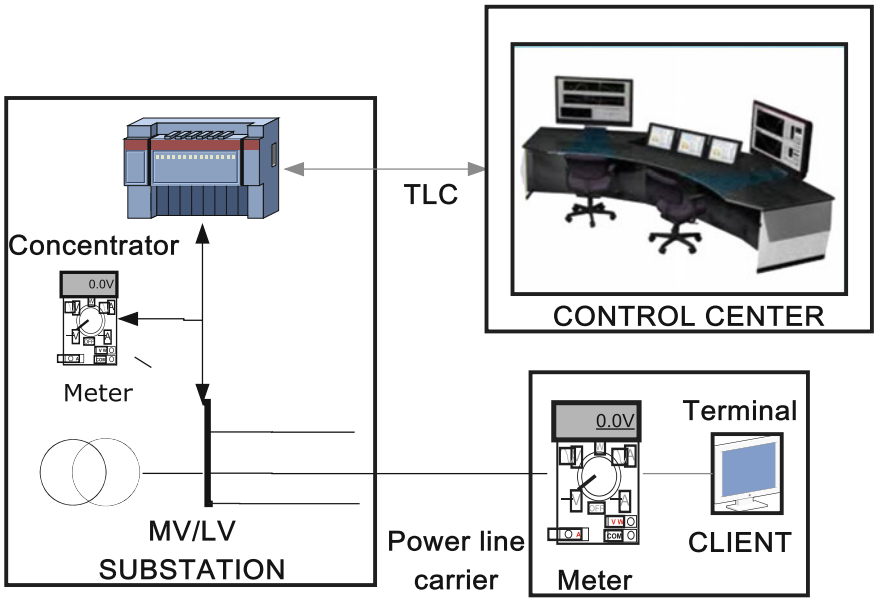



**Fig. 11.6** General structure of power distribution systems

- An automatic meter reader (*AMR*)
- A data concentrator unit (*DCU*)
- A supervisory control and data acquisition (*SCADA*) system
- Communication units.

The DCC is the primary subsystem of the power distribution system and must be scalable and amenable to further enhancements. To this end, power generation and distribution utilities use several "open-hardware" and "open-software" commercial components (e.g., operator workstations, a management server (with a back) for housing the software application and the alarm server, a data collection server, a router, etc.).

Today, the trend is to move from vertically integrated electric power systems towards unbundled model of *generation companies* (*GENCOs*), *transmission companies* (*TRANSCOs*), *distribution companies* (*DISCOs*), and *energy service companies* (*ESCOs*). Previously, all electric power distribution-related functions could be transparently coordinated along the complete supply chain. Today many power distribution companies manage third-party contacts by delivering bulk power from GENCOs and TRANSCOs to meters owned by ESCOs [51].

To make distribution automation more intelligent and cost-effective and to accomplish the goal of full-scale unbundling of power systems, attention is focused, among others, on the following issues:

- Communication protocols with interoperability capabilities
- Communication units that make the power system commercially viable
- Intelligent remote terminal units (RTUs)
- Algorithms that secure accurate control
- Intelligent instrumentation system.

A very important requirement for an electric power distribution system is to have the capability for restoration in cases emerging from faults in the network. Restoration of supply to the affected customer is a must and has to be done as quickly as possible. A solution to this problem can be found in [52].

## 11.8 Computer-Integrated Manufacturing

*Computer-integrated manufacturing* (**CIM**) is an advanced form of *automated manufacturing* that is concerned with the application of automation methods and tools to produce goods for the society in an automatic way. According to "Answers.com": "A computer-integrated manufacturing system is a computer-automated system in which the individual engineering, production, marketing, and support functions of a manufacturing enterprise are organized. Functional areas such as design, analysis, planning, purchasing, cost accounting, inventory control, and distribution are linked through the computer with floor functions such as materials handling and management, providing direct control and monitoring of all process operations."

Technologically, CIM is based on information technology and closed-loop methods and tools. Among the major components of CIM are the sensors (and smart sensors), actuators (prime movers), and real-time controllers. CIM is an enhancement of flexible manufacturing, in which the factory is able to adapt quickly to produce several different products, or where the product volumes can be changed quickly via computers to satisfy varying demand [53, 54]. A CIM system integrates the technical issues and the managerial issues of the production and involves the "human" (manager, engineer, operator) as a crucial component of the overall system, Today's demand for *discrete products* (machines, cars, very large integration circuits, refrigerators, air conditioning devices, etc.) is high and is very quickly increasing. Therefore, CIM systems and techniques based on information technology, management science, and control engineering play a crucial role for the human development, economic growth, and improvement of quality of life.

CIM systems involve at least two communicating computers, e.g., a supervisor computer, a robot arm control computer or a numerical machine (CNC) control computer, etc. Actually, CIM systems have a hierarchical structure with at least three levels, as shown on Fig. 11.7.

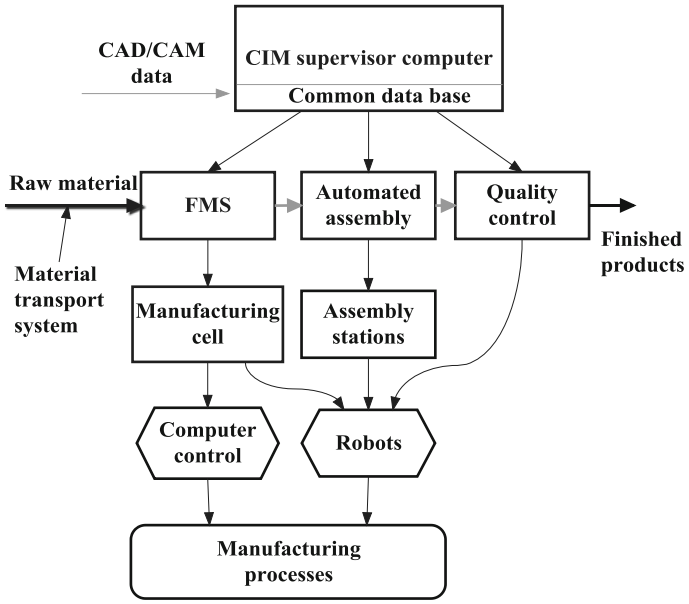



**Fig. 11.7** Typical hierarchical structure of a CIM system

At the lowest level, there are stand-alone computer controllers and industrial robots (manufacturing cells). The operation of several manufacturing cells is coordinated by the central (supervisor) controller via a materials handling system which constitutes the intermediate (or FMS) level of the CIM system.

According to Appleton [55], CIM involves three parts or points of view.

- *The demand for information*: the *user view* of CIM determined by the system's market environment
- *The supply of information*: the technology view of CIM, which is the outcome of pressures on the suppliers of technology
- *The enterprise view* of CIM: this provides a control structure that can maintain alignment between the dynamic user and technology views, and at the same time leads to the integration and consistency required by the enterprise as a whole.

The enterprise view of CIM involves

- Planning policies
- Project management procedures
- Data and technical standards
- Budgeting
- Performance testing and control.

Overall, the principal technical functions of CIM are as follows:

- Product design and design for assembly (DFA)
- CAD/CAM in terms of features
- Process planning, scheduling, and control
- Dynamic simulation of FMS
- Equipment selection
- Facility layout
- Fault detection/diagnosis/restoration
- Quality control and assurance.

The three primary challenges in developing a successful CIM system are as follows:

- Integration of components from different suppliers
- Integration of data, depending on the degree of automation
- Integration of process control with the human operation to secure smooth functioning of the overall system.

Devices and equipment used in CIM systems include CNC (computer numerically controlled) machines, PLCs (Programmable logic controllers), computers, robots, stand-by digital controllers, interfaces, monitoring devices, and computer networking.

The operation of modern high-level (intelligent) CIM systems is facilitated by artificial intelligence technologies and expert system tools [56] (see also Sect. 5.3.1.6). A framework specification that aims to assist in the creation of an integrated, common, flexible, modular object model leading to an open, multi-supplier CIM system environment is presented in [57]. Features that can be assured by following this framework include increased productivity, reduced cycle time and cost of new development, shorter factory start-up and ramp-up times, wider range of applications, and reduction in the meantime to repair (MTTR) of the system.

Figure 11.8 is a general diagram of the managerial operations performed in a CIM system, namely supply chain management, retail management, warehouse management, facilities management, customer relationship management, and financial functions.

Figure 11.9 shows an industrial robotic manufacturing cell in the car industry.

Figure 11.10 shows a view of an actual computer-integrated system involving a robot and a workpiece transfer unit.


**Fig. 11.8** Managerial and financial functions of a computer-integrated manufacturing system. (http://efficientcomputersystems.com/images/controlsoftware.gif)

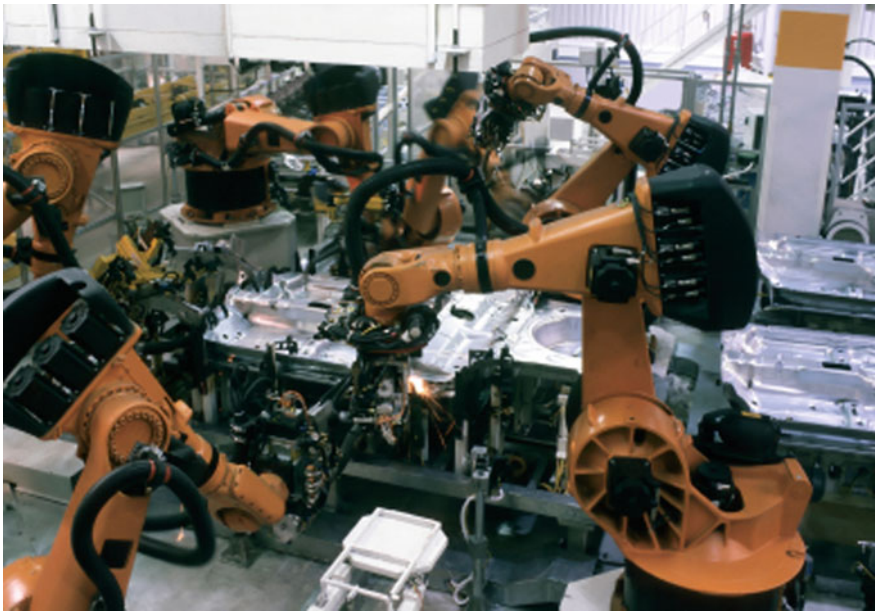


**Fig. 11.9** Industrial robots in a CIM factory. (http://www.ipofferings.com/drawings/5newpatentsforipofferings_com12/iStock_000009439397XSmall.jpg)

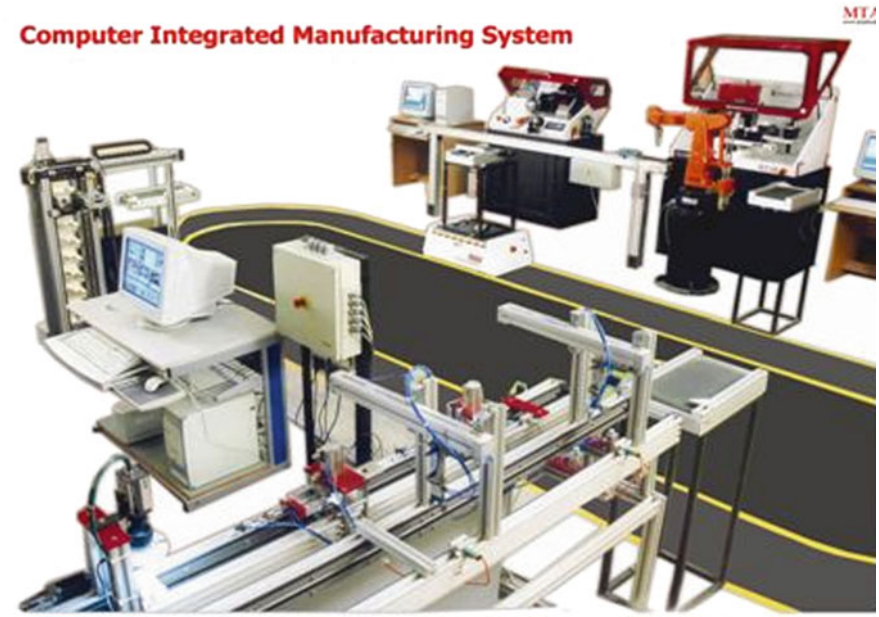



**Fig. 11.10** Example of the setup of a CIM system. (http://www.ogtte.com/images/prodimages/prod23.jpg)

## 11.9 Information Technology in Business: Electronic Commerce

The positive impact of information and communication technology (ICT) on business and management is very impressive. One of the first important ways this impact is manifested is the reduction of the critical role of distance. Many industries are changing their geographic distribution of work significantly. For example, European software companies have realized that can overcome the tight local market for software engineers by sending projects to Far Eastern countries where the wages are considerably lower. Companies can outsource their manufacturing to other countries relying on telecommunications to maintain marketing, research and development, and distribution teams in close cooperation with the manufacturing staff [58]. This implies that ICT helps to achieve a better division of labor among countries, which in turn influences the relative demand for various professional specialists in each nation. This also means that ICT enables various kinds of work and employment to be decoupled from one another. As a result, companies have now greater freedom in locating their business activities, creating stronger competition among regions in infrastructure, capital, labor, management, and other resource issues and choosing which tax system is most preferable.

The operation of the market has also changed using ICTs that enable 24-h access at low cost to almost any type of price and product information sought by the customers. One of the best examples to show the impact of ICT on companies and business is the *electronic commerce* (e-commerce), which allows a reduction in the cost of the sales activities, order placement and execution, customer support, staffing, inventory process, and distribution. This, together with the significant cost reductions, secures higher service quality with far fewer, but high-skilled, employees.

In addition, the faster the orders can be placed and delivered, the less the need for a large inventory. This reduction in inventory quantities is mostly beneficial to companies that produce products with a limited shelf life. Moreover, the distribution sector is directly affected by e-commerce which is a new process of supplying and delivering goods and services. The simultaneous developments in media and ICTs create a novel integrated supply chain for the production and delivery of information content of a multimedia nature. Of course, this affects the employment sectors as well by both eliminating and creating jobs.

The new type of doing business and the new style of management must be combined with information technology skills, especially front end applications with enterprise operations and applications. It has been widely recognized that computer networking, competitive websites, and complex data handling applications require skills at a level much higher than that needed by a platform-specific ICT job. Two books on the information technology impact on business and management education and organizational change are [59, 60]. A general discussion on the evolution and penetration of the Internet phenomenon and computer networking in modern life and society can be found in [61]. Particular methodological and technological issues on information systems (structure, operation, organization, development) were given in Sect. 5.4). A set of standards for the process of developing enterprise information systems is provided in [62].

## 11.10 Information Technology in Education

The influence of ICTs, including Internet and web-based multimedia, on e*ducation* is now visible in both developed and developing countries. The Internet offers, in most cases, complementary teaching and learning methods to the traditional face-to-face classroom instruction [63–65]. In science and engineering education programs, there are today available important results and platforms for e-course material with online theoretical and laboratory exercises, course management, virtual classes, etc., as well as virtual (simulation) laboratories and web-based remote physical laboratories. These web-based platforms can be classified into the following categories:

- E-course material and e-classroom environments
- Virtual laboratories

- Remote (physical) laboratories
- Combinations of the above.

Examples of such platforms in the control and robotics field are as follows [66]:

- Web-based study support environment for teaching automatic control [67]
- The automatic control telelab [68]
- Web-based control system design and analysis (WCDAS) [69]
- Recolab: A hybrid virtual/remote control laboratory [70]
- Internet-based laboratory for robotics and automation [71].

The general architecture of a virtual lab and remote lab is shown in Figs. 11.11 and 11.12, respectively.

In the virtual lab, the server station implements and controls the communication between the teacher/students and the plant simulator via suitable sockets. This architecture enables collaborative operation by the students who can operate the same simulator concurrently as a team, remaining online during the cooperative session.

The web-based remote lab involves the following:

- An access management system (**AMS**)
- A collaboration server (**CS**)
- An experimentation server (**ES**).

These components can be implemented using any suitable combination of available information technology tools (MATLAB, LabView, VRML, Java, etc.).

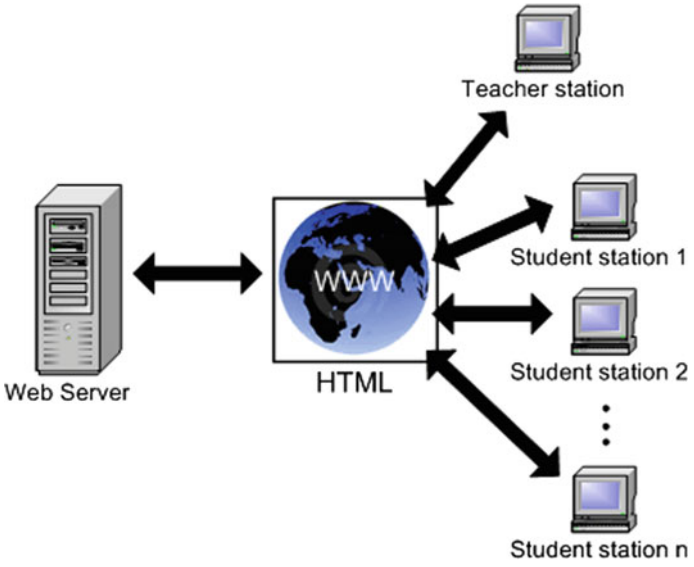



**Fig. 11.11** General architecture of a virtual web-based lab [66]

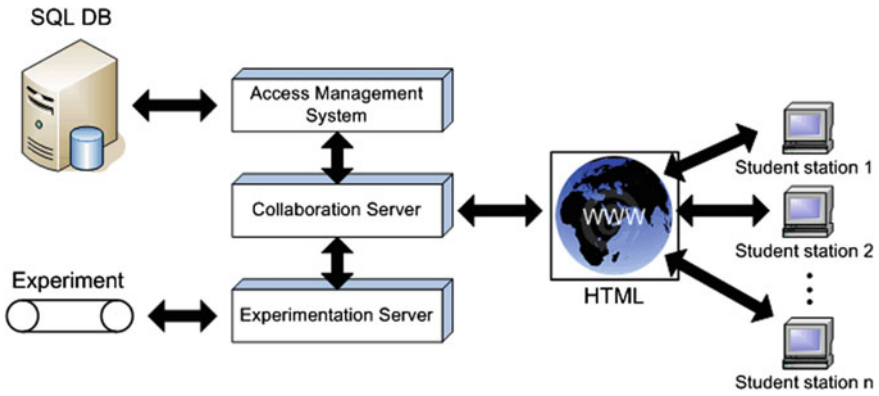



**Fig. 11.12** General architecture of a web-based remote lab (experiment) [66]

## 11.11 Information Technology in Medicine

The penetration of information technology (and informatics) into medicine started in the US in the 1950s, after the introduction of the digital computer. Today, many terms are used to represent the symbiosis of informatics with medicine. These are

- Medical informatics
- Biomedical informatics
- Health informatics.

All these terms represent a field that lies at the intersection of information technology and science, biology, and medical/health care [72, 73]. Whatever name is used, medical informatics deals with all issues of understanding and implementing effectively the organization, analysis, treatment, and use of information in health care. These issues involve the resources, equipment, and techniques needed for the optimization of acquisition, storage, retrieval, exploitation, and use of medical/healthcare data. Figure 11.13 shows a schematic of the medical/healthcare informatics branches or particular fields, extracted from Hersh's website [74]. The contents of this diagram are self-explanatory.

Closely related, but not identical, to medical/health informatics is the branch of *telemedicine*, a clinical medical application where medical data are transferred between remote places via interactive multimedia, for consultation and remote medical examinations and, sometimes, medical treatment [75–78]. Telemedicine usually refers to the provision of clinical services. There is an alternative term, i.e., *telehealth* or *e-health*, which includes both clinical and nonclinical services (e.g., administration, and medical education/research). The benefits of telemedicine (and e-health) for society are enormous, especially for remote/isolated places and communities. Subfields of telemedicine are all specializations and medical procedures/examinations that include the prefix "tele" (or "e"), for example, *teleradiology* (for radiology), *telesurgery* (for surgery), *telecardiology* (for

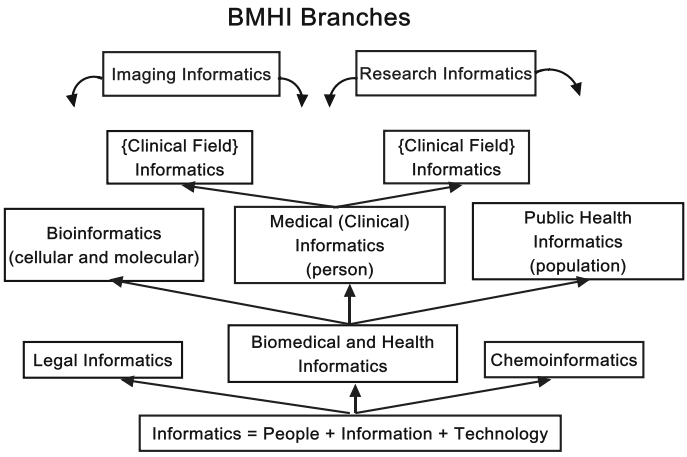



**Fig. 11.13** Network of biomedical and health informatics (BMHI) areas dealing with the individual person, population, biology, clinical process, etc

cardiology), *teleneurology* (for neurology), and so on. A quickly growing telemedicine service is the monitoring of blood pressure and cardiac functioning at home. This is useful for patients with chronic diseases and dysfunctions.

## 11.12 Information and Communication Technology in Transportation

The fundamental goals and objectives of the transportation community have not changed very much, but the challenges of current technologies that dominate the information and communication technology are greater than ever before. In transportation, major IT functions that need to be applied are [79]:

- Advanced user system interfacing
- Data management and data sharing
- Use of common semantics and standards
- Correct and accurate monitoring of the benefits of "technology transfer", including productivity gains
- Use of computer networking and web technologies (internets, intranets and Extranets).

Particular topics and types of IT that have contributed considerably to the advancement of transportation and need to be exploited further are as follows:

- Mobile communications
- Electronic data interchange (EDI)
- Very large-scale computing

- Global positioning systems
- Voice recognition
- Geographic information systems (GIS)
- Distributed and client–server technologies
- Airborne ground surveillance.

Transportation includes the following:

- Railway transportation
- Automobile transportation
- Sea transportation
- Aviation/Air transportation.

Automated/ICT-based railway systems have to incorporate the following in an integrated way [80–82]:

- Train traffic control subsystem
- Automatic train operation subsystem
- Electric power supply control subsystem
- Information transmission subsystem
- Automatic car inspection subsystem
- Supporting business management subsystem.

As an example, we mention the traffic control system of the Kobe City Subway which is of the decentralized type and uses, as the data transmission subsystem, the optical ADL-Net (Autonomous Decentralized Loop Net).

Information and communication technology is now applied to many various components of the automobile, such as engine control, transmissions, instrumentations, in-vehicle comfort, and in-vehicle entertainment. ICTs of automobile transportation include the following:

- Driver interface (DI)
- Advanced traveler information systems (ATIS)
- Collision avoidance and warning systems (CAWS)
- Automated highways systems (AHS)
- Vision enhancement system (VES)
- Advanced traffic management systems (ATMS)
- Commercial vehicle operation (CVO).

Collectively, the above functions and systems constitute the so-called *intelligent transportation system* (ITS), which, due to the programmable nature of integrated electronics (e.g., Field Programmable Gate Arrays (FPGAs) and the systems-on-a-chip (SOC) technologies), will enable further adaptability of the ITSs in various vehicle categories, driver capabilities, and environmental situations [83, 84]. Figure 11.14 shows a schematic of the basic functions of an ITS for electronic horizon, hazard messaging, collision management, safe speed, safe following, lane support, vulnerable road user recognition, safe distance, and correct direction.

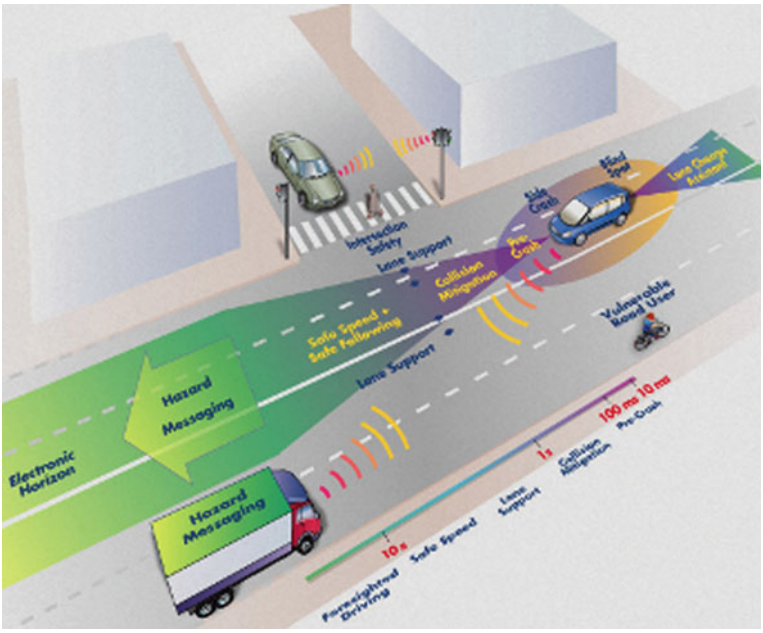



**Fig. 11.14** Illustration of an ITS. (http://www.em-its.eu/img/home.png)


**Fig. 11.15** Diagram of ETSI's virtual ITS for land, sea, and air transportation. (http://www.hitachi.eu/erd/research/ict/intelligent-transport/images/its01.jpg)

Figure 11.15 is a diagram illustrating ETSI's philosophy for a global land, sea, and air ITS design.

*Sea transportation* is performed by ships, which (together with the naval undersea vessels) require sophisticated IT-based automation. Issues that must be addressed include the following:

- Automatic roll stabilization
- Tracking of depth and obstacles (using solar)
- Localizing latitude and longitude for navigation (via GPS)
- Maneuvering supertankers through harbor channels and into port with complex terrain (using predictor displays working with GPS).

*Aviation systems* are automated to higher degrees in comparison to other man-made transportation systems (e.g., railway). Commercial aircraft have received the greatest attention for reasons of comfortability and the safety of the passengers. A very powerful facility added to the flight deck is the so-called *flight management system*, which extends manual control today via a *fly-by-wire*. A basic component of air transportation is *air traffic control* (**ATC**), which in our time is still implemented through *radio communication* equipment, using a "*duplex communication system*". Figure 11.16 presents an overall picture of the organization of the components/subsystems of a representative aviation system.

The *International Civil Aviation Organization* (**ICAO**) developed a world standard for aviation systems and suggested procedures for aviation regulatory agencies [85]. Conventional *air traffic controllers* (ATCs) are implemented via a network of stations around the world that employ two-way radios and "see" the

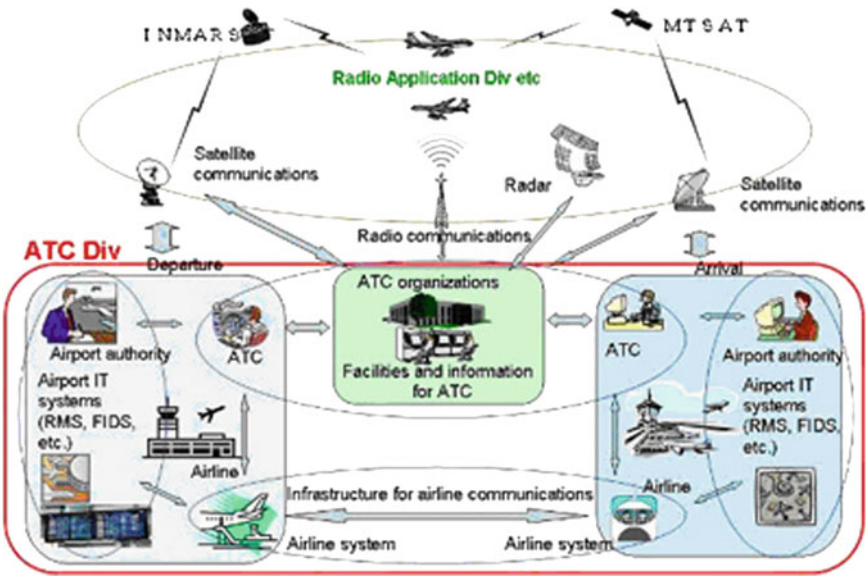



**Fig. 11.16** Architecture of aviation systems using ATC. (http://id.nec.com/en_ID/files/images/aeronautical1.jpg)

aircrafts through the radar. **ICAO** uses unique four-letter commercial airports' identification for ATC, whereas **IATA** (*International Air Transport Association*) uses three-letter codes primarily for travel agents and airline personnel. Under the free flight mode, pilots operating under *instruments flight rules* (**IFRs**) are able to select their aircraft's path, speed, and altitude in real time [86].

## 11.13 Concluding Remarks

Information manifestations in both natural and man-made systems have attracted the attention of humans throughout the evolution of humankind. On the life and biological side of information (*biocomputation*), we have two main avenues

- The study of the underlying natural/biological mechanisms of storing, processing, and transmitting information in living beings from cells to entire organisms. The purpose of this study is to understand better the informational aspects of life.
- The use of natural (biological) mechanisms of computation in the design and implementation of new types of man-made computational systems. The current achievements in this area (*DNA computers*) are encouraging and provide a very important challenge for our information society's future.

On the technological side, information and communication technology (ICT) is increasingly entering to the "core" of national and international competitive policies, due to its capacity as a primary player in the ongoing human growth, development, and modernization. Statistical studies and worldwide data support the view that ICT plays a crucial role in human society through innovation processes, novel products/services, and activities that enhance the competitive advantage of individuals, communities, and enterprises. Generally, ICT contributes to the rising of the *quality of life*. However, networked information technology has many, globally recognized, drawbacks especially in *private life*. This impact is considered in [87], where three different views on how society and networked information technology are changing each another are presented.

These three essays deal with the following issues:

- Has the ice man arrived? Tact on the internet (Jonathan Grudin)
- When information technology "goes social?" (Marti Hearts)
- Living networked in a wired world (Barry Wellman).

Jonathan Grudin points out that as ICT continually invades the more sensitive aspects of human life, we are forced to make it more concerned about special expectations and try to understand fully the nature of computer-assisted services.

Marti Hearts is concerned with the question, "how newly internetworked IT allows people acting in their own self-interest to indirectly affect the experiences of other people." She forecasts that people will attempt to trick or deceive systems that inherently support social activities, such as running auctions.

Barry Wellman introduces the term "*glocalization*" to express the process of simultaneously acting *intensely globally* and *intensely locally*. He describes the manner in which the structure of social networks influences the ways we live and work, and explains how computer-mediated communication (CMC) facilitates this *glocalization* transition process in social lifestyles and infrastructure. According to him, living in social networks (i.e., people and organizations connected via computer networks) differs from living in groups in the following aspects:

- It enhances the ability to be connected to many different *milieus* without decreasing the involvement in any one of them.
- It weakens the control of any one milieu over the individuals and reduces the commitment of any one milieu for a person's quality of life.
- It changes the types of interactions from the natural ones (based on gender, age, race, etc.) to those adopted over the course of life (e.g., lifestyles, voluntary desires, common norms, etc.).
- It weakens the ties that link and integrate social groups, without keeping them isolated within hard-bounded boxes.
- It increases the free choices and releases people from the feeling of "belonging" to the groups. In general, living in social networks reduces the identity and pressures of belonging to groups and increases the opportunities, surprises uncertainty, and globalized behavior of the individual members.

Tim Berners-Lee, the father of the Internet declared that: "One of the things I always remain concerned about is that the Internet remains neutral," and Alberto Ibargüen states: "The free flow of information is of paramount importance to communities in a democracy and maintaining the World Wide Web free is critical for the future of that free flow" (www.webfoundation.org).

The above points of view about IT and the Internet/WWW, as well as the views of many other workers in the field, warn us that both designers and users of IT and computer networking must pay particular attention to the issues of the social impact of these technologies [88–90]. Here, the *ethics* in the field of ICT is an important issue of human concern. There are already signs that this technology is a prime target for those wishing to abuse or misuse its very advantages. Unethical behavior by these people presents a real threat to social gains made possible by information technology. Among the serious ethical concerns involving ICT misuse are those that have to do with the issues of privacy and ownership [91–93]. In particular, Baird, Ramsower, and Rosenbaun have classified the contributions of their book [93] into four themes: (i) anonymity and personal identity in cyberspace; (ii) personal privacy in the light of extensive storage and dissemination of personal data; (iii) ownership of intellectual property and copyright law; and (iv) the impact of ICT on democracy and society. On this issue, Rogerson stated that: "*Communication without moral application is at best a wasted opportunity and at worst a dangerous threat to society and the rights of its citizens*" (Centre for Computing and Social Responsibility) [94].

## References


1. L. Adleman, Computing with DNA. Sci. Am. **279**(2), 54–61 (1998)
2. J. Bath, A. Turberfield, DNA Nanomachines. Nat. Nanotechnol. **2**, 275–284 (2007)
3. D. Endy, Foundations for engineering biology. Nature **438**, 449–453 (2005)
4. L. Kari, DNA computing, the arrival of biological mathematics. Math. Intell. **19**(2), 9–22 (1997)
5. M. Nagasaki, S. Onami, S. Miyano, H. Kitano, Biocalculus: its concept and molecular interaction. Genome Inf. **10**, 133–143 (1999)
6. D. Dasgupta (ed.), *Artificial Immune Systems and Their Applications* (Springer, Berlin, 1998)
7. V.B. Bajic, T.T. Wee, *Information Processing and Living Systems* (World Scientific, Singapore, 2005)
8. J. Collier, in *Information in Biological Systems*, ed. by P. Adriaans, J. Van Benthem. Handbook of Philosophy of Science, vol. 8, (Philosophy of Information) (North Holland, Dordrecht, 2008), pp 763–787
9. F. Crick, On protein synthesis. Exp. Biol. **12**, 138–163 (1958)
10. K. Lorenz, Analogy as a Source of Knowledge, *Nobel Lecture*, 12 Dec 1973
11. C.T. Bergstrom, M. Rosvall, The transmission sense of information. Biol. Philos. Online: Oct 1 2009
12. D. Campbell, Blind variation and selective retention in creative thought as in other knowledge processes. Psychol. Rev. **67**, 380–400 (1960)
13. K. Popper, *Objective knowledge: an evolutionary approach* (Clarendon Press, Oxford, 1979)
14. R. Boyd, P. Richerson, *Culture and evolutionary process* (University of Chicago Press, Chicago, 1985)
15. J. Sweller, Natural information processing systems. Evol. Psychol. **4**, 434–458 (2006)
16. J. Maynard Smith, E. Szathmary, *The Major Transitions in Evolution* (W.H. Freeman, Oxford, 1995)
17. J. Rissanen, *Stochastic Complexity in Statistical Inquiry* (World Scientific, Singapore, 1989)
18. C.S. Wallace, P.R. Freeman, Estimation and Inference by Compact Coding. J. Royal Statist. Society (Series B: Methodology) **49**, 240–265 (1987)
19. J.D. Collier, Entropy in evolution. Biol. Philos. **1**, 5–24 (1986)
20. J.D. Collier, hierarchical dynamical information systems with a focus on biology. Entropy **5**, 100–124 (2003)
21. K. Sterelny, P.E. Griffiths, *Sex and Death: an Introduction to Philosophy of Biology* (University of Chicago Press, Chicago, 1999)
22. P. Godfrey-Smith, in *Information in Biology*, ed. by D.L. Hull, M. Ruse. The Philosophy of Biology (Chap.6) (Cambridge University Press, Cambridge, 2008)
23. N. Shea, Representation in the tems. Biol. Philos. **22**, 313 (2007)
24. J. Maynard Smith, The concept of information in biology. Philos. Sci. **67**, 177–194 (2000)
25. P.E. Griffiths, Genetic information: a metaphor in search of a theory. Philos. Sci. **68**, 394–412 (2001)
26. P.E. Griffiths, in *Molecular and Developmental Biology*, ed. by P.K. Machamer, M. Silberstein. The Blackwell Guide to Philosophy of Science (Blackwells, Oxford, 2002), pp 252–271
27. M. Chi, R. Glaser, E. Rees, Expertise in Problem Solving, in *Advances in the Psychology of Human Intelligence*, ed. by R. Sternberg (Erlbaum, Hillsdale, NJ, 1982), pp. 7–75
28. W. Carroll, Using worked examples as an instructional support in the algebra classroom. J. Educ. Philos. **86**, 360–367 (1994)
29. A. Newell, H. Simon, *Human Problem Solving* (Prentice Hall, Englewood Cliffs, NJ, 1972)
30. E. Jablonka, M.J. Lamb, *Evolution in four dimensions: genetic, epigenetic, behavioral, and symbolic variation in the history of life* (MIT Press, Cambridge, MA, 2005)
31. H. Simon, Invariants of human behavior. Annu. Rev. Psychol. **41**, 1–20 (1990)

32. J. Hickman, F. Darema, W.R. Adrion, Biological Computation: how Does Biology Do Information Technology? *Report from NSF Workshop on Biological Computation* (2001). http://www.csd.uoa.ca/∼lila/biocamp.html
33. DNA Computers http://www.brookscole.com/chemistry-d/templates/student_resour
34. W. Liu, L. Gao, X. Liu, S. Wang, J. Xu, Solving the 3-SAT problem based on DNA computing. J. Chem. Info. Modeling **43**(6), 1872–1875 (2003)
35. J. Howard Price, *Israelis Develop DNA Computer* (The Washington Times, April 29, 2004). http://www.wisdom.weizmann.ac.il/∼udi/PressRoom/new_pages
36. J.R. Beniger, *The Control Revolution: technological and Economic Origins, of the Information Society* (Harvard University Press, Cambridge, MA, 1986)
37. D. Bell, *The Coming of Post-Industrial Society* (Basic Books, New York, 1976)
38. A. Alter, in *Knowledge Workers Need More Supervision*. CIO Insight http://www.cioinsight.com/article2/0,1397,1843978,00.asp. 5 Aug 2005. Also: http://www.shafee.com/my-papers.html
39. J.I. Salvaggio, *The Information Society: economic, Social and Structural Issues* (Erlbaum, Hillsdale, NJ, 1989)
40. N.R.C. Committee, *Information Technology in the Service Society: a Twenty-First Century Lever* (National Academics Press, Washington, DC, 1994)
41. K. Eason, *Information Technology and Organizational Change* (CRC Press/Taylor and Francis, Bristol, PA, 1988)
42. C.T. Marsden, *Regulating the Global Information Society* (Routlege/Taylor and Francis, London, 2000)
43. W.H. Dutton, *Society on the Line: information Politics in the Digital Age* (Oxford University Press, Oxford, 1999)
44. G. Berka, *Putting the Information Society to the Test: comments on the Research into Mechanized Communication* (Oldenbourg Verlag, Munich, 1994), pp 57–66
45. C. May (ed.), *Key Thinkers for the Information Society*, vol. 1 (British International, Studies Association, Routlege/Taylor and Francis, New York, London, 2002)
46. Office Automation, Webopedia. Com http://www.webopedia.com/TERM/O/office_automation.html
47. http://visual.merriam-webster.com/images/communications/office-automation/internet-uses.jpg
48. D. Basset, K. Clinard, J. Crainger, S. Purucker, D. Ward, tutorial course: distribution automation. *IEEE Publication*, 88EH0280-8-PWR, 1988
49. L. Grasberg, L.A. Osterlund, in *SCADA/EMS DMS: A Part of the Corporate IT Systems*. Proceeding of PICA-2001: power Industry Computer Applications, 22nd IEEE Power Energy Social Intelligence Conference Sydney, NSW, Australia, pp 141–147, 2001
50. M.K. Jbira, M.S. Smiai, in *Computer-Based System for Sustainable Development of the 'Saudi Electricity Company' Distribution Automation Network*. Proceeding of 18th National Computer Conference, Saudi Computer Society, 2006
51. R.P. Gupta, R.K. Varma, Power distribution automation: present status. Acad. Open Internet J., **15**, 1–10, 2005 www.acadjournal.com/2005/∼15/part1/p1/
52. S. Curcic, C.S. Ozveren, K.L. Lo, Computer—based strategy for the restoration problem in electric power distribution systems. Proc. IEEE, Gener., Transm. Distrib. **144**(5), 389–398 (1997)
53. J.-B. Waldner, *Principles of Computer-Integrated Manufacturing* (Wiley, New York, 1992)
54. T.C. Chang, R.A. Wysk, *An Introduction to Automated Process Planning Systems* (Prentice-Hall, Engle wood Cliffs, NJ, 1985)
55. D.S. Appleton, The state of CIM. Datamation **15**, 66–72 (1984)
56. S.G. Tzafestas, in *AI Techniques in Computer-Aided Manufacturing Systems*, ed. by H. Adeli. Knowledge Engineering, vol. II (Mc-Graw-Hill, New York, 1990), pp 161–212
57. D. Doscher (ed.), CIM Framework Specification, Version 2.0, Sematech. Technology Transfer 93061697 J-ENG, Jan (1998). http://www.semantech.org/docubase/documents/1697jeng.pdf

58. K.R. Lee, in *Impacts of Information Technology on Society in the New Century*. http://www.zurich.ibm.com/pdf/Konsbruck.pdf
59. B.Z. Barta, P. Juliff, *Place of Information Technology in Management and Business Education* (Business and Economics, Springer, Berlin, 1997)
60. K. Eason, *Information Technology and Organisational Change* (Taylor and Francis, London, 1988)
61. V.G. Cerf, in *Computer Networking: global Infrastructure for the 21st Century*. http://www.cs.washington.edu/homes/lazowska/cra/networks.html
62. Louisiana State University, Information Technology Services. Standards for Enterprise Information Systems and Solutions (Version 1.0). http://itsweb.lsu.edu/ITS_Security/files/item891.pdf
63. S.G. Tzafestas (ed.), *Web-Based Control and Robotics Education* (Springer, Dordrecht/Berlin, 2009)
64. M.M. Driscoll, Defining Internet-Wed-Based Training. J. Perform. Instr. **36**(4), 5–9 (1997). http://home.vicnet.net.au/~carlrw/net2000/ten_things_we_know.html
65. D. Mioduser, R. Nachmias, U. Lahav, A. Oren, Web-based learning environments: current pedagogical and technological state. J. Res. Comput. Educ. **33**(1), 55–76 (2000). http://muse.tau.ac.il/publications/wble-54.html
66. S.G. Tzafestas, in *Teaching Control and Robotics Using the Web*, ed. by S.G. Tzafestas. Web-Based Control and Robotics Education, (Springer, Dordrecht/Berlin, 2009), pp 1–38
67. G.J.C. Cpinga, M.H.G. Verhaegen, M.J.J.M. Van de Ven, Toward a web-based study support environment for teaching automatic control. IEEE Control Sys. Mag. **20**, 8–19 (2000). http://lcewww.et.tudelft.net/~babuska/Gerald.html
68. M. Casini, D. Prattichizzo, A. Vicino, The automatic control telelab: a user-friendly interface for distance learning. IEEE Trans. Educ. **46**(2), 252–257 (2003)
69. Q. Yu, B. Chen, H.H. Cheng, Web-based control system design and analysis. IEEE Control Sys. Mag. **24**, 45–57 (2004). http://www.softintegration.com/webservices/control/
70. R. Puerto, L.M. Limenez, O. Reinoso, C. Fernandez, and R. Neco, in *Remote Control Laboratory Using Matalab and Simulink: application to DC Motor*. Proceedings of the Second IFAC Workshop on Internet Based Control Education (IBCE 2004) (Grenoble, France, 2004), pp. 5–7
71. Drexel University, in *Laboratory Development for Robotics and Automation Education Using Internet Based Technology*, 2006. http://www.drexel.edu/goodwin/faculty/ASEENSFCCon06_final.pdf
72. P. O'Carrol, W. Yasnoff (eds.), *Public Health Informatics and Information Systems* (Springer, Berlin/New York, 2002)
73. E. Shortliffe, J. Cimino (eds.), *Biomedical Informatics: computer Applications in Health Care and Biomedicine* (Springer, Berlin/New York, 2006)
74. W. Hersh, in *What is Biomedical and Health Informatics?* http://www.billhersh.info/whatis/
75. N. Brown, in *Telemedicine Coming of Age*. http://tie.telemed.org
76. R. Higgs, *What is Telemedicine?* http://www.icucare.com/PageFiles/Telemedicine.pdf
77. http://tie.telemed.org
78. ADV Communications: Telemedicine http://www.adv.comms.co.uk/telemedicine/definition.htm
79. J.L. Western, B. Ran, Information Technology in Transportation, Key Issues and a Look Forward. Transportation Research Board, Committee on Information Systems and Technology (A5003) (Washington, DC, 2000). http://pubsindex.trb.org/view.aspx?sid=639264
80. H. Ihara, M. Nohmi, in *Current Status of Microcomputer Applications on Railway Transportation Systems*, ed. by S.G. Tzafestas, J.K. Pal. Real Time Microcomputer Control of Industrial Processes, (Kluwer, Dordrecht/Boston, 1990), pp 481–508
81. H. Matsumaru, Recent computer application systems for railways. Hitachi Rev. **31**(1), 1–6 (1984)

82. M. Yabushita, in *Autonomous Decentralization Concept and its Application to Railway Control Systems.* Proceeding of 34th IEEE Vehicular Technology Conference, vol. 34, pp 285–260, (1984)
83. M.M. Popp, B. Farder, in *Advanced Display Technologies, Route Guidance Systems and the Position of Displays in Cars*, ed. by A.G. Grade. Vision in Vehicles—III (Elsevier/North–Holland, Amsterdam, 1991), pp 219–225
84. L. Evans, *Traffic Safety and the Driver* (Van Nostrand Reinhold, New York, 1991)
85. ICAO, in *Annexes to the Convention on International Civil Aviation*, (ICAO, Montreal, Canada)
86. M.S. Nolan, *Fundamentals of Traffic Control* (Books Code Publ. Co., Pacific Grove, 1999)
87. M.A. Hearst, The changing relationship between information technology and society. IEEE Intelligent Systems, 7–17 Jan/Feb (1999)
88. L. Garton, B. Wellman, Social impacts of electronic mail in organizations: a review of the research literature. Comm. Yearb. **18**, 434–453 (1995)
89. J. Scott, *Social Network Analysis* (Sage, London, 1991)
90. B. Wellman, in *An Electronic Group is Visually a Social Network*, ed. by S. Kiesler. Culture of the Internet, (Erlbaum, Mahwah, NJ, 1997), pp 179–205
91. R.J. Severson, *The Principles of Information Ethics* (M.E. Sharpe Armonk, New York, 1997)
92. T. Ward Bynum, S. Rogerson (eds.), in *Computer Ethics and Professional Responsibility* (Wiley-Blackwell, Hoboken, NJ, 2003)
93. R.M. Baird, R.M. Ramsower, S.E. Rosenbaum (eds.), in *Cybernetics: social and Moral Issues in the Computer Age*, (Prometheus Books, 2000)
94. CCSR: Centre for Computing and Social Responsibility, http://www.ccsr.cse.dmu.ac.uk/

# Chapter 12
# Feedback Control in Life and Society

*Feedback is the fundamental principle that underlies all self-regulating systems, not only machines but also processes of life.*

—Arnolst Tustin

*Mankind's history has been a struggle against a hostile environment. We finally reached a point where we can begin to dominate our environment and cease being victims of the vagaries of nature in the form of fire, flood, famine, and pestilence… As soon as we understand this fact, our mathematical interests necessarily shift in many areas from descriptive analysis to control theory.*

—Richard Bellman

**Abstract** The aim of this chapter is to illustrate the role of feedback, negative and positive, in biological and societal systems and applications (technological, behavioral). Feedback, the third fundamental element of life and society, is a process which is based on energy, and exploits the information existing or generated in each particular case. The mathematical analysis of feedback is more easy to be made successfully for well-defined simple or complex man-made systems, and more difficult or incomplete for living and society systems. For the convenience of the reader the material of this chapter is presented via a number of selected biological, societal, and technological examples. These examples demonstrate that both negative and positive feedback is present and efficiently used by living organisms and human societies. Negative feedback offers the means for achieving stability and the goals of each case. Positive feedback is used whenever a purposeful oscillatory behavior is the desired goal. Negative feedback biological examples considered in this chapter are: temperature regulation, water regulation, sugar regulation, and hydrogen ion (pH) regulation. Positive biological feedback is illustrated by autocatalysis and auto-reproduction chemical reactions. Mathematical models and controllers in biological systems are provided for enzyme operation, biological rhythmic movement, insulin–glucose balancing, and cardiovascular-respiratory system. On the societal side, this chapter discusses four technological (hard) systems (process control, manufacturing systems control, air-flight control, and robotic systems control), and two types of soft systems, namely management



S. G. Tzafestas, *Energy, Information, Feedback, Adaptation, and Self-organization*, Intelligent Systems, Control and Automation: Science and Engineering 90,
https://doi.org/10.1007/978-3-319-66999-1_12

control and economic system control. In hard systems, the control means include prime movers and end effectors, whereas in soft systems, the means of control are regulation laws and rules posed by rulers, managers, and government.

**Keywords** Feedback control · Living systems · Internal environment
Homeostasis · Temperature regulation · Water regulation (osmoregulation)
Sugar regulation · Hydrogen-ion regulation · Hypothalamus · Pancreas
Insulin · Glucagon · Glycogen · Insulin-glucose system · Acidosis
Alkalosis · Autocatalysis · Biphasic modulation · Systems biology
Enzyme operation · Linear biological models · Biological rhythmic movement
Nonlinear biological models · Cardiovascular–respiratory system
Feedback control in society · Hard systems · Soft systems · Hard/soft systems

## 12.1 Introduction

*Feedback* is an intrinsic function in all biological, technological, and societal systems. In Chaps. 6 and 7, the history of the study of "*feedback*" was presented, and the classical and modern ways in which it is implemented were described. Feedback, the third pillar of life and society, is a function based on energy and exploiting the information that exists or is generated in each particular case. The bulk of analysis and design methodologies in the feedback and control field is more successfully applied to well-defined *man-made* simple or complex systems. *Living systems*, especially the multicellular ones, are very complex, and typically their study in a standard mathematical quantitative way is much more difficult than technological systems. The same is true for *society systems,* which have a behavioral structure and cannot be precisely mathematically modeled.

In this chapter, we review the role of feedback in both living organisms and societal systems. Both negative feedback and positive feedback are considered. The material is presented with the aid of selected representative examples that best illustrate the concepts and operational modes involved. Specifically, negative biological feedback is illustrated by temperature regulation, water regulation, sugar regulation, and hydrogen-ion regulation. Positive biological feedback is illustrated by autocatalytic or autoreproduction reactions, such as biphasic modulation and the development of sinks. On the methodological side, we present how linear and nonlinear mathematical control models can be used for the analysis of biological processes, in particular for enzyme operation, biological rhythmic movement, the insulin-glucose balance system, and the cardiovascular-respiratory system. Regarding the role of feedback in society's systems, we discuss four classes of technological (hard) systems, viz., process control, manufacturing control, flight/air-traffic control and robot control systems, and two classes of soft systems, namely, management control and economic control systems. Of course, the above examples do not exhaust the subject but sufficiently illustrate the role of feedback in life and society.

## 12.2 Feedback Control in Living Organisms

### *12.2.1 General Issues*

The study of living systems by feedback and control-science methods is actually a follow up to *Erwin Schrödinger's* work concerning the question "*what is life?*" [1]. Of course, the development of a general systems approach to biological processes goes back to the ancient times and the findings of Mendel and Darwin. Schrödinger considered and studied life processes from the point of view of *physics* and argued that specific life functions can be assigned to individual cells and molecules. This, together with his *macroscopic systems* viewpoint of life, can be considered as the first identifiable landmark of what is today called "*Systems Biology*" [2–8]. Systems biology courses are currently integrated into process dynamics and control curricula, particularly in chemical engineering departments [9].

The fact that the "*internal environment*" (*milieu interieur*) of living organisms is kept remarkably constant, despite the changes occurring in the external environment, was first revealed by *Claude Bernard*, the founder of modern experimental biology (1813–1878) [10]. For this capability of life, *Walter Cannon* (1871–1945) coined in 1932 the term *homeostasis* [11]. This term comes from the Greek words "*όμοιο*" (homeo = the same) and "*στάσις*" (stasis = stationary/standing). Since then, the concept of homeostasis (or homeostasy) has been a key concept of the cybernetics and systems biology field, meaning life's general feature of "*resisting to change*". Major contributors in this field include, among others: *W. Cannon* [11], *W. Hodgkin* [12], *A. Huxley* [12], *N. Wiener* [13], *L. Bayliss* [14], *M. Mesarovic* [15], and *L. Von Bertalanffy* [16]. A small set of important works in the field of systems biology and feedback control in living systems is provided in [17–27].

### *12.2.2 Negative Feedback Biological Systems*

*Negative feedback* is always applied in living systems in order to achieve homeostasis, i.e., maintain an equilibrium state (usually a dynamic equilibrium state) despite changes in the environment. The way negative feedback operates has been described in Chap. 6. Here, we will briefly describe the following negative feedback (*homeostasis*) systems in pure biological terminology without any mathematical modeling [24–32]:

- Temperature regulation
- Water regulation
- Sugar regulation
- Hydrogen-ion regulation

#### 12.2.2.1 Temperature Regulation

Mammals are warm blooded and so the enzymes involved require a certain temperature to function optimally. In addition, the water concentration of a cell and its chemical composition must be maintained at certain levels in order for the cellular operation to be normal. Humans operate normally (healthily) in the temperature range 37 ± 0.5 °C and may suffer serious consequences if their body temperature deviates by as little as 2 °C outside this region.

The body thermostat (temperature regulator) is a region of the brain called the "*hypothalamus*", which provides the temperature set point and plays a dominant role in the integration of temperature information. The variation of body temperature is detected by sensors in the skin (*peripheral sensors*) and in the hypothalamus (which measure the *core temperature*). The hypothalamus controls the so-called "*thyroid gland*", an endocrine organ. Increased sweating, perspiration for evaporative cooling, and the rate of breathing is corrective responses to reduce body temperature in case the body becomes too warm. Similarly, when the body is too cool, the hypothalamus stimulates shivering, a drop in metabolic rate, and repeated contractions of muscle fibers to generate heat in the body.

The above temperature regulation functions of the hypothalamus can be illustrated as shown in Fig. 12.1 [29].

#### 12.2.2.2 Water Regulation

The regulation of water concentrations in the bloodstream (or, as otherwise called, *osmoregulation*) controls the amount of water available to cells to absorb. The *total body water* (**TBW**) is about 70–75% of the weight of a person of 75 kg, but this percentage varies with age. A baby at birth has about 80% water, and an elderly person may have about 50% water. The *tonicity* of a solution in a living cell can be: *isotonic* (when the solution bathing a cell does not cause the cell to osmotically gain or lose water), *hypertonic* (when the solution contains more water than needed, in which case the cell loses water and shrinks), or *hypotonic* (when the solution contains smaller amounts of osmotically active particles, in which case the cell takes up water until it bursts).

The homeostatic water control process is shown in Fig. 12.2.

This process involves the following:

- Active negative feedback is a response to a change in water concentration.
- The water concentration is detected by "osmoreceptors" in the hypothalamus.
- The hypothalamus sends chemical signals to the pituitary gland next to it.
- The pituitary gland secretes *antidiuretic hormone* (**ADH**), which is received by the kidneys which maintain normal water levels.
- The hormone that reaches the kidney changes the tubules of the kidney to become more or less permeable to water (as a response to higher or less ADH concentrations, respectively).

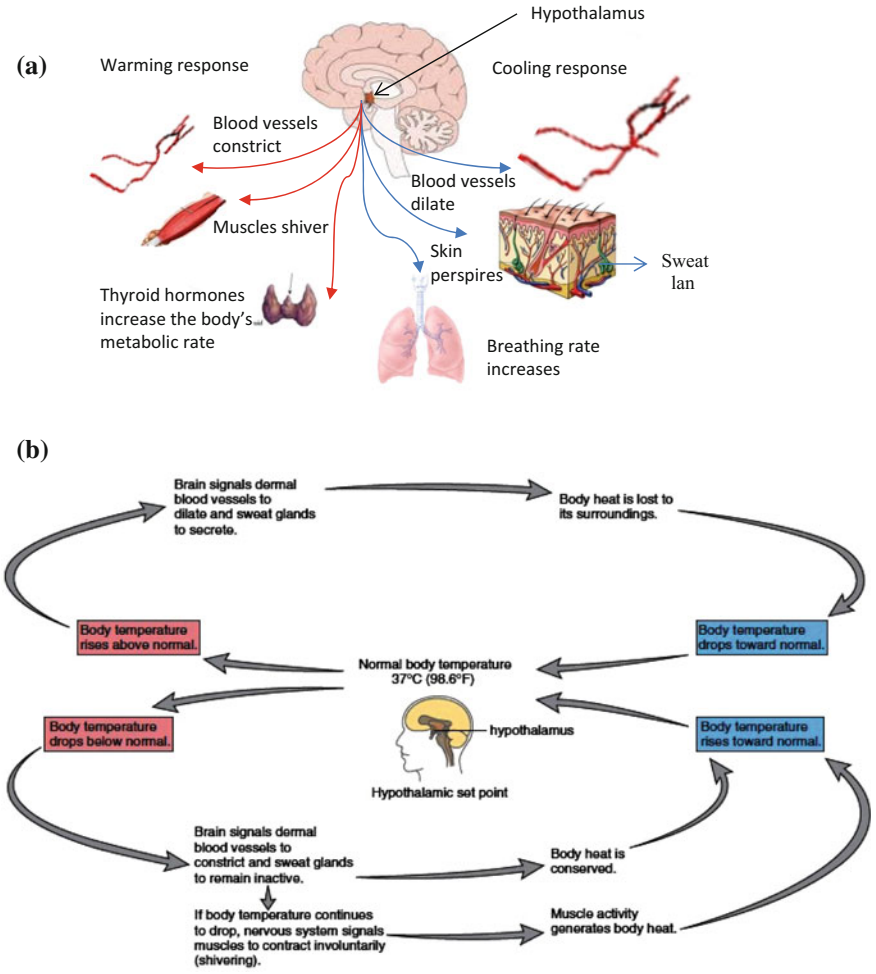
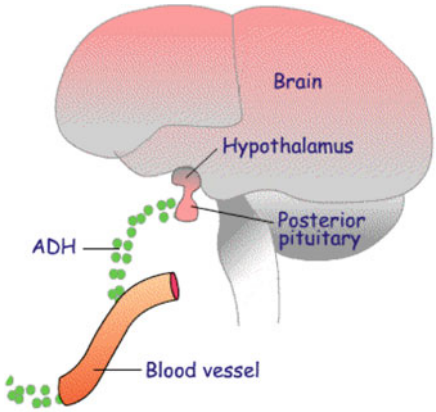


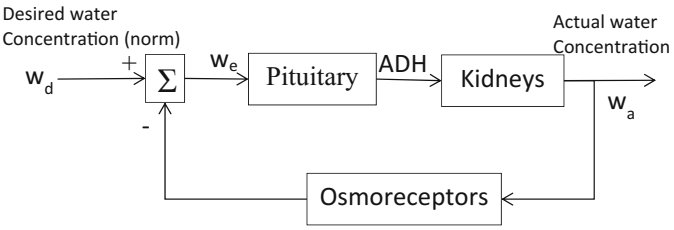

**Fig. 12.1** Representation of the temperature regulatory functioning of the hypothalamus, as control center: **a** Body organs involved in the process [29], **b** Feedback loop representation (http://encyclopedia.lubopitko-bg.com/images/Homeostasis%20and%20body%20temperature%20regulation.jpg). The reader is informed that web figures and references were collected at the time of the writing this book. Since some of them may no longer be valid due to change or removal by their creators, they may no longer be useful

#### 12.2.2.3 Sugar Regulation

To create the required amount of ATP, which is dynamically changing, the body needs appropriate amounts of *glucose* to maximize its energy-making capability. The hormones that are responsible for the control of the glucose level in the body are the following:

- Insulin
- Glucagon



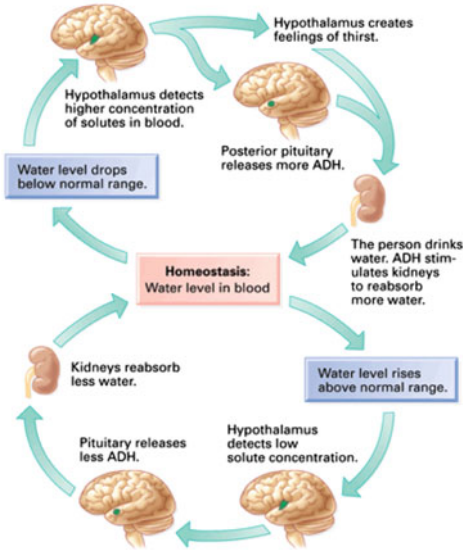




**Fig. 12.2** Water regulation process in humans. *Source* http://students.um.edu.mt/aaxi0003/img10D.gif, http://humanbiologylab.pbworks.com/f/1327381134/Water%20Balance.gif

The level of glucose in the blood is sensed by the receptors of the *pancreas*. Two kinds of pancreas cells release insulin and glucagon, which are directed to the liver in amounts that depend on the glucose concentration. Specifically:

- If the glucose level increases, less glucagon and more insulin are released by the pancreas targeting the liver.
- If the glucose level decreases, the pancreas releases less insulin and more glucagon targeting the liver.

*Glycogen* is the form in which glucose is stored in the liver. When insulin is released as a result of increased glucose levels (which leads to the production of more glycogen), the excess glucose can be stored as glycogen in the liver for later needs. Glucagon is released when the glucose levels fall to promote the conversion of glycogen into glucose and so compensate for the lack of glucose.

When a person has no ability to produce adequate insulin (which means that the conversion of glucose is not sufficient), she/he is said to suffer from *diabetes melitus*. In this case, the patient injects insulin after snacks and meals to keep the storage of glucose within the normal levels. In cases of emergency (known as "*fight-or-flight reactions*"), the body releases adrenaline to overpower the homeostatic regulation of glucose. Adrenaline is secreted by *adrenal glands*. This secretion leads to increased metabolism, breathing, and heart rate. After the emergency situation has ended, adrenaline levels decrease, and the control is taken again by the homeostatic regulation systems. Another type of diabetes is the so-called "*diabetes insipidus*", a situation where excess urine is excreted due to an inability to produce the ADH that increases the retention of water. Diabetes damages the eyes, the kidney, and nervous system, with consequent damage to the nerve fibers all over the body.

A schematic representation of the negative feedback control system for biological sugar is shown in Fig. 12.3.

Even people who control their sugar levels with insulin and other drugs have, statistically, reduced life expectancy.

#### 12.2.2.4 Hydrogen-Ion Regulation

The regulation of the hydrogen ions (**pH**) known as *acid–base balance* (**ABB**) is one of the most important homeostatic systems of the human body. The free hydrogen ions (the ions that are not bound to other molecules) are free to react. The concentration of hydrogen ions is expressed on the **pH scale** as

$$\mathrm{pH} = -\log_{10}(\mathrm{H}^+)$$

where $\mathrm{H}^+$ is the hydrogen-ion concentration. Clearly, each unit of pH corresponds to a tenfold change in the ion concentration. Neutral pH corresponds to seven units (i.e., $10^{-7}$ mol/L of free ions, which is the ion concentration of pure water). Solutions with pH lower than seven have more hydrogen ions and are *acidic*, whereas if $\mathrm{pH} > 7$ the solution contains fewer hydrogen, i.e., it is *alkaline*. The normal pH in

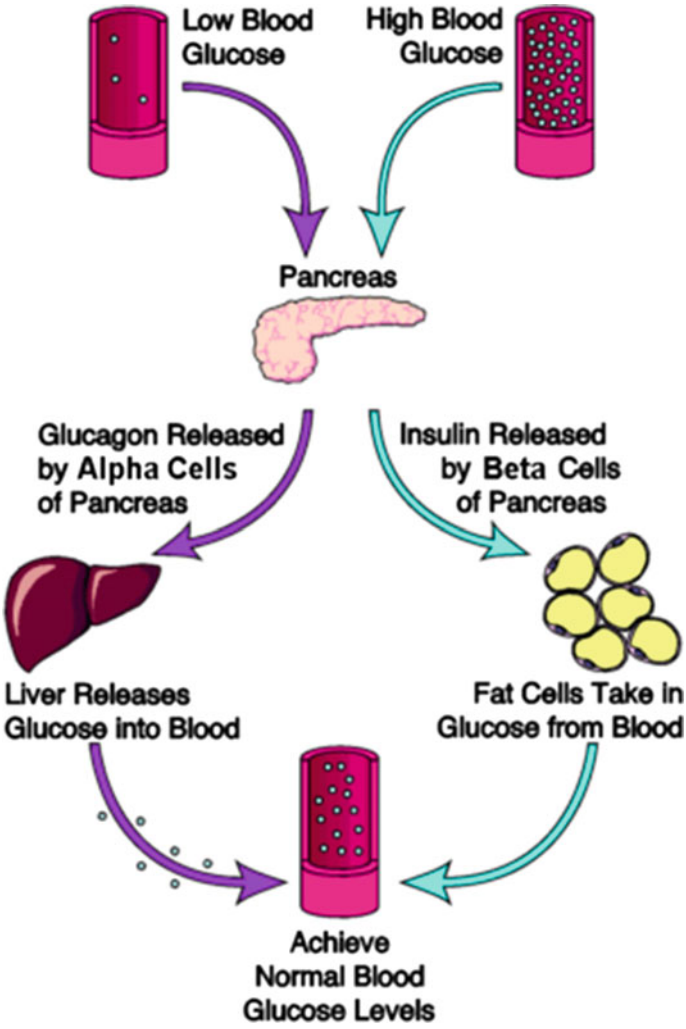



**Fig. 12.3** Sugar homeostatic system. *Source* http://www.endocrineweb.com/conditions/diabetes/normal-regulation-blood-glucose

the human blood is about 7.4 (actually *arterial* pH = 7.45, *venus* pH = 7.35). When pH is less than 7.35, we have *acidosis,* and if pH > 7.45, we have *alkalosis*. Of course, not all fluids in the human body have the same normal pH level.

Changes in pH influence strongly the excitability of nerves and muscles. For example, a fall in pH results in a strong effect on cardiac muscle, and an increase in pH is followed by an increased secretion of $H^+$ from the kidney that leads to increased *potassium ions* ($K^+$) in the blood and cardiac *arrhythmia* or *dysrhythmia*.

Without going into detail, the level of pH is controlled by carbon dioxide/bicarbonate $\left(\frac{CO_2}{HCO_3^-}\right)$. The $CO_2$ generated by tissue metabolism is released to the

atmosphere by the lungs. A decrease in pH of the blood is followed by a proportional rise in respiration and an increase in the rate and/or depth of breathing (Fig. 12.4a, b). This negative feedback is supported by the kidneys, which control the amounts of $H^+$ and $HCO_3^-$ excreted by the urines. Disturbances on acid–base are caused by changes in the $\frac{HCO_3^-}{CO_2}$ ratio. Respiratory acid–base disorders cause changes in the arterial partial pressure of $CO_2(pCO_2)$, whereas metabolic acid–base disorders produce changes in $H^+$ or $HCO_3^-$. Acid–base changes that are due to respiration can be corrected only by non-respiratory mechanisms. But, changes in pH caused by metabolic disturbances (except to renal causes) can be corrected by both respiratory and renal homeostatic mechanisms [30]. Figure 12.4c shows the acid–base diagram for both acidosis and alkalosis.

The normal range for $pCO_2$ is 35–45 mmHg and for $HCO_3^-$ is 22–28 mM. This diagram helps to identify a person's disorder, as follows:

**Acidosis** (pH < 7.35)

*Respiratory acidosis* if $pCO_2 > 45\,\text{mmHG}$
*Metabolic acidosis if* $\left[[HCO]_3^-\right] < 22\,\text{mM}$

**Alkalosis** (pH > 7.45)

*Respiratory alkalosis if* $pCO_2 < 35\,\text{mmHg}$
*Metabolic alkalosis if* $\left[[HCO]_3^-\right] > 28\,\text{mM}$

In respiratory acidosis, if $\left[[HCO]_3^-\right] > 28\,\text{mM}$ then there occurs renal compensation, and in metabolic acidosis of $pCO_2 < 35\,\text{mmHg}$ then we have respiratory compensation. Correspondingly, in respiratory alkalosis we have renal compensation when $\left[[HCO]_3^-\right] < 22\,\text{mM}$, and in metabolic alkalosis there is respiratory compensation of $pCO_2 > 45\,\text{mmHg}$. Details on the mechanism of acid–base balance can be found in [30].

## *12.2.3 Positive Feedback Biological Systems*

### 12.2.3.1 General Discussion

*Positive Feedback* is applied in living systems to efficiently utilize the deviation of a parameter $y$ from its initial value $y_0$ after certain conditions have been reached. Positive feedback promotes the fast auto-excitation of endocrine and nervous systems (particularly important is stress situations) and also is a basic factor in morphogenesis, growth, and development of organs. All of these positive feedback processes lead to a quick departure from the initial state.

One beneficial example of (self-limited) positive feedback in life takes place in *childbirth*, during the contractions of the *uterus*. The stretching of the uterus triggers the secretion of the hormone *oxytocin,* which stimulates uterine contractions

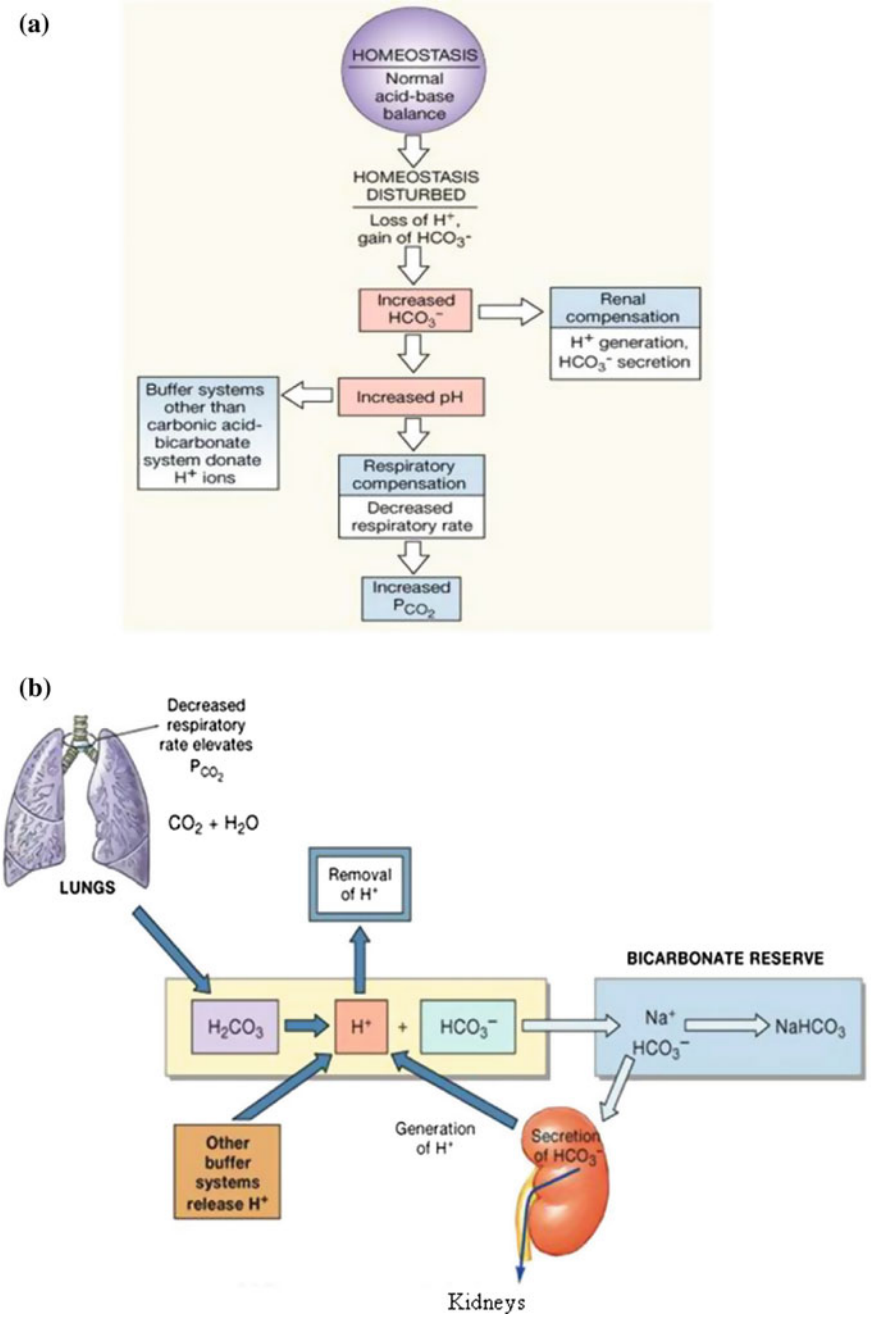



**Fig. 12.4** **a** Hydrogen-ion regulation (homeostasis) system, **b** Response to alkalosis, **c** Acid–base diagram (pH vs. $pCO_2$) [30]. The normal range is shown in the square area in the middle (a, b). *Source* http://www.austincc.edu/apreview/NursingPics/FluidPics/Picture18.jpg

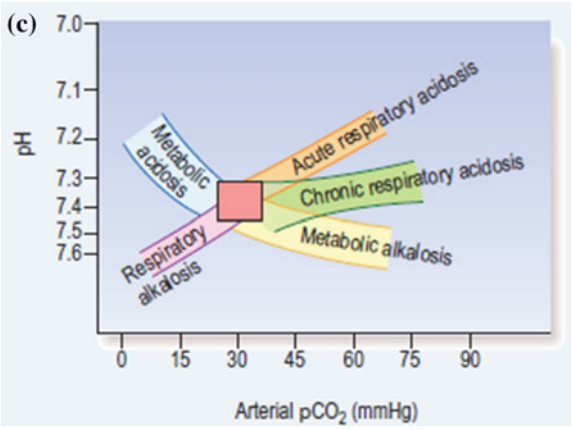



**Fig. 12.4** (continued)

and speeds up labor. The signals from the *fetus* regulate the series of events. The main event that initiates the positive feedback is the pressure of the fetus on the *cervix*. This process continues growing until the uterus succeeds to expel the baby. After that, the pressure on the cervix is relieved, the oxytocin secretion stops, and the uterine contractions cease. This means that the positive feedback loop is terminated by the birth of the baby, which signals the *self-terminating* condition.

However, very often positive feedback in living organisms lead to opposite results of homeostasis, i.e., to internal instability which in many cases can lead to fatal consequences. As an example of this is the *myocardial infarction* (*heart attack*) that is initiated when a small part of heart tissue has died and the heart pumps an insufficient quantity of blood. This has the consequence that the heart muscle itself is deprived of blood flow and starts to die, which may lead to the person's death.

An important biological process that involves positive feedback is the *build-up of binary biological memory* [33]. A common model found in many bistable genetic systems is a model of two interacting positive feedback loops. Bistable genetic systems possess a discontinuity of expression states, in which two distinct, stable steady states are obtained without the existence of stable intermediate states. The state which is occupied each time is determined by the history of the system. According to theoretical and experimental studies, there are two absolutely required conditions for getting *genetic bistability*. Firstly, some form of positive feedback controlling *gene expression* must be functioning, and second, the kinetic order or sensitivity of the system to the positive feedback element must be high [34, 35]. If these minimal requirements are satisfied, bistability can be anticipated to occur under certain environmental conditions. Examples of bistable biological processes are the genetic system that controls progression through the cell cycle; mammalian calcium signal transduction; and eukaryotic chemotaxis. *Ferrel* found that a suitable coupling of a positive feedback loop with a double negative feedback loop in a

system of opposing enzymes may lead to bistability over a wide range of conditions [35]. In general, distinct positive feedback loops that employ different biochemical mechanisms can be linked to give powerful genetic bistability, with scalability capability depending on factors that affect one of the feedback loops [33].

In the following, we discuss *autocatalysis,* which is a general class of positive-feedback biochemical processes encountered in a wide variety of cases.

#### 12.2.3.2 Autocatalysis

*Autocatalytic* (*or autoreproduction*) *reactions* take place when a product catalyzes a reaction and promote its own creation. Actually, biological systems can be regarded as complex reaction networks with multiple interconnected pathways and feedback loops. To study such biological systems, we must analyze their constituent unit processes. *Autocatalysis* is one of these basic processes in biology in its direct form. Here, the reaction product catalyzes its own creation in a positive feedback manner, i.e., the product of a pathway strengthens some or all of the initial steps in this pathway. The direct form of autocatalysis occurs at the single-molecule level. Specifically, proteins are synthesized as amino-acid polymers, and other proteins contribute to the folding of this chain towards the active conformation. The active forms of some of these proteins catalyze their own folding and activation.

In general, we call a set of chemical reactions "*collectively autocatalytic*" if the reaction products of some of these reactions are catalysts for sufficient for the other reactions, which assures that the totality of the chemical reactions is self-sustaining, provided that it is supported by energy and food molecules.

The simplest autocatalytic reaction is the second-order reaction

$$X + Y \rightarrow 2Y$$

which has the rate $r = \lambda[X][Y]$.

In the reaction, a molecule of compound $X$ interacts with a molecule of compound $Y$, and the $X$ molecule is converted into $Y$ molecule. Thus, the reaction's product involves the original $Y$ molecule plus the $Y$ molecule produced by the reaction. The time variation of $X$ and $Y$ has the following form:

$$[X] = \frac{[X]_0 + [Y]_0}{\left\{1 + \left(\frac{[Y]_0}{[X]_0}\right) e^{\left([X]_0 + [Y]_0\right)\lambda t}\right\}}$$

$$[Y] = \frac{[X]_0 + [Y]_0}{\left\{1 + \left(\frac{[X]_0}{[Y]_0}\right) e^{-\left([X]_0 + [Y]_0\right)\lambda t}\right\}}$$

The plot of $[Y]$ has the sigmoid form of Fig. 12.5. If the product of a chemical process has a sigmoid form, then there is a high probability that the process is autocatalytic.

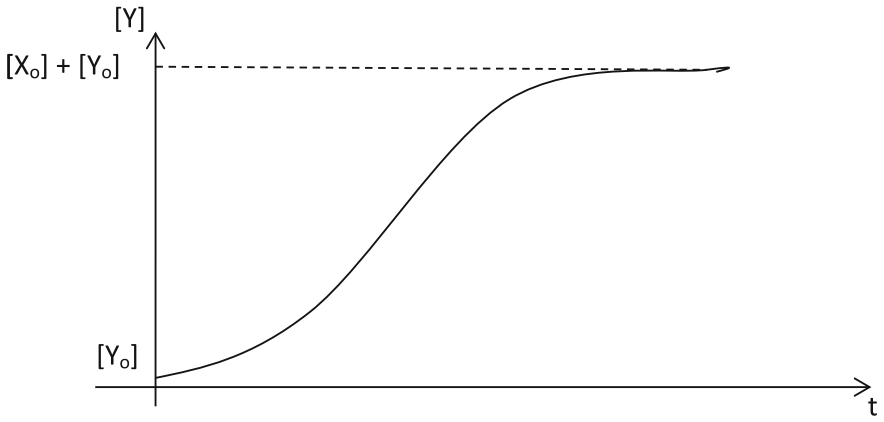



**Fig. 12.5** Sigmoid form of the product concentration Y in the autocatalytic reaction

The sigmoid form is explained as follows. Initially, the reaction rate is small since, at the star, the amount of catalyst is small. Then the rate is progressively increasing as the amount of catalyst increases, and finally the rate is again decreasing due to a decrease of the reactant concentration.

Another idealized autocatalytic reaction model is the following:

$$X + Y \rightarrow 2Y, \quad Y + Z \rightarrow 2Z, \quad Z \rightarrow F$$

In this model, we have a coupled pair of autocatalytic reactions in which the concentration of $X$ is much larger than its equilibrium value. Thus, the forward reaction rate is extremely higher than the reverse reaction rate, which can be neglected. The rate equations of this coupled autocatalytic process are as follows:

$$\frac{\mathrm{d}}{\mathrm{d}t}[Y] = \lambda_1[X][Y] - \lambda_2[Y][Z]$$

$$\frac{\mathrm{d}}{\mathrm{d}t}[Z] = \lambda_2[Y][Z] - \lambda_3[Z]$$

where $\lambda_1, \lambda_2$ and $\lambda_3$ are the rate coefficients of the three reactions. These equations are known as *Lotka–Volterra equations*.

Since the first autocatalysis models of *Lotka* [36], *Monod and Jacob* [37], and *Prigogine* [38], autocatalysis has been applied to model enzymatic and developmental processes (protein synthesis, oncogenesis, and physiology). Some examples of autocatalytic processes are the vinegar syndrome, the binding of oxygen by hemoglobin, the haloform reaction, the biphasic modulation, and the development of sinks. In the following, a short discussion on the last two processes is provided.

**Biphasic modulation** The effects exerted by *immunomodulators* on a dysfunctional immune system can be regarded as a reconstruction of homeostasis or

homeorhesis. One of the mechanisms used to model the biphasic modulation is the exertion mechanism of the peptide preparation Immax A [39]. In [40], the phenomenon of *biphasic modulation* is described as an autocatalytic process. Nonlinear autocatalytic processes contain instabilities that are the subject of catastrophe and singularity theory [41, 42]. Biphasic modulation influences the immunologic functions, in particular to patients with immunological dysfunctions. In this sense, the control of nonspecific response given by the first-line human defense system is related to the modulation of the phagocytic activity of neutrophils. The modulation of *phagocytosis* is actually an autocatalytic process. The changes in phagocytic activity can be determined from photon-counting time series via the classical perturbation coefficient CPC:

$$\mathrm{CPC} = \left(1 - \mathrm{I}_p/\mathrm{I}_n\right)100\%$$

where $\mathrm{I}_p$ and $\mathrm{I}_n$ denote the integrated intensity I, for peptide treated or native (untreated) neutrophils [39].

In this case, CPC represents the reaction rate:

$$\mathrm{CPC}(c) = \left[bc/\left(K + c + gc^2\right)\right] - a$$

where $c = C + c_0$ is the concentration of the peptide preparation, $c_0$ is a correction coefficient for the concentration, $K$ is a Michaelis constant, $a$ and $b$ are constant parameters, and $g$ is an autocatalytic interaction coefficient. The plot of the autocatalytic function CPC($c$), for the in vitro phagocytic process of human neutrophils by the peptide *Immax A* has the form shown in Fig. 12.6 [40].

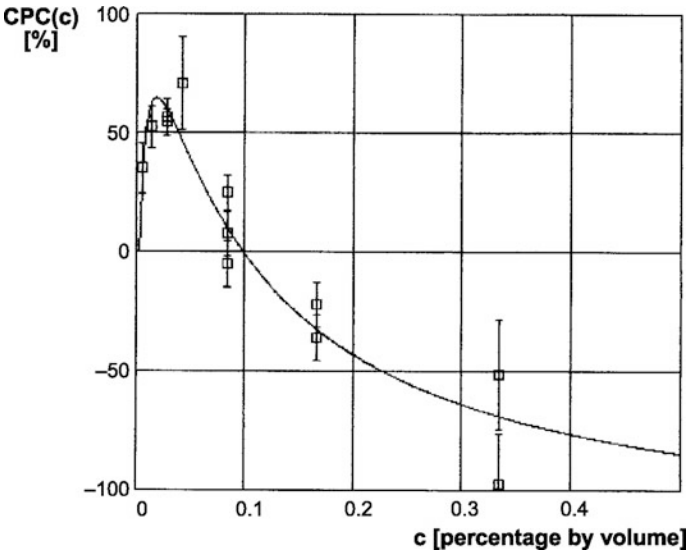



**Fig. 12.6** Biphasic modulation plot $\left(\lim_{c\to\infty} CPC(c) = b - a\right)$

The interaction coefficient g can be considered as a control parameter. In terms of $g$, the evolution of the function CPC($c$) is distinguished in three cases (see Fig. 12.7) namely:

- For $g < 0$, CPC($c$) is nonmonotonic and has a discontinuity in the concentration $c$.
- For $g = 0$, CPC($c$) is monotonic.
- For $g > 0$, CPC($c$) is nonmonotonic and has a maximum and a minimum.

The details of the analysis of this process are given in [40]. The general conclusion is that the autocatalysis model of the first-line human defense can be used to study the control performance in the treatment of immunologic dysfunctions and diseases. This model shows that a quantitative control of the phagocytic activity of neutrophils is presently possible.

**Development of sinks** The differential development of sinks, which depend on a common resource pool, can be regarded as the outcome of an autocatalytic feedback process of flows of resource units into them. The feedback is positive, i.e., the stronger a sink is relative to its competition, the greater is the probability to obtain more resources as a nonlinear function of its resource-drawing ability and sink size. Examples of such developing sinks are tree branches, leaves on the branches, fruits in inflorescences, and seeds in fruits. In [43], the sink strength-dependent model predictions of the subsequent development of the initial asymmetry of growing leaves, when their resource drawing is enhanced, are tested in the following way. The resource-drawing ability of the leaves of Mestha (*Hibiscus cannabinus L.*) is artificially enhanced by external application of growth regulators. It is shown that the results are in agreement with the autocatalytic model of *Ganeshaiah and Uma Shaanker* [44].

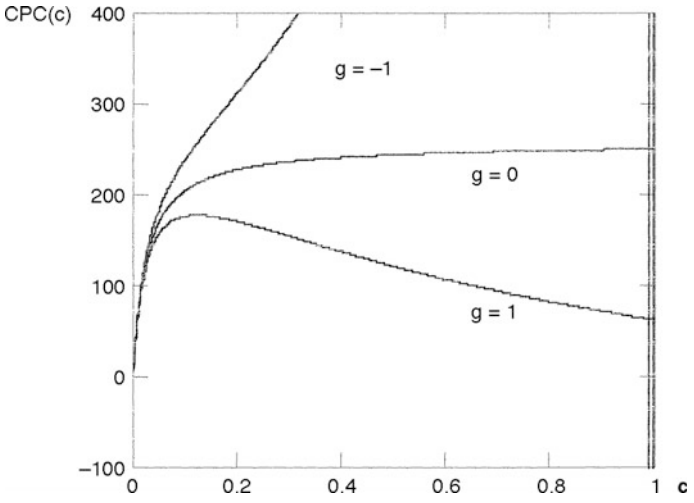



**Fig. 12.7** The three courses of CPC(c) for values $g = -1$, $g = 0$, and $g = 1$ (with $K = 0.017$)

## 12.3 Systems and Control Methods for Biological Processes

In this section, we discuss how systems and control methods have been used (and can be used) for the modeling, analysis, and investigation of biological feedback control mechanisms that are inherent in living systems, such as those considered in Sect. 12.2. The interaction of systems and control methods with biology has naturally led to the emergence of the new combined "*systems biology*" field [2, 45, 46]. Historically, the systems and control field of biological systems was one of the later research interests of *Arnold Tustin*, who recognized the need for a control theoretic foundation of biology as early as 1952 [2, 47, 48]. According to *E. D. Sontag* [45], the systems biology field provides control theorists and engineers a large variety of opportunities and challenges. These challenges can be classified as:

- Contribution of signal processing and system-control techniques to the design of instrumentation for high-accuracy biological measurements and manipulation.
- The use of available system and control techniques (modeling, identification, sensitivity analysis, optimal control, adaptive and robust control, etc.) for the analysis and solutions of problems of great importance and interest to biologists [49].
- Inspiration for new ideas for control and sensor engineering from biological systems and research (e.g., genomic research, digital information coded in DNA, etc.).
- Formulation of entirely new theoretical systems and control problems motivated by systems biology research. Examples of these problems (e.g., cells as I/O systems, monostable/bistable biological systems, etc.) are discussed in [45].

### *12.3.1 System Modeling of Biological Processes*

Biological processes have been modeled by both linear and nonlinear dynamic models. Linear models have been used in very simple cases, but in reality biological processes are nonlinear and need nonlinear modeling.

#### 12.3.1.1 Linear Modeling

Linear modeling will be illustrated by two examples, viz., *enzyme operation* [3] and *biological rhythmic movement* [50].

**Enzyme operation** Consider an enzyme, with intracellular concentration $c$, produced with rate $u_p(c)$ and degraded with rate $u_d(c)$. Under the simple (yet common) assumption that $c$ is uniform within the cell, the evolution of the enzyme concentration is described by the following linear dynamic model:

$$dc(t)/dt = u_p(c) - u_d(c)$$

Two issues that must be studied are the following:

- Steady-state solution
- Sensitivity analysis

The steady-state response is determined by the condition $dc(t)/dt = 0$. Suppose that the enzyme promotes its own production with rate:

$$u_p(c) = Vc/(K_m + c)$$

where $V$ is the maximum production rate and $K_m$ is a *Michaelis–Menten* type constant. Now, assuming that $u_d(c) = \mu_d c$, where $\mu_d$ is the decay constant, the above dynamic model for the enzyme concentration becomes:

$$dc/dt = Vc/(K_m + c) - \mu_d c$$

and the steady-state value of $c$ is found to be $c = V/\mu_d - K_m$. The plots of $u_p(c)$ and $u_d(c)$ versus $c$ are shown in Fig. 12.8. Clearly, the steady state of $c$ corresponds to the intersection of these plots (where $u_p(c) = u_d(c)$ and $dc/dt = 0$).

**Sensitivity analysis** Sensitivity is the opposite of robustness and is expressed by the so-called *sensitivity coefficient*. Consider the parameter $\mu_d$ and the output

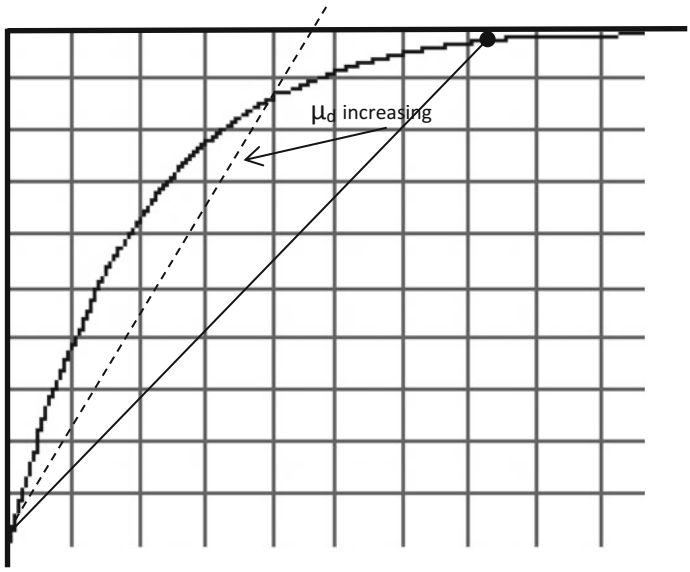



**Fig. 12.8** Linear enzyme concentration model. Increasing $\mu_d$ moves the steady state to a new value

(variable) $c$. The sensitivity coefficient $S_{c,\mu_d}$ of the variation of $c$ corresponding to a variation of $\mu_d$ is defined as

$$S_{c,\mu_d} = \lim_{\Delta\mu_d \to 0} \frac{\Delta c/c}{\Delta\mu_d/\mu_d} = \frac{\mu_d}{c}\frac{\mathrm{d}c}{\mathrm{d}\mu_d} = \frac{\mathrm{d}\,\ln c}{\mathrm{d}\,\ln \mu_d}$$

This formula indicates that $S_{c,\mu_d}$ is equal to the slope of the loglog plot of $c$ versus $\mu_d$. It is now clear that a system is good of its sensitivity coefficient has a magnitude less than 1 (i.e., if the fractional change of the valuable $c$ is smaller than the fractional change of $\mu_d$). Otherwise, the system is not acceptable. In our example, if we use the values $V = 1\ \mathrm{M}/\ \mathrm{sec}, K_m = 1\mathrm{M}$ *and* $\mu_d = 0.1\,\mathrm{sec}^{-1}$, it is found that:

$$S_{c,\mu_d} = -1.11$$

This means that changes of $\mu_d$ around the value $\mu_d = 0.1\,\mathrm{M}$ will be increased by 11%, a really undesired result. Actually, in this example no feasible values of $V$, $K_m$, and $\mu_d$ can give sensitivity coefficient of magnitude less than 1. Here is exactly where the use of negative feedback is beneficial. Indeed, in this example, if $u_p(c)$ itself is proportional to $c$, then $u_d(c) = \mu_d c^2$. This means that there is negative feedback which shows that the enzyme is responsible for its own dilution.

**Biological rhythmic movement** Rhythmic movements are inherent to animal locomotion (walking, crawling, trotting, swimming, etc.). These rhythmic movements are determined by oscillatory neural networks known as *central pattern generators* (**CPG**), the dynamics of which depend upon the connectivity of the network and the nonlinear characteristics of the individual neurons [50]. The animal's movement remains stable in a complex and changing environment through negative feedback based on sensory information. A general study of oscillatory (rhythmic) phenomena in physics and biology is provided in [51], where the emphasis is given to the cell-membrane oscillations modeled by a linear (distributed parameter) wave model. In [52], the role of proprioceptive feedback in the form of position information is studied.

The closed-loop system involves a CPG, a mechanical system, and a sensory system. The CPG drives the mechanical system with its rhythmic electrical patterns. The sensory system detects the position of the mechanical system and sends the information to the CPG via synaptic input. For simplicity, the mechanical system that represents a pair of antagonistic muscles driving a mass was modeled by a low-pass filter and the motor nerves (which actually have a sigmoid characteristic) were modeled by a linear gain. Thus, the resulting mechanical system was a second-order system with transfer function (see Sect. 6.5.5, Fig. 6.10):

$$H(s) = \frac{1/m}{s^2 + 2\zeta\omega_n s + \omega_n^2}$$

The sensors were simplified by assuming the output of each sensor is equivalent to the absolute value of a half-wave rectified version of the position. The feedback synaptic signal (current) $I_{\text{fb}}$ has the form [52]:

$$I_{\text{fb}} = g_{\text{fb}} \tanh(S_{\text{fb}} x_3)[E_{\text{fb}} - V_{\text{SN}}]$$

where $g_{\text{fb}}$ is the maximum conductance, $S_{\text{fb}}$ determines the value at which the conductance saturates, $x_s$ is the output of the sensor, $E_{\text{fb}}$ is the synaptic reversal potential, and $V_{\text{SN}}$ is the membrane potential. The feedback is *ipsilateral inhibition*, so as a while a neuron of the half-center oscillator causes the mass to move in a particular direction, is inhibited by the feedback. This means that the feedback is negative [51]. The general conclusion of this study is that a possible mechanism for the control of biological rhythmic movements is through the mechanical properties of the musculoskeletal system rather than direct use of the neural activity [53, 54]. Given the fact that harmonic/oscillatory behavior is present in biological systems (as in all physical systems), in particular in neurological systems, the frequency-domain control methods were proved useful for studying communication and control in such systems [55, 56].

#### 12.3.1.2 Nonlinear Modeling

Actually, there is not a unique nonlinear model covering all dynamic biological processes. The models differ not only in the type of the nonlinearities involved, but also in the nature of the models, namely:

- Lumped-parameter (ordinary differential equation) models
- Time-delay dynamic models
- Distributed-parameter (partial differential) models
- Stochastic models (of the above forms)
- Integral equation models
- Integrodifferential equation models

As discussed in Sect. 7.11. All these models can be time invariant or time varying.

For example, a general model for a cell-biological system is [3]:

$$\dot{x} = f(x, a)$$

where $x$ is the concentration vector, $f$ describes the reaction rates, and $a$ is a vector of reaction-like constant parameters. An alternative model has the form:

$$\dot{x} = sv(x, a)$$

in which the structural and dynamical parts of the system are separated. Here, $S$ is the *stoichiometric matrix* representing the structure of the chemical reactions, and

$\upsilon(x, a)$ is the vector of reaction rates, corresponding to the dynamics of the system [53, 54]. The identification of the system is performed using the interaction graph of the Jacobian $\partial f/\partial x$. Frequently, in biochemical reaction networks in cells is that their structure, determined by $S$, is more completely known that the exact reaction mechanism that determines the reaction rate $\upsilon(x)$.

**Insulin-Glucose System**

A feedback biological system that has received special attention is the insulin-glucose dynamic system [6, 57–63]. As we discussed in Sect. 12.2.2.3, diabetes is characterized by high blood glucose levels resulting from insufficient metabolization by insulin. Glucose is a monosaccharide produced when digestion breaks down ingested food. It is the main energy source of the body. Insulin is a hormone made by the pancreas and performs the following operations:

- Uptakes glucose from blood by muscle, liver, and fat tissue cells
- Stores glucose in liver
- Regulates the use of fat as an energy source
- Promotes the protein synthesis and general growth of the body

A *minimal model* of glucose kinetics was proposed by *Bergman* [63]. This model involves two differential equations that describe the nonlinear dynamics of the insulin-to-glucose relationship during an **IVGTT** (*Intravenous Glucose Tolerance Test*):

$$\mathrm{d}g(t)/\mathrm{d}t = -p_1[g(t) - g_b] - x(t)g(t), \quad g(0) = p_{\mathbf{0}}$$

$$\mathrm{d}x(t)/\mathrm{d}t = -p_2x(t) + p_3[i(t) - i_b], \quad x(0) = 0$$

where $g(t)$ is the deviation of glucose plasma concentration from its based value $g_b$ (in mg/dl), $x(t)$ is the insulin-excitable glucose uptake rate (in $\mathrm{min}^{-1}$), $i(t)$ is the deviation of insulin plasma concentration from its based value $i_b$ (in μU/ml), $p_1$ and $p_2$ are the glucose kinetic parameter, and $p_3$ is a parameter that influences the insulin sensitivity. This model does not incorporate the effect of the secretion of insulin from the pancreas in response to an elevation in blood glucose concentration. Therefore, it is an open-loop model which can be used through the application of suitable experiment *IVCTT protocols*.

But actually, the glucose metabolism process is a closed-loop system. To express this fact, a third dynamic equation is added and the overall model becomes:

$$\mathrm{d}g(t)/\mathrm{d}t = -p_1[g(t) - g_b] - x(t)g(t), \quad g(0) = g_{\mathbf{0}}$$

$$\mathrm{d}x(t)/\mathrm{d}t = -p_2x(t) + p_3[i(t) - i_b], \quad x(0) = 0$$

$$\mathrm{d}\upsilon(t)/\mathrm{d}t = -\mu\upsilon(t) + \lambda\Theta[g(t)]$$

where $\upsilon(t)$ is the secreted insulin by the pancreatic beta cells in response to an increase in plasma glucose concentration, and $\Theta c[g(t)]$ is the following threshold function [60]:

$$\Theta[g(t)] = \begin{cases} g(t) - \theta, & g(t) \geq \theta \\ 0, & \text{otherwise} \end{cases}$$

Here, $\theta$ represents the glucose concentration above which insulin is secreted.

This model can be used not only for analyzing the insulin-to-glucose dynamics during **IVGTT**, but also for determining the "*insulin sensitivity*" and "*glucose effectiveness*" at individual level.

Besides the above insulin-glucose model, many other complex and effective models have been suggested, e.g., the Degaetano and Arino model [64] and the Sturis–Polonsky–Mosekilde–Van Cauter model [65] for the I*VGTT test*. This test consists of injecting by IV a bolus of glucose and frequently sampling the glucose and insulin plasma concentrations afterward, for a period of about three hours. The DeGaetano and Arino model has a unique, asymptotically stable equilibrium point, and good numerical fitting to individual data as shown in Fig. 12.9. For example, in [60], this feedback-enhanced minimal model was used as a reference model for computing the accuracy and effectiveness of an integral (Volterra-type) model estimated from input–output data. This comparison showed that both models are equivalent. Therefore, the use of data-driven models that do not depend on simplified a priori assumption required in standard compartmental models can be used for preparing effective insulin-glucose protocols.

The Sturis et al. model was developed for modeling the slow (ultradian) oscillations in insulin secretion and studying the reasons for slow oscillations in insulin supply:

$$\frac{\mathrm{d}G}{\mathrm{d}t} = G_{in} - f_2(G(t)) - f_3(G(t))f_4(I_i(t)) + f_5(x_3(t))$$

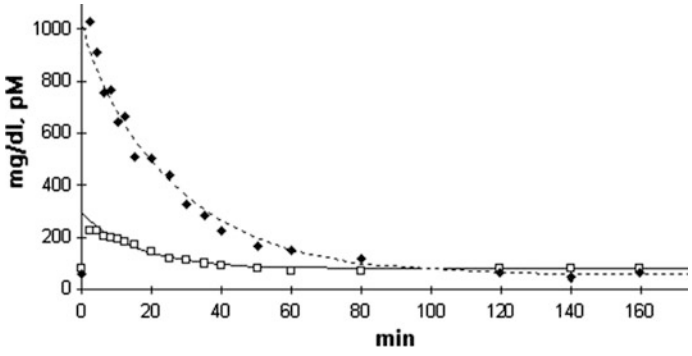



**Fig. 12.9** Time courses of blood glucose (lower plot) and plasma insulin for a patient (solid and dashed lines represent the model, and black/solid squares represent the data). Both curves show a grossly exponential evolution

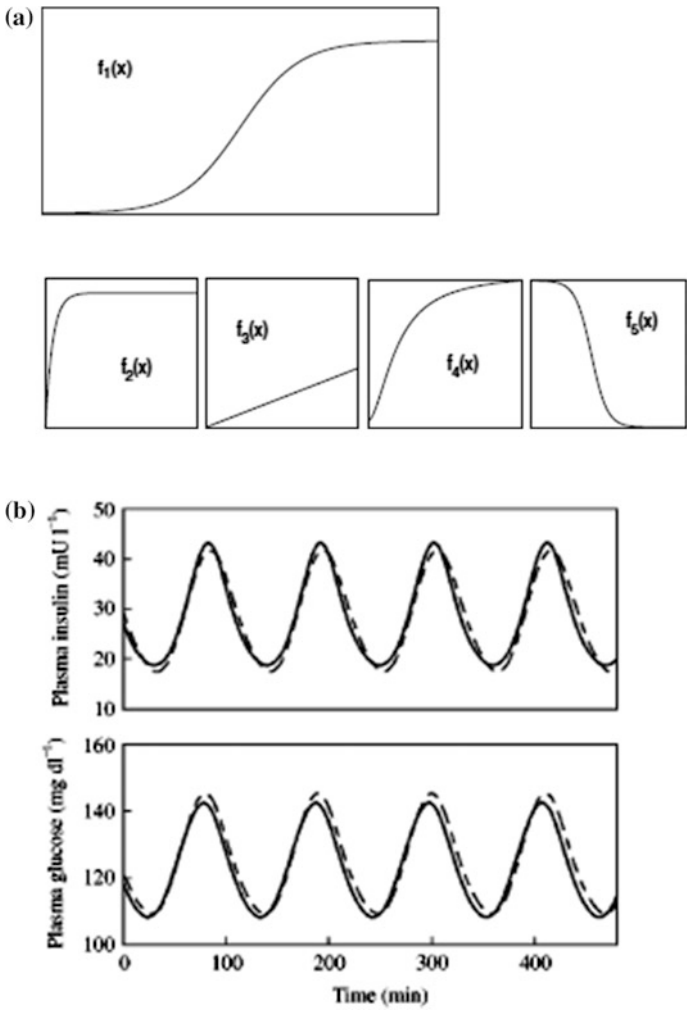



**Fig. 12.10** **a** Shapes of function $f_i$ ($i$ = 1, 2, 3, 4, 5), **b** Simulation plots with glucose infusion rate of 216 mg/min

$$\frac{\mathrm{d}I_p}{\mathrm{d}t} = f_1(G(t)) - E\left(\frac{I_p(t)}{V_p} - \frac{I_i(t)}{V_i}\right) - \frac{I_p(t)}{t_p}$$

$$\frac{\mathrm{d}I_i}{\mathrm{d}t} = E\left(\frac{I_p(t)}{V_p} - \frac{I_i(t)}{V_i}\right) - \frac{I_i(t)}{t_i}$$

$$\frac{\mathrm{d}x_1}{\mathrm{d}t} = \frac{3}{t_d}\left(I_p(t) - x_1(t)\right)$$

$$\frac{dx_2}{dt} = \frac{3}{t_d}(x_1(t) - x_2(t))$$

$$\frac{dx_3}{dt} = \frac{3}{t_d}(x_2(t) - x_3(t))$$

where $I_p$ and $I_i$ is the insulin in plasma and intercellular space, respectively, and $x_1, x_2, x_3$ represent the delayed effect of insulin on hepatic glucose production with time $t_d$. The shapes of the functions $f_i$ $(i = 1, 2, 3, 4, 5)$ are as shown in Fig. 12.10a [66, 67].

Numerical analysis revealed that oscillations are due to a bifurcation in the model. The oscillations depend on the hepatic glucose production. Self-sustained oscillations occur when the hepatic glucose time delay is in the range 25–30 min (see Fig. 2.10b) [68]. Increasing the rate of infusion does not affect the frequency of oscillations, but it results in an increase in their amplitude. A comprehensive review of the available insulin-glucose models is provided in [59], including time-delay and distributed-parameter models and the available software.

An optimal controller methodology in $H_2/H_\infty$-space was presented in [69]. This includes a linear quadratic controller and a disturbance-rejection linear quadratic (minimax) controller. In [65], a neural network technique is provided for a multilayer (MLN) feedforward network and a polynomial network (PN), with Levenberg–Marquadt (LM) learning. The MLN model was proved to be superior and suitable for use as a guide for designing insulin patient-treatment protocols.

**Cardiovascular-respiratory system** A very important biological system for human life is the cardiovascular-respiratory system. This system has been extensively studied both by pure biological/medical techniques and system-control mathematical methods. A few studies of the respiratory system are presented in [70–74], and some mathematical models of the coupled cardiovascular-respiratory control systems are presented in [75–79].

Here we will briefly review the model presented in [78]. This model consists of the following 13 first-order differential equations:

$$V_{\downarrow}(A_{\downarrow}(CO_{\downarrow}2))P_{\downarrow}(a_{\downarrow}(CO_{\downarrow}2))(t) = 863F_{\downarrow}p(t)(C_{\downarrow}(u_{\downarrow}(CO_{\downarrow}2))(t) - C_{\downarrow}(a_{\downarrow}(CO_{\downarrow}2))(t))$$

$$\begin{aligned} V_{\downarrow}(A_{\downarrow}(O_{\downarrow}2))P_{\downarrow}^{\cdot}(a_{\downarrow}(0_{\downarrow}2))(t) = {} & 863F_{\downarrow}p(t)(C_{\downarrow}(u_{\downarrow}(0_{\downarrow}2))(t) - C_{\downarrow}(a_{\downarrow}(0_{\downarrow}2))(t)) \\ & + (V_{\downarrow}A)^{\cdot}(t)(P_{\downarrow}(t)) \end{aligned}$$

$$V_{T_{CO_2}}\dot{C}_{u_{CO_2}}(t) = MR_{CO_2} + F_s(t)\left(C_{a_{CO_2}}(t) - C_{u_{CO_2}}(t)\right).$$

$$V_{T_{O_2}}\dot{C}_{u_{O_2}}(t) = -MR_{O_2} + F_s(t)\left(C_{a_{O_2}}(t) - C_{u_{O_2}}(t)\right).$$

$$C_{\text{as}}\dot{P}_{\text{as}}(t) = Q_l(t) - F_s(t).$$

$$c_{\text{us}}\dot{P}_{\text{us}}(t) = F_s(t) - Q_r(t).$$

$$c_{\text{up}}\dot{P}_{\text{up}}(t) = F_p(t) - Q_l(t)$$

$$\dot{S} = \sigma_l(t).$$

$$\dot{S}_r = \sigma_r(t).$$

$$\dot{\sigma}_l = -\gamma_l\sigma_l(t) - a_lS_l(t) + \beta_lH(t).$$

$$\dot{\sigma}_r = -\gamma_r\sigma_r(t) - a_rS_r(t) + \beta_rH(t).$$

$$\dot{H}(t) = u_1(t).$$

$$\ddot{V}_A(t) = u_2(t).$$

where the meaning of the symbols is shown in Tables 12.1 and 12.2.

**Table 12.1** Respiratory parameters and variable

| Symbol | Meaning | Unit |
|---|---|---|
| $C_aco_2$ | Concentration of bound and dissolved $CO_2$ in arterial blood | $l_{STTPD}\ l^{-1}$ |
| $C_ao_2$ | Concentration of bound and dissolved $O_2$ in arterial blood | $l_{STTPD}\ l^{-1}$ |
| $C_vco_2$ | Concentration of bound and dissolved $CO_2$ in mixed venous blood entering the lungs | $l_{STTPD}\ l^{-1}$ |
| $C_vo_2$ | Concentration of bound and dissolved $O_2$ in the mixed venous blood entering the lungs | $l_{STTPD}\ l^{-1}$ |
| $MRco_2$ | Metabolic $CO_2$ production rate | $l_{STTPD}\ min^{-1}$ |
| $MRo_2$ | Metabolic $O_2$ consumption rate | $l_{STTPD}\ min^{-1}$ |
| $P_aco_2$ | Partial pressure of $CO_2$ in arterial blood | mmHg |
| $P_Ao_2$ | Partial pressure of $O_2$ in arterial blood | mmHg |
| $P_vco_2$ | Partial pressure of $CO_2$ in mixed venous blood | mmHg |
| $P_vo_2$ | Partial pressure of $O_2$ in mixed venous blood | mmHg |
| $P_I$ | Partial pressure of inspired gas | mmHg |
| B | Brain compartment | – |
| $u_2$ | Control function $u_2 = \bar{V}_A$ | $l_{BTPS}\ min^{-2}$ |
| $\dot{V}_A$ | Alveolar ventilation | $l_{BTPS}\ min^{-1}$ |
| $\ddot{V}_A$ | Time derivative of alveolar ventilation | $l_{BTPS}\ min^{-2}$ |

(continued)

**Table 12.1** (continued)

| Symbol | Meaning | Unit |
|---|---|---|
| $V_{ACO_2}$ | Effective $CO_2$ storage volume of the lung compartment | $l_{BTPS}$ |
| $V_{AO_2}$ | Effective $O_2$ storage volume of the lung compartment | $l_{BTPS}$ |
| $V_{TCO_2}$ | Effective tissue storage volume for $CO_2$ | l |
| $V_{TO_2}$ | Effective tissue storage volume for $O_2$ | l |
| $l_p, l_c$ | Cut-off thresholds | mmHg |

**Table 12.2** Cardiovascular parameters and variables

| Symbol | Meaning | Unit |
|---|---|---|
| $a_l$ | Coefficient of $S_l$ in the differential equation for $\sigma_l$ | $\text{min}^{-2}$ |
| $a_r$ | Coefficient of $S_r$ in the differential equation for $\sigma_r$ | $\text{min}^{-2}$ |
| $A_{\text{pesk}}$ | $R_s = A_{\text{pesk}} C_{v_{O_2}}$ | mmHg min $\text{l}^{-1}$ |
| $\beta_l$ | Coefficient of $H$ in the differential equation for $\sigma_l$ | mmHg $\text{min}^{-1}$ |
| $\beta_r$ | Coefficient of $H$ in the differential equation for $\sigma_r$ | mmHg $\text{min}^{-1}$ |
| $C_{\text{as}}$ | Compliance of the arterial part of the systematic circuit | l $\text{mmHg}^{-1}$ |
| $C_{\text{ap}}$ | Compliance of the arterial part of the pulmonary circuit | l $\text{mmHg}^{-1}$ |
| $C_{\text{us}}$ | Compliance of the venous part of the systematic circuit | l $\text{mmHg}^{-1}$ |
| $C_{\text{up}}$ | Compliance of the venous part of the pulmonary circuit | l $\text{mmHg}^{-1}$ |
| $F_p$ | Blood flow perfusing the lung compartment | l $\text{min}^{-1}$ |
| $F_s$ | Blood flow perfusing the tissue compartment | l $\text{min}^{-1}$ |
| $H$ | Heart rate | $\text{min}^{-1}$ |
| $\gamma_l$ | Coefficient of $\sigma_l$ in the differential equation for $\sigma_l$ | $\text{min}^{-1}$ |
| $\gamma_r$ | Coefficient of $\sigma_r$ in the differential equation for $\sigma_r$ | $\text{min}^{-1}$ |
| $P_{\text{as}}$ | Mean blood pressure in arterial region: systematic circuit | mmHg |
| $P_{\text{ap}}$ | Mean blood pressure in arterial region: pulmonary circuit | mmHg |
| $P_{\text{vs}}$ | Mean blood pressure in venous region: systematic circuit | mmHg |
| $P_{\text{vp}}$ | Mean blood pressure in venous region: pulmonary circuit | mmHg l |
| $Q_l$ | Left cardiac output | l $\text{min}^{-1}$ |
| $Q_r$ | Right cardiac output | l $\text{min}^{-1}$ |
| $R_p$ | Resistance in the peripheral region of the pulmonary circuit | mmHg $\text{min}^{-1}$ |
| $R_s$ | Peripheral resistance in the systematic circuit | mmHg $\text{min}^{-1}$ |
| $S_l$ | Contractility of the left ventricle | mmHg |
| $S_r$ | Contractility of the right ventricle | mmHg |
| $\sigma_l$ | Derivative of $S_l$ | mmHg $\text{min}^{-1}$ |
| $\sigma_r$ | Derivative of $S_r$ | mmHg $\text{min}^{-1}$ |
| $u_l$ | Control function, $u_1 = H$ | $\text{Min}^{-2}$ |
| $V_{\text{str},l}$ | Stroke volume of the left ventricle | l |
| $V_{\text{str},r}$ | Stroke volume of the right ventricle | l |
| $V_0$ | Total blood volume | l |

The respiratory part of the system involves the first four equations and is modeled using two subsystems the lung compartment (the first two equations) and a general tissue compartment (the next two equations). The lung compartment represents mass balance for $CO_2$ and $O_2$, and the third and fourth equations are the state equations for $CO_2$ and $O_2$, respectively. The fifth and sixth equations describe the blood-mass balance in the systemic artery and vein components. The seventh equation stands for the balance in the pulmonary-vein component. This model is indeed quite complex and will not be fully explained here. The relation of blood flow F and blood pressure (fifth through seventh equations) are given by the hydraulic resistance (Ohm's) law:

$$F_s(t) = \frac{[P_{as}(t) - P_{us}(t)]}{R_s(t)}$$

$$P_{ap}\big(u - P_{up}(t)\big]/R_p$$

where $P_a$ is the arterial blood pressure, $P_u$ is the venous pressure, and $R$ is vascular resistance. The cardiac output $Q$ is given by the mean blood flow over the length of a pulse, i.e.,:

$$Q(t) = H(t)V_{\text{str}}(t)$$

where $H$ is the *heart rate*, and $V_{\text{str}}$ is *stroke volume*. The relation between $V_{\text{str}}$ and blood pressure is:

$$V_{\text{str}} = S(t)[cP_u(t)/P_a(t)]$$

where $S$ is the contractibility, $P_u$ is the venus filling pressure, $P_a$ is the arterial blood pressure opposing the ejection of blood, and c is the compliance of the relaxed ventricle.

The coupling of the respiratory and cardiovascular subsystems is approximately modeled by the equation:

$$R_s(t) = A_{\text{pesk}}C_{\text{uo}_2}(t)$$

where $A_{\text{pesk}}$ is a constant. This equation is actually a basic local control mechanism for changing vascular resistance. The influence of heart rate $H$ and ventilation rate $\dot{V}_A$ on the system is imposed via the control signals $u_1$ and $u_2$.

In [78] a stabilizing controller was designed to drive the cardiovascular and respiratory system from a steady state "initial disturbance" to a steady state "final equilibrium".

The system model can be expressed in compact form by the state-space equation:

$$\dot{x}(t) = g(x,t) + \mathrm{B}\ u(t), \quad x(0) = x^0$$

which is linear in the control vector $u(t)$, where

$$x(t) = \left(P_{a_{\mathrm{co}_2}}, P_{a_{\mathrm{co}_2}}, C_{u_{\mathrm{co}_2}}, C_{u_{\mathrm{o}_2}}, P_{\mathrm{as}}, P_{\mathrm{us}}, P_{\mathrm{up}}, S_l, S_r, \sigma_l, \sigma_r, H, \dot{V}_A\right)^T$$

The objective of the control was to minimize the cost functional:

$$\int_0^\infty \left(q_{\mathrm{as}}\left(P_{\mathrm{as}}(t) - P_{\mathrm{as}}^f\right)^2 + q_c\left(P_{a_{\mathrm{co}_2}}(t) - P_{a_{\mathrm{co}_2}}^f\right)^2 + q_0\left(P_{a_{o2}}(t) - P_{a_{\mathrm{o}_2}}^f\right)^2 + q_1 u_1(t)^2 + q_2 u_2(t)^2\right)\mathrm{dt}$$

where $u_1(t)$ is the heart rate variation $\dot{H}$ and the ventilation rate $\ddot{V}_A$, i.e.,:

$$u = [u_1(t), u_2(t)]^T = \left[\dot{H}, \ddot{V}_A\right]^{\mathrm{T}}$$

The control law considered has the linear feedback state-vector form:

$$u(t) = -F_m x(t)$$

where $F_m$ is the feedback gain matrix. The matrix $F_m$ was computed for the linearized system around the final state $x_f$, and so the control is suboptimal for the full nonlinear system.

Omitting the details (see [78]), this model and controller provide satisfactory predictions for *the 4NREM sleep* state in humans starting from the initial *awake* state. This model has also been used for the transfer problem from *rest* to *non-aerobic exercise* with very good results.

## 12.4 Feedback Control in Society

### *12.4.1 General Issues*

Society involves three types of feedback control systems:

- *Hard systems*, i.e., physical and technological systems that can be accurately described by mathematical dynamic models,
- Soft systems, i.e., behavioral and social systems,
- *Hard–soft systems,* which involve the biological processes of life.

The *hard–soft* (biological) systems were studied in Sects. 12.2 and 12.3. Here we will consider four representative examples of hard systems and two classes of soft systems, namely managerial systems and economic systems. Another class of soft systems is the so-called class of *social-control systems*. This class refers to the societal mechanisms used to regulate and control individual and group behavior so

as to conform to the rules of a society or social group. Sociology classifies social controls in two main types:

- Internalization of norms and values
- External ratifications, which can provide positive feedback (reward) or negative feedback (punishment)

*Social control* may be either formal or informal [80, 81]. *Informal social control* is manifested by the social values that are present in individuals (e.g., ethical values and religious values). Traditional societies use mainly informal social control means involved in their particular cultures by which social order is established. *Formal social control* is imposed via laws, rules, and regulations enforced by government and state organizations authorized to enforce formal sanctions, such as fines and imprisonment [80]. Sociologists agree that informal social control is fundamental for keeping social order, but they also recognize the need of formal society control means as societies become more complex. According to *Edward Ross*, belief systems can more strongly influence and control human behavior than laws and rules imposed by the government.

Today, large-scale hard technological/industrial systems contain not only the hard technological issues, but also soft managerial and economic processes that are studied by principles and techniques borrowed from the class of soft systems. On the other hand, many soft systems are studied via a combination of human-behavior concepts and mathematical models by which their investigators attempt to quantify the interrelations and dynamic evolution of the various subprocesses involved.

## *12.4.2 Hard Technological Systems*

The following four categories of hard technological systems will be investigated:

- Process control systems
- Manufacturing control systems
- Aircraft flight and traffic-control systems
- Robotic systems

### 12.4.2.1 Process Control Systems

*Process control* is a term that mainly embraces the control of continuous physical and chemical systems (e.g., the food processing industry, oil and gas refineries, room-temperature control, plastics industry, power generation utilities, etc.) [82–88]. A chemical process industry example is shown in Fig. 12.11.

Complex process control systems involve a large number of controlled variables and control inputs or manipulated variables. Actually, a large number of interacting/coupled control loops exist that can be designed by using multivariable control

**Fig. 12.11** View of a chemical process industry

techniques, such as those discussed in Chaps. 6 and 7. Process control installations contain various components, such as control valves, prime movers, general-purpose controllers, and sensors that measure the system variables (flow sensors, pressure sensors, temperature sensors, etc.). The overall structure of the control system has the general form shown in Fig. 6.4, of course with particularities characterizing each system. The controllers typically used in industry belong to the following categories:

- Two-term and three-term controllers (**PIs**, **PIDs**)
- Programmable logic controllers (**PLCs**)
- Supervisory controllers (e.g., **SCADA**)
- Embedded controllers (**EC**s)
- Distributed controllers (**DCs**)
- Networked controllers (**NCs**)
- Special-purpose sophisticated/intelligent controllers (**ICs**)

*Distributed controllers* are usually used for single-point processing over a small geographic area. *SCADA* (supervisory control and data acquisition) systems are used for large-scale and distributed-management control. *Embedded controllers* are central computer-embedded controllers (hardware and software) that oversee, manage, and monitor the system operation and especially secure its safe operation. *Networked control* is the control achieved through the communication of several control computers (controllers), sometimes separated by very long distances. Here the methods of computer communications are employed (Sect. 4.3.6). Special-purpose and intelligent controllers include predictive controllers, adaptive controllers, knowledge-based controllers, etc.

Process control systems are distinguished in:

- **Continuous** (represented via smooth and uninterrupted variables in time)

- **Discrete** (production or manipulation of discrete products, e.g., metal stamping, care spare-parts, etc.). Discrete process control is the subject of discrete manufacturing control, which is discussed in the next paragraph
- **Batch** (combination of specific quantities of raw materials to produce a quantity of end products)
- **Hybrid** (systems that have elements of more than one of the above types)

The technology layers of process control systems involve the following in a top-down hierarchy:

- Plant and information management
- Events, alarms, and risk management
- Operator supervision and control
- Executive control (actuators and control logic)
- Data highway (connectivity)
- Sensory level (input/output interface)

#### 12.4.2.2 Discrete Manufacturing Control Systems

*Manufacturing systems* constitute a major class of technological systems that need the synergy of techniques and tools from such disciplines as sensor technology, control engineering, management science, and computer and information technology. New discrete manufacturing systems are continuously being designed, tested, and put into operation via these technologies [89–92]. In all cases, the minimal set of the goals of a computer-aided manufacturing system include the following:

- Increased product quality
- Increased worker satisfaction
- Increased customer satisfaction
- Increased productivity
- Increased flexibility
- Decreased lead times
- Decreased costs
- Guaranteed reliability and fault tolerance

These requirements are met through a suitable synergy of the technological and human resources, as is the case in every management control system.

The manufacturing processes involve the following top-down hierarchical operations:

- Product planning at the organization level
- Product design and hardware/software assignment and scheduling at the coordination level
- Product generation at the execution level
- Machine control and actuators

The algorithms and procedures at the various levels can be modified according to the nature of the manufacturing system and the tastes of the designer. Robots and numerical machines are an integral part of discrete manufacturing systems.

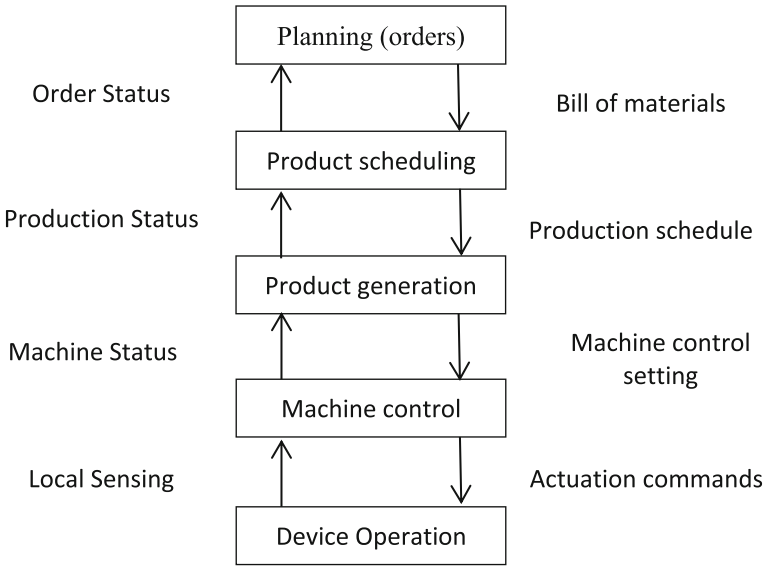



**Fig. 12.12** Hierarchical manufacturing control (Commands from top to bottom, feedbacks from bottom to top)

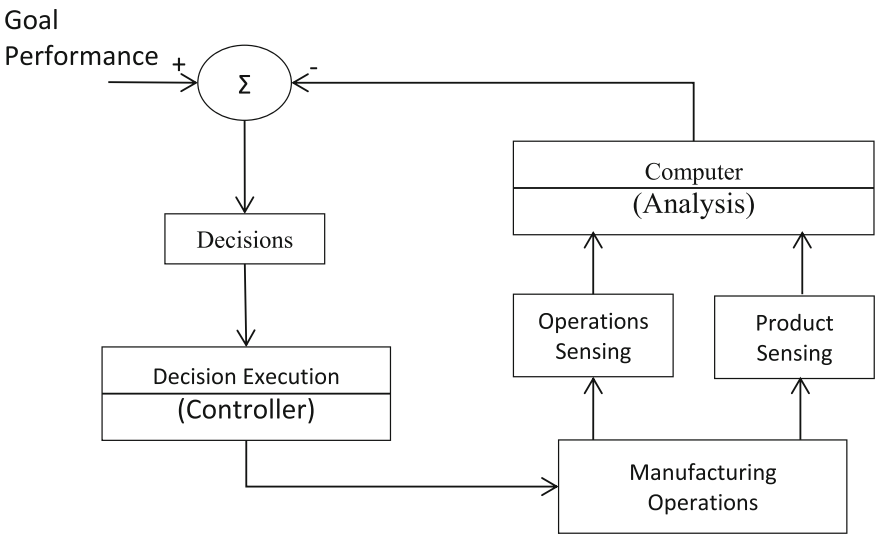



**Fig. 12.13** Standard manufacturing feedback control loop

The above operations are in general interlinked as shown in Fig. 12.12 [93].

A typical feedback control loop in a manufacturing system has the form shown in Fig. 12.13, which is self-explanatory.

Information technology helps in many ways to improve the competitive advantage of a company employing *computer integrated manufacturing* (**CIM**). The three major ways are:

- By changing the structure of the company to create new rules of competition
- By creating competitive advantage through the development of new ways of outperforming the company's competitors
- By creating an entirely new business, often from within a company's existing operations.

The feedback from any lower level to its superior level(s), implies a systematic approach to the overall operation of a manufacturing company which involves plant functions, production functions, business functions, and administrative functions. The interfaces of the above functions (activities) with the CIM system are work-stations or interactive terminals for the people and instrumentation for the equipment.

Computer integrated manufacturing involves several subfields (processes) that are interconnected in a tree representation as shown in Fig. 12.14.

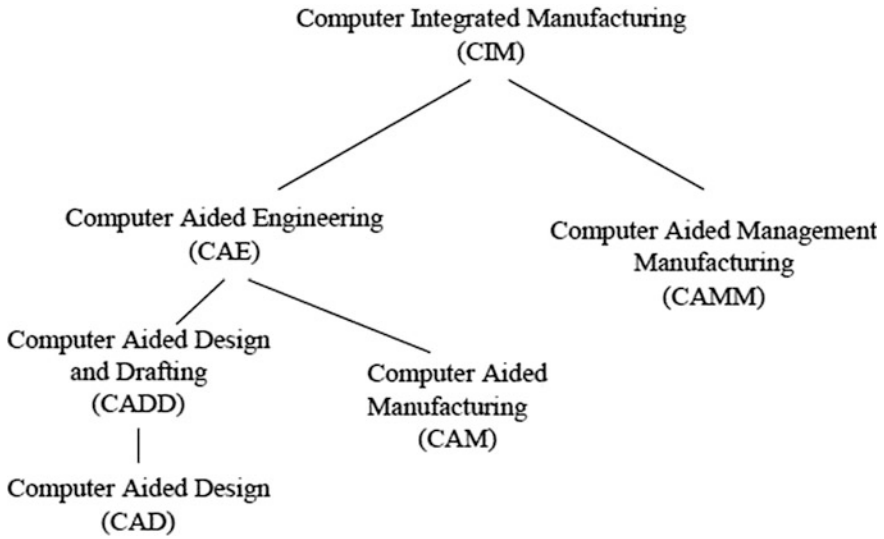
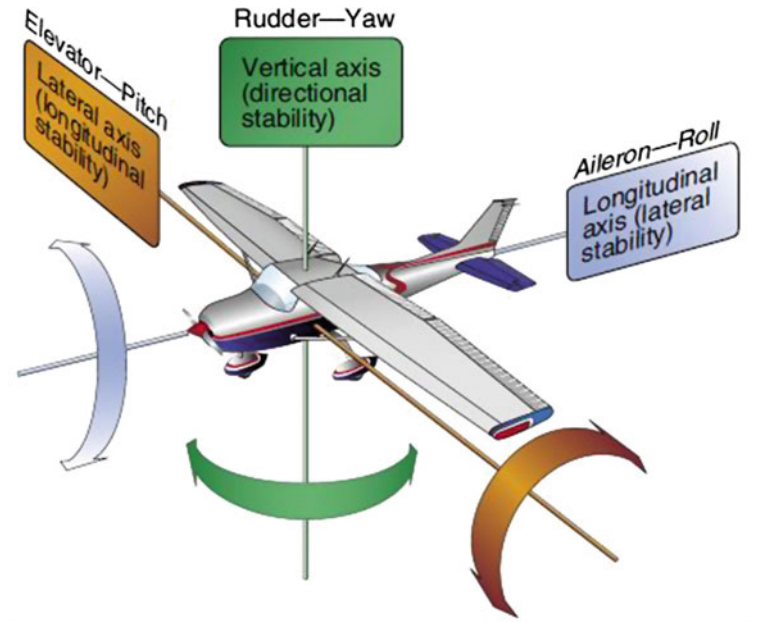



**Fig. 12.14** Subfields of computer integrated manufacturing. http://collections.infocollections.org/ukedu/collect/ukedu/index/assoc/h2398e/p18_55.gif

### 12.4.2.3 Aircraft Flight and Traffic-Control Systems

*Aircraft flight-control systems* play a very important role in our modern society [94–97]. The primary variables that are controlled in aircraft systems are roll, pitch, and yaw. The control is implemented by a set of mechanical and electronic equipment that enables the aircraft to be flown with exceptional accuracy and reliability. The controller consists of cockpit control, sensors, prime movers (mechanical, hydraulic, electric) and control computers. The aircraft is steered by a system of flaps called *control surfaces*, namely, *ailerons*, *elevators*, and *rudder*, which are moved by a control stick and pedals. The ailerons control the *roll angle*, the *elevators* control the *pitch angle*, and the *rudder* turns the *yaw angle*. These three rotations and the corresponding stability types are illustrated in Fig. 12.15.




| Primary Control Surface | Airplane Movement | Axes of Rotation | Type of Stability |
|---|---|---|---|
| Aileron | Roll | Longitudinal | Lateral |
| Elevator/ Stabilator | Pitch | Lateral | Longitudinal |
| Rudder | Yaw | Vertical | Directional |

**Fig. 12.15** Roll, pitch, and yaw. The three possible rotation ways of an aircraft and the corresponding stability types. *Source* www.aboutflight.com/up-content/uploads/2011/08/5-4.png

The control surfaces cannot be controlled by the pilot directly (there are too many), but via a *flight-control system* which receives simple inputs from the pilot and selects the surfaces to move and the amount of movement. This movement is performed by control prime actuators (hydraulic or electric movers or motors) in response to the movement signals provided from the flight controller connected to the pilot's control yoke or stick [97].

The changes of position, direction, and speed are measured by proper sensors (gyroscopes, accelerometers) (Fig. 12.16a, b). The outputs of the sensors are sent to the control computer which, according to the control laws adopted, computes the control commands and sends them to the actuators.

All the above elements are shown collectively in the overall aircraft flight feedback control system of Fig. 12.16c [97].

Commercial aircraft (Fig. 12.17a) have received the greatest attention for reasons of comfort and safety of the human passengers. Unfortunately, despite the sophisticated cockpit automation, such as the *flight management system* (**FMS**), the pilot has a high workload and may take erroneous actions. For this reason, much research and developmental efforts are concerned with the flight-deck operations, which continuously lead to new modes for the pilot to understand, in order to minimize the confusion over states and modes in automated cockpits [98, 99].

The modern method for *air-traffic control* (**ATC**) is by using radio-transmitting and receiving equipment for the pilot-controller communication. To eliminate so-called *navaid interference,* the aircraft transmitters used a different frequency than the ground-based navaids. This two-frequency system is called a "*duplex communication system*". The radio-frequency bands allocated to aeronautical communications are determined by international agreements. These frequency bands exist mainly in *high frequency* (**HF**), *very-high frequency* (**VHF**), and *ultra-high frequency* (**UHF**). Due to some drawbacks, the duplex system was abandoned, and today all **ATC**s worldwide use the "*simplex system*", which enables pilots to communicate with controllers using one discrete frequency (Fig. 12.17b) [100].

#### 12.4.2.4 Robotic Systems

Robotic systems are systems that give more flexibility to industrial production and can be used in a large variety of societal applications. They are advanced automation systems which use the electronic computer as their basic feedback control element. The term 'robot' (robota) was coined by the Czech writer Karel Capek (1921) and means 'forced labor'.

The industrial robotic period started in the USA with the development of the "Unimate robot" put into operation by the robot company UNIMATION (from UNIversal autoMATION) in 1961. This company was founded by George Devol, Jr. and Joseph Engelberger. Further historical landmarks of robotics include the development at Stanford of the first mobile robot "Shakey" (1970), the Stanford Cart (1979), the HONDA "humanoid" walking robot ASIMO, the "Roomba" robot (a vacuum cleaner), the German Aerospace Company's "Rollin Justin" robot (a

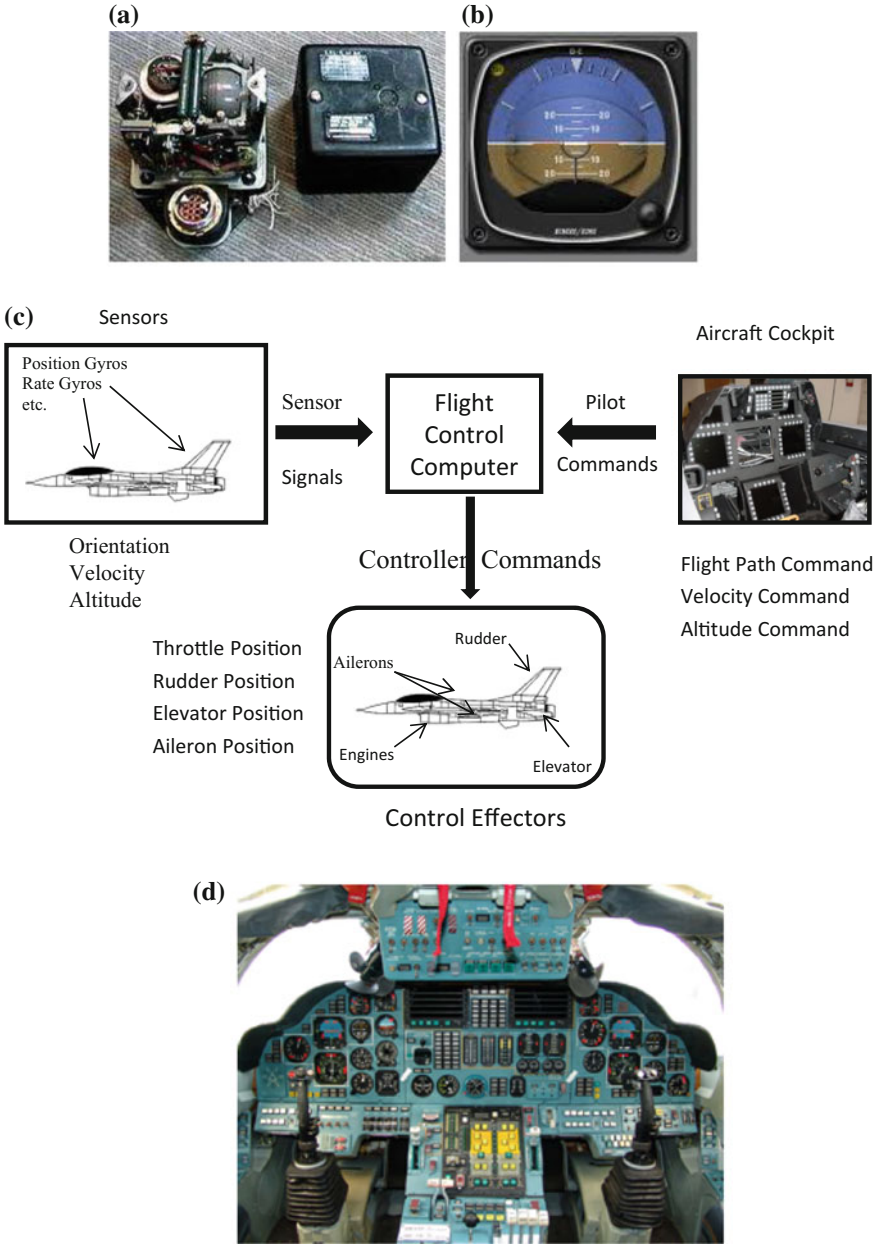



**Fig. 12.16** **a** Gyrocompasses used in ships and aircrafts; **b** Artificial horizon-autopilot (shows the pitch of the aircraft); **c** Overall structure of aircraft flight-control system; **d** A typical aircraft cockpit

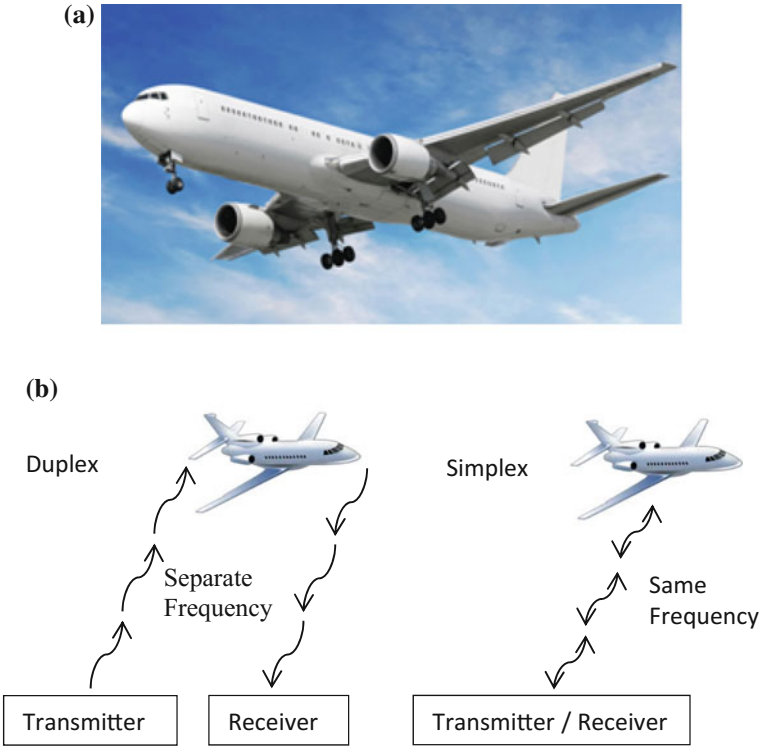



**Fig. 12.17** **a** A modern commercial aircraft, **b** Duplex and **s**implex transmission in ATC systems

robot that can prepare and serve drinks), the robot Pioneer 3-DX (a fully programmable, multifunction, mobile robot), and multi-legged robots (e.g., the quadruped robot "Kotetsu", the hexapod robot spider, etc.). Current generation robots possess "intelligence", i.e., as we say, they are "intelligent robots", being able to think and make decisions. They assume this intelligence via several methods of "artificial Intelligence" (AI) and "sensing".

Today, there is a wonderful world of robots that can move, walk, see, speak, express emotions, and perform dexterous and delicate intelligent tasks. This world is continuously expanding to satisfy the industrial, service, medical, assistance, companionship, and entertainment requirements of our society. Technologically, the design and construction of modern robots are based (non-exhaustively) on mechanics (kinematics, dynamics), sophisticated feedback control (nonlinear, adaptive, robust, intelligent methods), electronics and power amplification, computation, artificial and computational intelligence, sensing, and signal processing [101–103].

The principal categories of modern robots are: fixed-place industrial robots or robotic manipulators (Fig. 12.18a), mobile robots and manipulators (Fig. 12.18b), which can move around their environment, medical robots, which are divided into

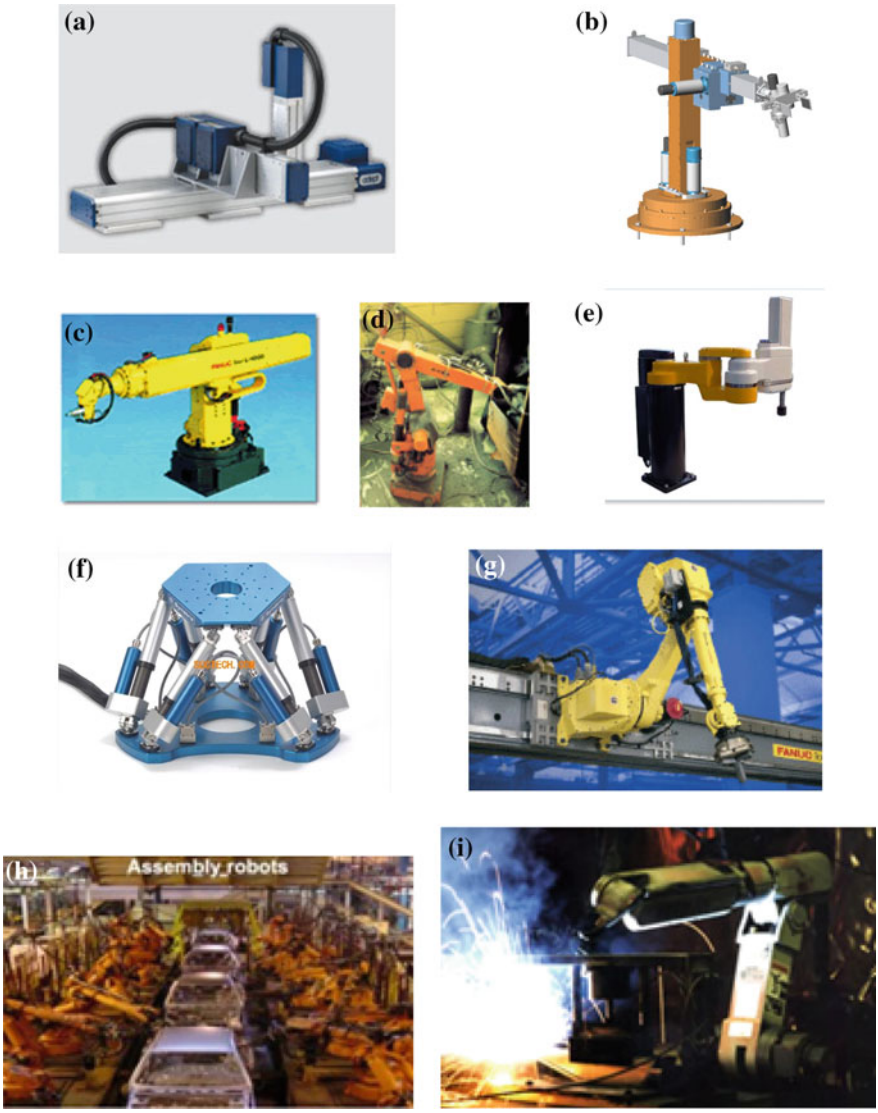



**Fig. 12.18** Examples of industrial robots: **a** Cartesian, **b** cylindrical, **c** spherical, **d** articulate, **e** SCARA, **f** parallel/Stewart, **g** gantry robot, **h** articulated assembly robots at work, **i** welding robot at work. *Source* http://www.roboticsbible.com/wp-content/uploads/2011/10/assembly-robot-300x154.jpg http://www.defiancemetal.com/Images/Equipment/Robot_weld.jpg

"macrorobots" (surgical robots, rehabilitation robots) and "microrobots" (image-driven robots for angioplasty operation, brain aneurism repair, etc.), telerobots which can perform non-repetitive distant operations (e.g., telesurgery, telemonitoring, and submarine or space operations), service robots which perform


**Fig. 12.19** Examples of mobile robots: **a** Pioneer-3 differential drive, **b** tricycle, **c** car-like, **d** omnidirectional/universal drive, **e** omnidirectional/mecanum drive, **f** omnidirectional/synchro drive, **g** skid steering/tracked robot

several home and domestic operations (e.g., autonomous floor cleaning, garbage collection), social or socialized robots for emotional interaction of impaired and elderly persons, therapeutic assistance, and entertainment, and war robots for semiautonomous or autonomous combat (unmanned aircrafts, autonomous weapons, etc.). Figures 12.18, 12.19, 12.20, 12.21, 12.22 and 12.23 give some examples of industrial robotic manipulators and medical, assistive, and socialized robots.

**Fig. 12.20** A surgical robot at work

**Fig. 12.21** Autonomous robotic wheelchairs

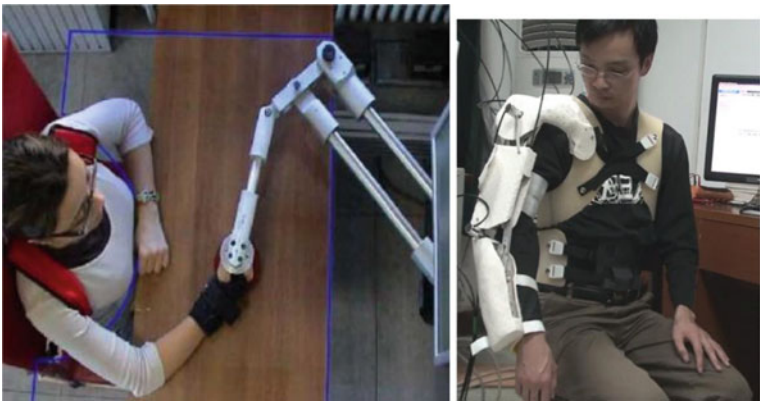


**Fig. 12.22** Rehabilitation robots

**Fig. 12.23** KASPAR humanoid socialized robot interacting with an autistic child

### *12.4.3 Soft-Control Systems*

The following soft-control systems, which primarily operate on a human-behavioral level, will be discussed:

- Management control systems
- Economic control systems

#### 12.4.3.1 Management Control Systems

The term *management control system* (**MCS**) was coined by *Robert N. Anthony* [104] as a system that collects and employs information to evaluate the performance of various resources involved in organizations. These resources include the following:

- Physical/technological resources
- Economic resources
- Human resources

An **MCS** considers the overall organization as an integrated system and includes the study of the influence of the resources in implementing the organization's strategies toward the achievement of its goals. **MCSs** always involve decision-making methods and procedures and optimization algorithms [105–114]. The concepts, subproblems, and techniques of the general decision-making and control design of management systems are shown in block diagram form in Fig. 12.24. Decision-making is actually a selection process among available

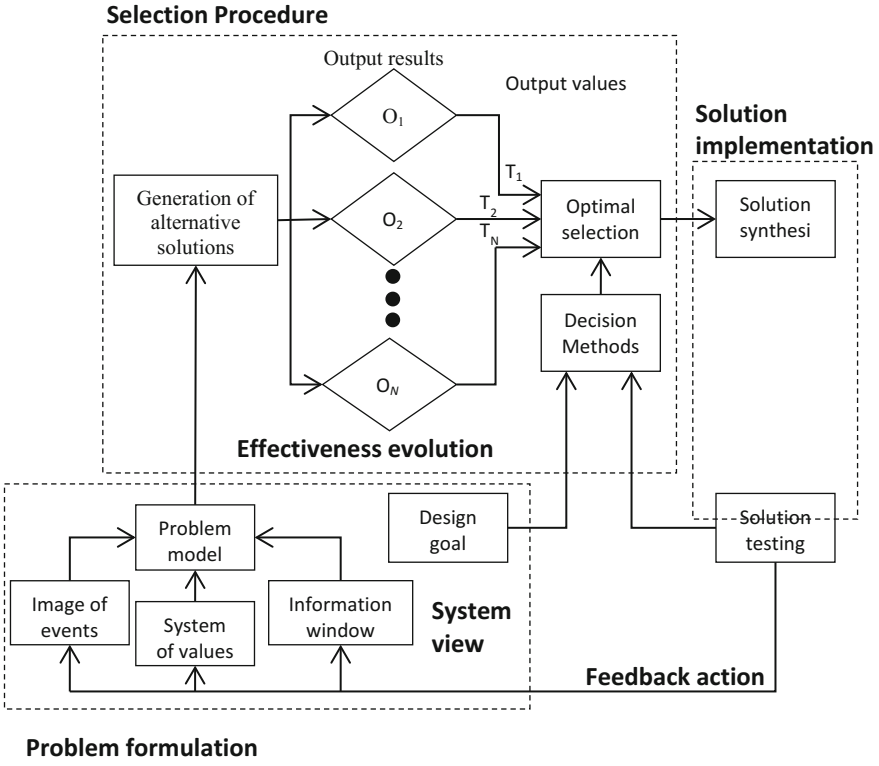



**Fig. 12.24** Structure of the general feedback decision and control model

alternative solutions or procedures. Control is the process of using the decision-making results for the initiation and implementation of the corrective actions. Management control is the control that refers to organizations and enterprises. The feedback control is the process by which the desired goals are achieved.

Actually, the structure in Fig. 12.24 embraces both technological and managerial systems. The technological and managerial (behavioral) operations are interacting and must be considered in totality (as discussed in the CIM example). This structure involves three general stages, namely: (i) *Problem formulation*, (ii) *Selection process* (optimal design/plan), and (iii) *implementation of the design*, which are divided into several steps as follows:

***Problem formulation***

- System view
- Modeling
- Requirements/design goals

***Selection procedure***

- Generation of alternative solutions
- Effectiveness evaluation of the solutions
- Optimal solution selection

***Design implementation***

- Synthesis/implementation of the selected solution
- Testing of the system
- Control

The *system view* reflects all our experience and information from the past and involves the *view of facts* (from the past), the *information window* (that keeps only the data which are relevant to the problem/system at hand), and the *system of values* upon which the system design will be based.

The *modeling step*, which is the core of the optimal decision-making process, is the process of constructing or developing a model of the system which will be based on the *system view*. Here, the system is any union of cooperative and interacting components. The mathematical models of the systems contain differential equations, algebraic equations, and the constraints of the variable/quantities involved. As we have seen in Chaps. 6 and 7, dynamic models can be described either as transfer functions (in the complex frequency domain) or as integral or integrodifferential equations. Thus, the *control design* can be made either by classical control techniques or by modern control techniques in $L_2/L_\infty$ spaces.

The *design goals* vary from case to case and may involve minimum energy consumption, maximum economic (or other) profit, minimum time control, desired final-state achievement, decoupling/non-interacting control, etc.

The *generation* of alternative solutions involves alternative procedures, alternative plans/programs/schedules/price policies, alternative design/selection processes, etc.

The *evaluation stage* requires the availability of suitable decision models and quantification of the inputs and outputs. The *optimal solution selection* can be performed by two-valued and multi-valued decision theory, and by dynamic/static optimization theory (see Chap. 7).

The *synthesis and implementation* of the selected solution may reveal practical problems not anticipated at the beginning. Therefore, very often we correct the solution using *feedback procedures* and *controllers* based on the input/output measurements. In all cases, a suitable detection/measurement mechanism is needed.

Feedback control is applied at several levels (enterprise level, national level, international level) and for several processes within the society. Some examples of such processes are the following:

- Resource allocation
- Inventory of products and raw materials
- Economic transactions
- Human factors and psychology
- Legal and ethical frameworks

By 1998, *Anthony's* definition of management control as "the process by which managers influence other members of the organization to implement the organization strategies," included human-behavioral issues [72].

Management control systems involve both formal and informal information subsystems and processes. Therefore, management control depends on issues like "cost accounting", "financial accounting", "regulatory compliance", etc. According to *Simons* [105], "taking an isolated view of features such as accounting systems or managerial behavior will leave the theory weak in both explanatory and predictive power". Based on empirical evidence, Simons looked at management control systems from the viewpoint of how managers use control systems under various conditions, taking normative design rationality as given. He also explicitly links, in his framework, controls with both strategy formulation and implementation.

In quantitative procedures, management employs the models and techniques of *operation research* (OR) [106, 107], and also economic-theory techniques. Some techniques used in MCS are:

- Program management techniques
- Capital budgeting
- Just-in-Time (JIT) techniques
- Target costing
- Activity-based costing
- Total quality management (TQM)
- Benchmarking

In 1995, Anthony actually defined management control as separate from strategic control and operational control. Langfield–Smith [108] argued that this is an artificial separation that may no longer be useful in an environment where employee empowerment has become popular. Actually, management control involves formal and informal controls, output and behavior controls, market, bureaucracy and clan controls, administrative and social controls, and results, action, and personnel controls.

#### 12.4.3.2 Economic Control Systems

Over the years, economic systems have been investigated with aid of "*theories*" and "*models*" [115]. *Kevin Hoover* (1995) [116] indicated that "model is a ubiquitous term in economics, and a term with a variety of meanings." Similarly, the term "theory" is not interpreted in a broadly unique way among economists. According to *Daniel Klein* and *Pedro Romero* (2007) [117], after defining "*model*", "*theory*" has a higher normative status than "model", and also «a theory does not require a "model", and a "model" is not sufficient for a "theory"». But these authors leave the term "theory" undefined. According to *Robert Goldfarb* and *Jon Ratner* (2008) [115]: «A widespread use of "theory and model" is that "theory" is a broad

conceptual approach (as in "price theory") while '*models*', typically in mathematical (including graphical) form, are applications of a theory to particular settings and/or represent explorations of different sets of assumptions conditionally allowable by the theory approach» In this way, for example, "*price theory models*" have a fully understandable and standard meaning. This view of the relation of "*economic theory*" and "*economic model*" is adopted in most classical textbooks on economics. Of course, this does not mean the term "model" or "theory" are able to carry all the weight of competing possible interpretations. According to *Kevin Hoover,* the term "model" has a variety of meanings, and so we have many types of economic models, including the following opposing classes:

- Evaluative/interpretive models
- Observational models

*W.E. Deming* states: "Without theory, experience has no meaning. Without theory, one has no questions to ask. Hence without theory, there is no learning. Theory is a window into the world. Theory leads to prediction. Without prediction, experience and examples teach nothing. To copy an example of success, without understanding it with the aid of theory, may lead to disaster" (Economics for Industry, Government, Education, p. 106, 1993).

The application of *feedback control* is done using a "model" which is formulated on the basis of a "theory" and tested using real data. Currently, there are many textbooks and research books concerned with the application of formal control theories (linear, nonlinear, stochastic, adaptive, etc.) based on specific mathematical models, e.g., [118–120].

Economic models were formulated and applied on two levels:

- Microeconomic level
- Macroeconomic level

In all cases of feedback control, the main question is to find a (good) control goal. Usually, this goal is related to an efficient (optimal or suboptimal) use of relevant resources (time, capital, material, human) in relation to the product of service provided. The above aspects hold in both the microeconomic and macroeconomic levels. However, the tools that are appropriate for the two economic levels are not the same, for example:

- Classical macroeconomic control tools are legislations that guarantee certain minimal constraints (e.g., minimum wages), short-term economic policies that address current problems (e.g., price limiters), and monetary interventions of the central bank, etc.
- Classical microeconomic control tools include managers that constraint the interactions of employees through a specified organization chart, the amount of allowed resources (capital, material, space, input), process control, i.e., real-time control procedures using rule sets (*Tayloristic rules* [121]), supervision procedures, and output (product) measurements.

On the microeconomic level, large companies perform their business processes within specified *organizational units* (business units) in order to create products or services that are supplied to the market. The steps followed in these business processes are determined by the *operational structure* of each company.

Economic actions are based on negative feedback which assures the achievement of a predictable equilibrium for prices and market shares. Negative feedback tends to stabilize the economy because any major changes will be offset by the very reactions they generate. According to conventional economic control theory, equilibrium marks the "*best*" outcome possible under the circumstances.

Here, it is useful to note that economic systems also exhibit positive feedback effects that need to be observed and taken care [121]. These self-reinforcing processes may appear on all economic levels. Typically, increased production brings additional benefits, producing more product units implies gaining more experience in the manufacturing process and understanding how to produce additional units even more cheaply. Thus, companies and countries that gain high volume and experience in a high-technology industry can exploit advantages of lower cost and higher quality that may make it possible for them to shut other companies or countries out.

In general, economic systems are divided into four types:

- Traditional economic systems
- Command or planned economic systems
- Market-driven economic systems
- Mixed economic systems

This distinction is based on the way these systems operate and are controlled.

*Traditional systems* tend to follow long-established patterns and so they are not very dynamic. The quality-of-life standards are static, and the individual persons don't have much occupational or financial mobility. Typically, in many traditional economies, community interests have higher priority than private interests, although in some of them some kind of private property is respected under a strict set of conditions and obligations.

*Command economic systems* are fully controlled by the government, which decides how to distribute and use the available resources. Government regulates and control prices, wages, and sometimes, what kind of work the citizens do. Socialism is the main example of a command economy.

In *pure market (neoliberal) economies,* it is the interaction of individual and companies in the market environment that determines how resources are used and goods are distributed. Here it is the individual who selects how to invest his/her own resources, what jobs to perform, what services or goods to produce, and what to consume. In a pure market economy, the government has no involvement in economic life.

Today, in most developed countries, the economy is of the *mixed* type. Of course, the kind of *mix* differs from state to state. Usually, the mix between public

and private sectors is not static but changes adaptively according to particular conditions each time. Some representative references on control in economics are [122–126].

## 12.5 Concluding Remarks

In this chapter, we have illustrated the role of feedback in biological and societal (technological and behavioral) systems. It was demonstrated that both negative and positive feedback are present and used. Negative feedback offers the means for achieving the system equilibrium state that assures the given goals in each case. Positive feedback is purposefully used whenever an oscillatory behavior is the system's goal. In living systems, the existence of feedback is intrinsic and has been established by nature itself via adaptation and evolutionary processes.

In man-made hard systems, the feedback is incorporated into the system using the classical and/or modern control theory techniques (presented in Chaps. 6 and 7), and designed so as to achieve the desired performance specifications. The main control means are suitable prime movers and end effectors. In human-behavior systems, the feedback control means are either informal cultural and ethical rules or formal rules (regulations, laws) imposed by rulers, managers, and government.

A class of feedback that is implemented by cooperation of humans and machines is the so-called "*biofeedback*". Biofeedback is used for measuring and controlling a person's specific and bodily functions (heart rate, blood pressure, muscle tension, breathing mode) or obtaining particular skills such as driving, playing games, music, art, or for rehabilitation purposes, etc. [127–130]. All the above show that "*feedback*" is actually one of the basic pillars of life and society, naturally inherent or purposefully embedded by the human designer.

## References

1. E. Schrödinger, *What is Life?* (Canto Edition) (Cambridge University Press, Cambridge, U. K., 1967)
2. P. Wellstead, Systems biology and the spirit of Tustin, IEEE Control Sys. Mag., pp. 57–71, (February 2010)
3. G.T. Reeves, S.E. Fraser, Biological systems from an engineer's point of view. PLoS Biol. **7** (1), 32–35 (2009)
4. Engineering principles in biological systems. http://meetings.cshl.edu/meetings/pastprograms/2006pastprograms/engineprogram.pdf
5. A. Daskalaki, *Handbook of Research on Systems Biology Applications in Medicine*, vol. 1, Medical Information Science Reference (Hershey, New York, 2008)
6. V.Z. Marmarelis, *Nonlinear Dynamic Modeling of Physiological Systems* (IEEE Press/Wiley, New York, 2004)
7. H.T. Milhorn, *The application of control theory to physiological systems*. http://home.comcast.net/~milhorn1 http://www.milhornbooks.com/control.html

8. T.E. Bellows, T.W. Fisher et al., *Handbook of Biological Control: Principles and Applications of Biological Control* (Academic Press, New York, 1999)
9. R.S. Parker, F.J. Doyle III, M.A. Henson, Integrating Biological Systems in the Process Dynamics and Control Curriculum, *Ch E Division of ASEE*, pp. 181–188, (2006)
10. C.G. Gross, Claude bernard and the constancy of the internal environment. Neuroscientist **4**, 380–385 (1998)
11. W.B. Cannon, *The Wisdom of the Body* (Norton, New York, 1932)
12. A.L. Hodgkin, A.F. Huxley, A quantitative description of membrane current and its application to conduction and excitation in nerve. J. Physiol. **117**, 500–544 (1952)
13. N. Wiener, *Cybernetics: Control and Communication in the Animal and the Machine* (Hermann, Paris and MIT Press, Boston MA, 1948)
14. L.E. Bayliss, *Living Control Systems* (English University Press, London, 1966)
15. M.D. Mesarovic, *Systems Theory and Biology* (Springer, Berlin, 1968)
16. L. Bertanlafly, *General Systems Theory: Foundations, Development Applications* (Brazillier, New York, 1969)
17. U. Stelling, U. Saner, Z. Szallasi, F. Doyle, J. Doyle, Robustness of cellular functions. Cell **118**(6), 675–685 (2004)
18. H. Kitano, Toward a theory of biological robustness. Mol. Sys. Biol. **3**(137), 1–7 (2007)
19. A. Cornish-Bowden, Putting the systems back into systems biology. Perspect. Biol. Med. **49** (4), 475–489 (2006)
20. J. Keener, J. Sneyd, *Mathematical Physiology* (Springer, Berlin, 1998)
21. A. Cornish-Bowden, *Fundamentals of Enzyme Kinetics* (Portland Press, New York, 2004)
22. Y-K. Kwon, K.-H. Cho, Boolean dynamics of biological networks with multiple coupled feedback loops, Biophys. J. **92**(8), 2975–2981 (2007)
23. N. Barkai, S. Leibler, Robustness in simple biochemical networks. Nature **387**, 913–917 (1997)
24. R. Thomas, R. D'Ari, *Biological Feedback* (CRC Press, Boca Raton, FL, 1989)
25. D.E. Koshland Jr., A. Goldbeter, Amplification and adaptation in regulatory sensory systems. Science **217**, 220–225 (1985)
26. O. Cinquin, J. Demongeot, Positive and negative feedback striking a balance between necessary antagonists. J. Theor. Biol. **216**(2), 229–241 (2002)
27. K.S. Saladin, Homeostasis, biology encyclopedia. http://www.biologyreference.com/Ho-La/Homeostasis.html
28. Physiological Homeostasis, Biology online. http://www.biology-online.org/4/1_physiological_homeostasis.htm
29. Control, Regulation and Feedback. http://bcs.whfreeman.com/thelifewire/content/chp41/41020.html
30. P. Revest, Physiological control mechanisms and homeostasis. http://www.medicaltextbooksrevealed.com/files/12688-53.pdf
31. K.S. Saladin, *Anatomy and Physiology: The Unity of Form and Function* (McGraw-Hill, Dubuque, IA, 2001)
32. W.W. Blessing, *The Lower Brainstem and Bodily Homeostasis* (Oxford University Press, Oxford, 1997)
33. D.-E. Chang, S. Leung, M.R. Atkinson, A. Reifler, D. Forger, A.J. Ninfa, Building biological memory by linking positive feedback loops. Proc. Nat. Acad. Sci. U.S.A. **107**(1), 175–180 (2010)
34. D. Angeli, J.E. Ferrel Jr., E.D. Sontag, Detection of multistability, bifurcations, and hysteresis in a large class of biological—feedback systems. Proc. Nat. Acad. Sci. U.S.A. **101**, 1822–1827 (2004)
35. J.E. Ferrel Jr., Feedback regulation of opposing enzymes generates robust, all-or-none bistable responses. Curr. Biol. **18**, R244–R245 (2008)
36. A.J. Lotka, Contribution to the theory of periodic reactions. J. Phys. Chem. **14**, 271–274 (1910)

37. J. Monod, F. Jacob, General conclusions: teleonomic mechanisms in cellular metabolism, growth and differentiation, cellular regulatory mechanisms. Cold Spring Harbour Symp. Quant. Biol. **26**, 389–401 (1961)
38. R. Prigogine, Lefever, symmetry breaking instabilities in dissipative systems. J. Chem. Phys. **48**, 1665–1700 (1968)
39. B. Kochel, Modulatory effects of peptide nucleic acids on human neutrophil activity in vitro manifested by phagocyte luminescence, Wroclaw University of Medicine Report/Lecture Wroclaw, 1996. Also: In: *Experimentia,* vol. 48, pp. 1059–1069 (1992)
40. B. Kochel, Control of the first-line human defense system: an autocatalytic model. Kybernetes **28**(4), 430–440 (1999)
41. T. Broker, L. Lander, *Differential Germs and Catastrophes* (Cambridge University Press, Cambridge, 1975)
42. R. Thom, *Structural Stability and Morphogenesis*, (Benjamin/Cramming, Reading, 1975)
43. K.N. Ganeshaiah, R. Vasudeva, R. Uma Shaanker, Development of sinks as an autocatalytic feedback process: a test using the asymmetric growth of leaves in Mestha (*Hibiscus Cannabinus L.*). Ann. Bot. **76**(1), 71–77 (1995)
44. K.N. Ganeshaiah, R. Uma Shaanker, Frequency distribution of seed number per fruit in plants: a consequence of self organizing process? Curr. Sci. **62**, 359–365 (1992)
45. E.D. Sontag, Some new directions in control theory inspired by systems biology. Syst. Biol. **1**(1), 9–18 (2004)
46. P.A. Iglesias, B.P. Ingalls (eds.), *Control Theory and Systems Biology* (The MIT Press, Cambridge, 2010)
47. A. Bennet, *A History of Control Engineering 1930–1955* (IEE/Peter Peregrinus, London, pp. 133, 1993)
48. G.T. Reeves, S.E. Fraser, Biological systems from an engineer point of view. PLoS Biol. **7**, e21 (2009). doi:10.1371/journal.pbbio.1000021
49. C.W. Swan, *Applications of Optimal Control Theory in Biomedicine* (Marcel Dekker, New York, 1984)
50. E. Marder, R. Calabrese, Principles of rhythmic motor pattern generation. Physiol. Rev. **76** (3), 687–717 (1996)
51. V. Vuksanovic, C. Radenovic, B. Belesin, Oscillatory phenomena, processes and mechanisms-physical and biolocal analogy, *IPPA 1998 Conf.*, BIBLID 0021-3225, **34**, 247–258 (1998)
52. M.F. Simoni, S.P. DeWeerth, Sensorimotor feedback in a closed –loop model of biological rhythmic movement control. Proc. 24th Annual Conf. on Engineering in Medicine and Biology, p. 2561, Houston, 23–26 Oct 2002
53. S. Waldherr, T. Eissing, F. Allgower, Analysis of feedback mechanisms in cell-biological systems. Proc. 17th World IFAC Congress, pp 15861–15866, Seoul Korea, 6–11 July 2008
54. R. Heinrich, S. Schuster, *The Regulation of Cellular Systems* (Chapman and Hall, New York, 1996)
55. A. Schitztler, J. Gross, Normal and pathological oscillatory communication in the brain. Natl. Rev. Neurosci. **6**, 285–295 (2005)
56. R.E. Dolmetsch, K. Xu, R.S. Lewis, Calcium oscillations increase the efficiency and specificity of gene expression. Nature **392**, 933–936 (1998)
57. J. Sturis, K.S. Polonsky, E. Mosekilde, E. Van Cauter, Computer model for mechanisms underlying ultradian oscillations of insulin and glucose. Amer. J. Physiol. Endocrinol Metab **260**, E801–E809 (1991)
58. E.D. Lehmann, A physiological model of glucose-insulin interaction in type 1 diabetes melitus. J. Biomed. Eng. **14**(3), 235–242 (1992)
59. A. Makroglou, J. Li, Y. Kuang, Mathematical models and software tools for the glucose-insulin regulatory system and diabetes: an overview. Appl. Numer. Math. **56**(3), 559–573 (2006)

60. G.D. Mitsis, M.G. Markakis, V.Z. Marmarelis, Nonlinear modeling of the dynamic effects of infused insulin on glucose: comparison of compartmental with volterra models. IEEE Trans. Biomed Eng. **56**(10), 2347–2358 (2009)
61. M.G. Markakis, G.D. Mitsis, G.P. Papavassilopoulos, V.Z. Marmarelis, Model predictive control of blood glucose in type 1 diabetes: the principal dynamic modes approach, Proc. IEEE Conf. on Engineering in Biology Society, pp. 5466–5469 (2008)
62. G.D. Mitsis, V.Z. Marmarelis, Modeling of nonlinear physiological systems with fast and slow dynamics i: methodology. Ann. Biomed. Eng. **30**, 272–281 (2002)
63. R.N. Bergman, Y.Z. Ider, C.R. Bowden, C. Cobelli, Quantitative estimation of insulin sensitivity. Amer. J. Physiol. **236**, E667–E667 (1979)
64. A. De Gaetano, O. Arino, Mathematical modeling of the intravenous glucose tolerance test. J. Mathem. Biol. **40**, 136–168 (2000)
65. R. Abu Zitar, Towards neural network model for insulin/glucose in diabetics–II. Informatica **29**, 227–232 (2005)
66. H. Wang, J. Li, Y. Kuang, Mathematical modeling and qualitative analysis of insulin therapies. Math. Biosci. **210**, 17–33 (2007)
67. J. Li, Y. Kuang, C.C. Masa, Modeling the glucose-insulin regulatory system and ultradian insulin. J. Theor. Biol. **242**, 722–735 (2006)
68. I.-M. Tolic, E. Mosekilde, J. Sturis, Modeling the insulin-glucose-feedback system: the significance of pulsatile insulin secretion. J. Theor. Biol. **207**, 361–375 (2000)
69. L. Kovacs, B. Paláncz, Z. Benyo, Classical and modern control strategies in glucose—insulin stabilization, http://mycite.omikk.bme.hu/15987.pdf
70. M.C.K. Khoo, A Model of Respiratory Variability During Non-REM Sleep, ed. by G.D. Swanson, F.S. Grodins, R.L. Hughson, *Respiratory Control: A Modeling Perspective* (Plenum Press, New York, 1989)
71. K. Wasserman, B.J. Whipp, R. Casaburi, Respiratory Control During Exercise, ed. by A. P. Fishman, N.S. Cherniak, J.G. Widdicombe and S.R. Geiger, *Handbook of Physiology*, Section 3: The Respiratory System, vol. III, Control of Breathing Part 2, (American Phys. Society, Bethesda, Maryland, 1986)
72. J. Trinder, F. Whitworth, A. Kay, P. Wilkin, Respiratory instability during sleep onset. J. Appl. Physiol. **73**(6), 2462–2469 (1992)
73. J.J. Batzel, H.T. Tran, Modeling instability in the control system for human respiration: applications to infant non-rem sleep. Appl. Math. Comput. **110**, 1–51 (2000)
74. M.C.K. Khoo, R.E. Kronauer, K.P. Strohl, A.S. Slutsly, Factors inducing periodic breathing in humans: a general model. J. Appl. Physiol. **53**(3), 644–659 (1982)
75. L.B. Rowell, *Cardiovascular Control* (Oxford University Press, New York, 1993)
76. H. Tsuruta, T. Sato, M. Shirataka, N. Ikeda, Mathematical model of the cardiovascular mechanics for diagnostic analysis and treatment of heart failure: part 1, model description and theoretical analysis. Med. Biol. Eng. Comp. **32**(1), 3–11 (1994)
77. F.S. Grodins, Integrative cardiovascular physiology: a mathematical model synthesis of cardiac and blood vessel hemodynamics. Q. Rev. Biol. **34**(2), 93–116 (1959)
78. J.J. Batzel, F. Kappel, S. Timischi-Teschi, A cardiovascular-respiratory control system model including state delay with application to congestive heart failure in humans. J. Math. Biol. **50**(3), 293–335 (2005)
79. J.J. Batzel, F. Kappel, D. Schneditz, H.T. Tran, *Cardiovascular and Respiratory Systems: Modeling Analysis and Control* (Cambridge University Press, Cambridge, 2010)
80. S. Poore, Overview of social control theories, The Hewett School http://www.hewett.norfolk.sch.uk/curric/soc/crime
81. C. Livesay, Informal Social Control http://www.sociology.org/uk/p2s5an4.htm
82. F. Aftalian (Translator O. Theodor), *A History of the International Chemical Industry* (University of Pennsylvania Press, Philadelphia, 1991)
83. J. Hahn, D.P. Guillen, T. Anderson, Process control systems in the chemical industry: safety vs. security (Preprint), *Idaho National Laboratory*, INL/COW-05-00001, April 2005

84. T.M. Stout, T.J. Williams, Pioneering work in the field of computer process control. IEEE Ann. Hist. Comput. **17**(1), 6–18 (1995)
85. SIMATIC PCS7, The Process Control System SIMATICPCS7, *Siemens AG*, 2009
86. S.G. Tzafestas (ed.), *Microprocessor in Signal Processing Measurement and Control* (D. Reidel, Dordrecht, 1983)
87. S.G. Tzafestas, J.K. Pal (eds.), *Real-Time Microcomputer Control of Industrial Processes* (Kluwer, Boston/Dordrecht, 1990)
88. N.G. Leveson, M.P.E. Heimdahl, H. Hildreth, J.D. Reese, Requirements specification for process control systems. IEEE Trans. Softw. Eng. **20**(9), 684–697 (1994)
89. S.G. Tzafestas (ed.), *Advances in Manufacturing: Decision, Control and Information Technology* (Springer, London/Berlin, 1999)
90. S. Lee, A. Wysk, J.S. Smith, Process planning interface for shop floor control architecture for computer. Integrated manufacturing, Int. J. Prod. Res. **33**(9), 2415–2435 (1994)
91. P.G. Ranky, *Computer Integrated Manufacturing: An Introduction with Case Studies* (Prentice Hall, Englewood Cliffs, N.J., 1986)
92. A. Kusiak, *Intelligent Manufacturing Systems* (Prentice Hall, Englewood Cliffs, NJ, 1990)
93. D. McFarlane, S. Sarma, J.L. Chirn, C.Y. Wong, K. Ashton, *The Intelligent Product in Manufacturing Control* (Proc IFAC World Congress, Barcelona, 2002)
94. J. Taylor, *The Lore of Flight* (Universal Books, London, 1990)
95. C.R. Spitzer, *The Avionics Handbook* (CRC Press, Boca Raton/London, 2001)
96. Airbus A380 Cockpit. http://www.airlines.net
97. Introduction to Aircraft Control, NASA. http://dcb.larc.nasa.gov
98. N.B. Sarter, D.D. Woods, Pilot interaction with cockpit automation: operational experiences with the flight management system. Intern. J. Aviat. Psychol. **2**, 303–321 (1992)
99. N.B. Sarter, D.D. Woods, C.E. Billings, Automation Surprises, ed. by G. Salvendy, *Handbook of Human Factors and Ergoeconomics* (Wiley New York, 1997), pp 1926–1943
100. M.S. Nolan, *Fundamentals of Traffic Control* (Books Cole Pub. Co, Pacific Grove, 1999)
101. S. Nof, *Handbook of Industrial Robots* (J. Wiley, New York, 1999)
102. S. Tzafestas, *Introduction to Mobile Robot Control* (Elsevier, New York, 2013)
103. S. Tzafestas, *Sociorobot World: A Guided Tour for All* (Springer, Berlin/New York, 2015)
104. R. Anthony, D. Young, *Management Control Systems* (Irwin, McGraw Hill Chicago, 1999)
105. R.L. Simons, *Levers of Control: How Managers Use Innovative Control Systems to Drive Strategic Renewal* (Harvard Business School Press, Boston, MA, 1995)
106. F.S. Hillier, G.J. Lieberman, *Introduction to Operations Research*, 8th edn. (McGraw-Hill, New York, 2005)
107. S.G. Tzafestas, *Optimization and Control of Dynamic Operational Research Models* (North-Holland, Amsterdam, 1982)
108. K. Langfield-Smith, Management control systems and strategy: a critical review. Acc. Organ. Soc. **22**(2), 207–232 (1997)
109. R. Anthony, V. Govindarajan, *Management Control Systems* (McGraw-Hill, Chicago, Irwin, 2007)
110. D. Otley, Management control in contemporary organizations: towards a wider framework management. Acc. Res. **5**, 289–299 (1994)
111. J. Maciariello, C. Kirby, *Management Control Systems to Attain Control* (Prentice-Hall, New Jersey, 1994)
112. ICMR, *Management Control Systems: Overview*, 2nd edn, IBS Center for Management Research: An Integrated Approach to Performance and Compliance (IBS-CDC, Hyderabad, India, 2010)
113. J.C. Lere, K. Portz, Management control systems in global economy, The CPA Journal http://www.nysscpa.org/cpajournal/2005/905/essentials/p.62.htm
114. D.M. Daley, *Strategic Human Resource Management: People and Performance Management in the Public Sector* (Prentice-Hall, Upper Saddle River, NJ, 2002)

115. R.S. Goldfarb, J. Ratner, Economics in practice: follow-up theory and models: terminology through the looking glass. Econ. J. Watch **5**(1), 91–108 (2008)
116. K. Hoover, Facts and artifacts: calibration and the empirical assessment of real-business-cycle models. Oxford Econ. Papers New Ser. **47**(1), 24–44 (1995)
117. D. Klein, P. Romero, Model-building versus theorizing: the paucity of theory. J. Econ. Theory Econ. J. Watch **4**(2), 241–271 (2007)
118. P. Kopacek (ed.), *Advanced Control Strategies for Social and Economic Systems* (Elsevier, Amsterdam, 2005)
119. P. Chen, S.M.N. Islam, *Optimal Control Models in Finance: A New Computational Approach (Applied Optimization)* (Springer, London/Berlin, 2004)
120. G.C. Chow, *Analysis and Control of Dynamic Economic Systems* (Krieger Publ. Co., Malabar, FL, 1986)
121. F.W. Taylor, On the art of cutting metals. ASME J. Engrg Ind. **28**, 31–35 (1906)
122. J.H. Westcott, A.G. J. MacFarlane, J. Mason, Application of control theory to macroeconomic models and discussion, Online ISSN1471-2946, *The Royal Society* (2010)
123. V.Z. Belenky, A.M. Belostovsky, Control of economic systems under the process of data improvement. J. Econ. Dyn. Control **12**(4), 609–633 (1988)
124. A. Leijonhufvud, Models and theories. J. Econ. Methodol. **4**(2), 193–198 (1997)
125. A. Mas-Collell, M. Winston, J.R. Green, *Microeconomic Theory* (University Press, Oxford, 1995)
126. W.B. Arthur, Positive feedbacks in economics, Sci. Am. **262**, 99–109, Feb 1990
127. Biofeedback Technology: A Prospectus. http://www.athealth.com/Practitioner/particles/Guest_CoopersteinMA.html
128. Biofeedback information from the Mayo Clinic. http://www.mayoclinic.com/health/biofeedback/SA00083
129. Biofeedback from Zing Health in Australia. http://www.singhealth.com.au/bio-feedback.php
130. Biofeedback.http://en.wikipedia.org/wiki/Biofeedback

# Chapter 13
# Adaptation and Self-organization in Life and Society

*For the source of any characteristic so widespread and uniform as this adaptation to environment we must go back to the very beginning of the human race.*

Ellsworth Huntington

*The survival of the fittest is the ageless law of nature, but the fittest are rarely the strong. The fittest are those endowed with the qualifications for adaptation, the ability to accept the inevitable and conform to the unavoidable, to harmonize with existing or changing conditions.*

Dave E. Smalley

**Abstract** The aim of this chapter is to demonstrate the role of adaptation and self-organization in life and society. The range of adaptation is very wide and includes, among others, animal physiology adaptation, immigrant adaptation, animal fertility adaptation, emotional adaptation, adaptation to stress, etc. Self-organization is an intrinsic process taking place in both biological and societal systems. In both cases, the rules of self-organization are determined on the basis of local information only, without information from a global level. Examples of self-organizing biological systems or patterns include a raiding column of army ants, a termite mound, pigmentation patterns on shells, etc. This chapter illustrates the presence of adaptation and self-organization through a number of representative examples, namely: adaptation of animals, adaptation of ecosystems, adaptation of immune systems, adaptation of socio-ecological and general societal systems, self-organization of knowledge management, and self-organization of technological and man-made systems (traffic lights control, WWW, multiagent robotic systems, bio-inspired systems). The above examples demonstrate clearly that adaptation and self-organization are fundamental processes for the survival of living organisms and societies, and the optimal operation of hard and soft man-made systems.

**Keywords** Adaptation · Self-organization (S-O) · Life · Human society
Biological system · Genetic evolution · Fractal scaling (power law)
Self-organizing social system · Social dimension · Institutional dimension
Economic dimension · Environmental dimension · Animal adaptation
Complex adaptive system (CAS) · Ecosystem as CAS · Immune system adaptation
social ecological system · Stock market as CAS · Internet-based S-O voting system



S. G. Tzafestas, *Energy, Information, Feedback, Adaptation, and Self-organization*,
Intelligent Systems, Control and Automation: Science and Engineering 90,
https://doi.org/10.1007/978-3-319-66999-1_13

Knowledge pillars · Knowledge-based society · Man-made S-O systems
S-O traffic lights control

## 13.1 Introduction

In Chaps. 8 and 9, we have presented the concepts of adaptation and self-organization including the relevant definitions, the historical landmarks of their study, and the properties possessed by adaptive/complex adaptive systems and self-organizing systems in nature and society. We have seen that adaptation is manifested as genetic adaptation, structure adaptation, physical/physiological adaptation, function adaptation, and, in general, evolution adaptation. According to *Julian Huxley:* "adaptation is nothing else than arrangements subserving specialized functions, adjusted to the needs and the mode of life of the species or type… Adaptation cannot but be universal among organisms, and every organism cannot be other than a *bundle of adaptations*, more or less detailed and efficient, coordinated in greater or lesser degree [1, p. 420]." A presentation of historical perspectives on adaptation can be found in [2], a global view of adaptation is provided in [3], and a critique of the evolutionary thought on adaptation and natural selection is offered in [4]. Today there is a vast bibliography on "adaptation" and "complex adaptive systems," where the topics of life, society, and science that it covers are revealed and discussed.

A few of them are discussed in [5–15]. They range from animal physiology adaptation [6], immigrants adaptation [8], human fertility adaptation [9], adaptation to stress [10], adaptation in the mind-brain and physiology concepts [11], technology adaptation of e-society [12], emotional adaptation [13], to the view and study of the Web and supply networks as complex adaptive systems [14, 15].

As explained in Chap. 9, self-organization is a process, in which "*patterns*" at the global level of a system are the result of numerous interactions among the components of its lower levels. The rules that determine the interactions among the system's components are executed on the basis of local information only, without information from the global level. The patterns of the global level are an emergent property imported into the system by an external ordering entity.

Examples of self-organizing biological patterns are: a raiding column of army ants, the complex architecture of a termite mound, pigmentation patterns on shells, etc. Living systems obey, in addition to the physical laws, genetic programs that are the result of genetic evolution. This adds an extra dimension to self-organization in biological systems because the fine-tuning of the rules of local interactions is controlled by natural selection. A complex phenomenon possesses self-organizing complexity only if it is governed by some kind of fractal (power law) scaling. Of course, power scaling may be applicable only over a limited range of scales.

Self-organization is also a feature of human society. Some representative references dealing with self-organization in society and technology are [7–16]. Self-organization in modern society involves four basic dimensions, namely: the social dimension, economic dimension, institutional dimension, and environmental dimension, which are interrelated and interacting as shown in Fig. 13.1.

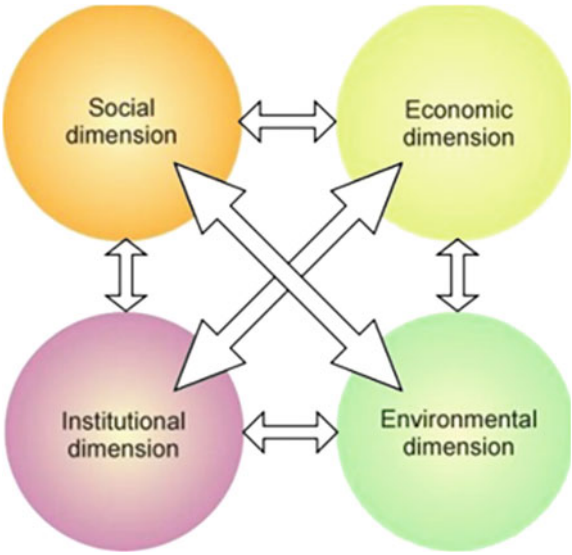



**Fig. 13.1** The four principal interconnected dimensions of social self-organization http://www.eolss.net/CF03-1.jpg). The reader is informed that Web figures and references were collected at the time of writing the book. Since then some of them may no longer be valid due to change or removal by their creators, they may no longer be useful

Complex adaptive systems theory offers the tools to analyze how large-scale self-organization arises and is maintained in many physical, biological, natural, and societal systems.

An important example, very crucial for modern human life on Earth, is the area of ecosystems and the biosphere. Understanding how changes, at one level of biological organization, influence the patterns or mechanisms occurring at another level, and how the cross-scale interactions lead to adaptation and self-organization can considerably help the management efforts that aim to assure manipulation and rehabilitation/restoration of damaged ecosystems.

The chapter presents the following examples of adaptation and self-organization in life and society: adaptation of animals, ecosystems, climate change, immune systems, social-ecological systems, stock market, general society systems, knowledge management, and man-made self-organizing systems design.

## 13.2 Adaptations of Animals

Adaptations of animals are among the best examples of adaptation in nature [5, 17]. Both animals and plants are continually adapting to their habitats. The habitats of plants and animals offer extremely diverse living conditions of temperature and water availability over the widely spaced areas of Earth. For example, more than 99% of Antarctica is covered by ice. Antarctica is very cold, and only a small number of plants grow there (e.g., algae, mosses, and lichens). Animals that have adapted to live in Antarctica obtain their food from the sea or migrate to leave the

continent during the winter. By adaptation, an animal's body changes to help the animal to live and survive in its environment. The physical characteristics of animals help them to find food, live safely, survive the weather conditions, etc. These characteristics are collectively called *adaptations.* Adaptations in each species has developed slowly over many generations, i.e., they are the result of evolution.

Some general examples of animals' physical adaptations are the following:

- Wing-flapping mode
- Bird's beak shape
- Nose and ear's shape
- Type of für
- Color of the für
- The number of fingers
- The locomotion style, etc.

The wing of a bird ends in a set of digits. The wing surface is made up of flight feathers aligned laterally. The pectoral muscles of a bird are purposefully located below the wings and provide the locomotive driving power for the bird flight. Birds use warm rising currents of air to stay afloat without using much energy. Birds are able to orient themselves using landmarks, the Sun and the Earth's magnetic field for locating true North. Night birds learn how to orient themselves by the positions of the stars, using celestial navigation; birds migrate annually, typically from breeding to nonbreeding grounds, relocating to areas with abundant food and returning to their breeding grounds when the food is again abundant to breed and bring up their young.

In the following, we provide a short list of animal adaptations in order to illustrate their variety and usefulness:

**Polar bears** Their color is white to blend in with the snow and ice. They have under their skin a thick layer of fat to keep warm in their cold environment. Their large paws enable the polar bear to walk in the snow.

**Penguins** These flightless birds are excellent swimmers using their webbed feed. They live on pack ice and the oceans around Antarctica. They keep warm using their thick skin and the large amount of underlying fat. They have streamlined bodies to reduce drag in the water and flipper-shaped wing to be able to "fly" underwater at speeds reaching 15 mph.

**Camels** They have many adaptations that enable them to live in desert environments that are dry and hot. They have long eyelashes and nostrils that can open and close for protection against the sand blown around by winds. They can live for more than a week without water and without food for long periods (many months). When they find water, they can drink up to 40–45 L.

**Fennec fox** This is the only carnivore able to live in a desert habitat without free water. To this end, their kidneys are adapted to function with only little water. They take moisture from their food. They have thick fur to insulate them from the cold

desert nights. They have sandy-colored fur for camouflage and thick fur on the soles of their feet for insulation against the hot sand of the desert.

**Lions** Some of their adaptations are: heavily muscled forelimbs and shoulders for capturing large prey, eyes set in the front of head for depth perception and good estimation of distances, rough tongues to peel the skin of prey animal away from the flesh and the flesh from bone, belly skin for protection against kicking by prey, and forepaws equipped with long retractile claws for easy grabbing and holding of prey.

**Giraffes** Their long neck help them to feed from treetops and detect predators. Their hearts are extremely large and powerful to pump blood up their long necks to their brains (about 2–3 times more powerful than a human's heart). A giraffe can drink up to 15–18 L of water. The spots on their fur are for camouflage among the trees. Giraffes get water from the dew on the leaves that they eat. Their long and tough tongues enable them to pull leaves from the branches without being hurt by the thorns during foraging.

**Dolphins** They have keen hearing ability and high mobility for protection from predators. For better protection, they swim in groups. Their bodies are streamlined to enable them to moving fast to catch food and escape from predators. They also have greater intelligence than most of the other mammals.

## 13.3 Ecosystems as Complex Adaptive Systems

*Ecosystems* represent one of the classes of natural systems that possess the features of complex adaptive systems outlined in Chap. 8. As we saw in Sect. 10.3, ecosystems are *open thermodynamic systems with respect to energy*. Ecosystems are controlled in two well-balanced ways:

- **Bottom-up control** Here, it is the nutrient supply to primary producers that controls the operation of the ecosystem.
- **Top-down control** Here, it is predation and grazing by higher trophic levels on lower trophic levels that controls the ecosystem operation.

If the nutrient supply is increased, the autotrophs' production increase is propagated via the food web, and so all the other trophic levels will expand accordingly. If there is an increase in predators, fewer grazers will result, which leads to more primary producers as fewer of them are eaten by grazers. This means that the control of population size and overall productivity "*travels*" from the top levels of the food chain down to the bottom trophic levels. Actually, in real ecosystems, neither mechanism completely occurs, but both of these controls take place in any ecosystem at any time. In order to see how an ecosystem will behave or adapt under several situations (e.g., a climate change), we have first to understand how both *bottom-up* and *top-down* control mechanisms are operating. Prey and predators

have reciprocal roles that are specified by the feedback loops in which they participate. To be part of a food web *ipso facto* is to belong to a system of feedback loops that establish pathways of energy flow from living being to living being.

One of the early studies on the consideration of ecosystems and biosphere as complex adaptive systems was conducted by *Simon Levin* [18, 19]. He demonstrated that ecosystems possess the four CAS properties suggested by Holland (complexity, nonlinearity, flows, and diversity). *Bonabeau* [20] classified social insect colonies (ants, termites, etc.) as complex adaptive systems and demonstrated that they possess all the CAS properties presented by *Levin* [18]. Social insect colonies are composed of hundreds to millions of genetically similar individuals. These individuals interact locally but collectively to produce large-scale patterns of colonies dynamics. A similar view of ecosystems as complex adaptive systems was inherent in [21]. A simple adaptive system is a *flock of birds*. Actually, there is no bird-in-chief directing the behavior. Individual birds have some degree of decision-making capacity, but all the flight decisions have to follow simple rules, such as:

- Align flight to match the neighbors
- Avoid collision with neighbors or obstacles
- Fly an average distance from neighbors.

These rules, despite their simplicity, lead to very complex adaptive flocking behaviors. Figure 13.2 shows an example of bird collective movement and two other cases of self-adaptive collective movement in nature.

*Janssen* [22] uses genetic algorithms as a modeling tool for adaptation and management in two different cases, namely: (i) how the evolution of drug resistance alters malaria dynamics, and how individual-level variability in humans changes group responses to elevated atmospheric carbon dioxide. These two case studies help to explain how relatively simple CAS techniques may lead to the emergence of fresh perspectives on complex management problems that cannot be easily addressed by standard ecological models.

In general, CAS techniques provide insight into large- and cross-scale ecological interactions and help in the successful analysis of the role of adaptation in driving system dynamics and responses to new situation.

## 13.4 Adaptation to Climate Change

In this section, we will discuss one particularly critical, current problem of ecological adaptation, namely the problem of *adaptation to climate change*. The study of this adaptation is very complex and provides many challenges. One of them is the requirement to get information about impacts on climate change, as well as their secondary effects. All the approaches to understanding the potential impacts of climate involve, or are dominated by, uncertainty. Any attempt to face these

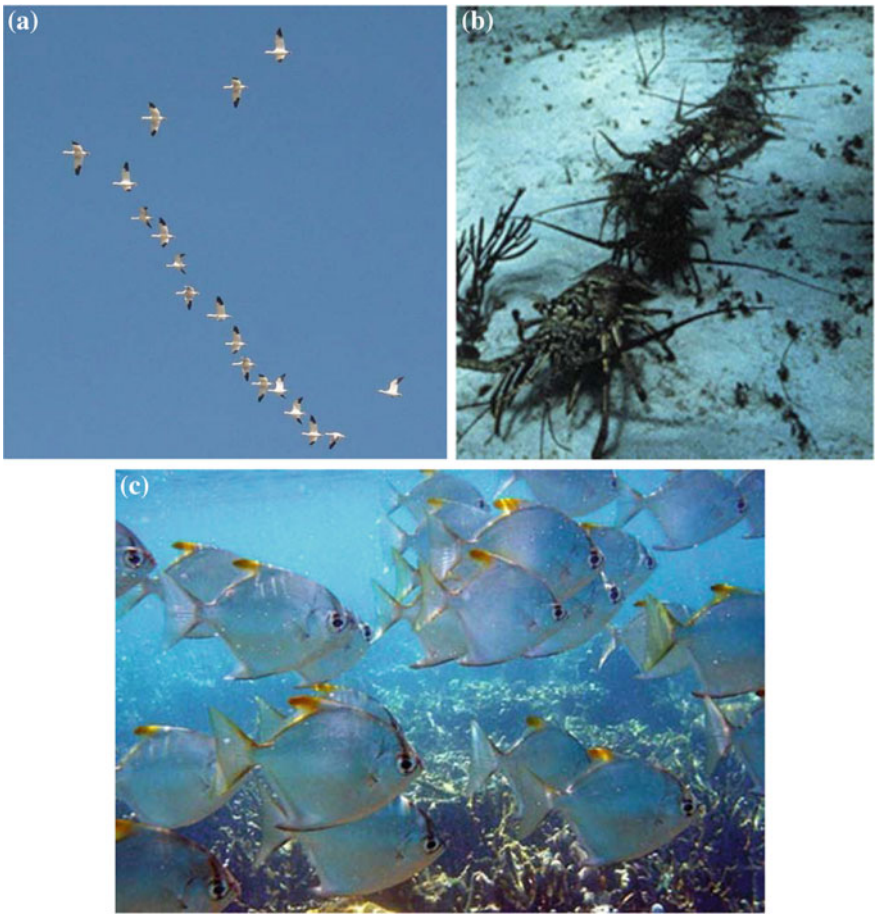



**Fig. 13.2** Three examples of collective movement: **a** birds, **b** social insects, **c** fish (www.irit.fr/TFGSO/DOCS/TFGSO_Mano.ppt, http://www.cs.tufts.edu/~paulina/images/fish.jpg)

uncertainties needs the design of adaptation policies that would be successful under a wide variety of future climate conditions, i.e., they should be "*robust*" against the uncertainties. Of course, it would be difficult to develop adaptation options that address simultaneously a wide range of drier and wetter conditions. These drier and wetter conditions need to be handled by different adaptation schemes. It should be noted that it is not always necessary to justify adaptation actions. If the weather seems to follow a well-known trend, there is no need for detailed climatic data for deciding the policies to be followed.

Clearly, each household, community, and society needs to design an adaptation strategy that fits its own specific conditions. This can be done via an enabling national policy/legislation framework and functional and environmentally conscious institutions. In any case, the resources allocated must be sufficient to meet the minimal, administrative, societal levels.

According to the *Environmental Change Institute* (**ECI**) [23], the following issues must be studied and faced:

- How to make successful decisions about adaptation to climate change.
- The effectiveness of international environmental agreements.
- The role of hybrid schemes of governance in environmental risk management (co-governance, public-private partnerships, social-private partnerships).
- Motivators of behavioral responses to environmental risk (human motivation, social values and culture, social and economic characteristics, attitudes towards the environment, etc.).

Of course, the above challenges of adaptation to climate change appear differently in developing and developed countries, although there are the following *common concerns* [24]:

- The need to shift from studying the impact of climate change to increased understanding of how to make adaptation occur
- The need to examine adaptation needs and identify priorities
- The relative roles of adaptation and mitigation actions
- The need to clarify the relationship between climate-change adaptation policies and the mainstream of development and financial support
- What funding mechanisms and organizations can be used for delivery at national and international levels.

**Climate Change in Developed Countries**

- Developed countries have accepted the need to meet obligations (financial and other) towards covering the cost of the accumulated greenhouse gas.
- The financial mechanisms should deliver effectively for their taxpayers.
- The minimum conditions for accessing the required funding must be met.
- There should be no proliferation of new funds under the Convention.

**Climate Change Issues of Developing Countries**

- Aspects of equity and justice about the damage caused by climate change to vulnerable countries due to emissions from rich developed countries are of primary concern.
- Developed countries must deliver on their obligations under the Convention for finance, technology, and capacity building.
- The additional costs for adaptation to climate change should be covered.
- Governance of financial mechanisms should be transparent and include an equitable and balanced representation by all parties and “direct access” to funding to all recipient countries.
- Support should be provided, not in a fragmented manner, but through international organizations (e.g., the United Nations Development Framework Convention on Climate Change (*UNFCC*).

In [25], the *Intergovernmental Panel on Climate Change* (**IPCC**) of **EPA** (*Environmental Protection Agency*, USA) provides a discussion on climate-change adaptation strategies and states that "adaptation alone is not expected to cope with all the projected effects of climate change, and especially not over the long term as must impacts increase in magnitude." IPCC's definition of adaptation is: "the adjustment in natural or human systems in response to actual or expected climatic stimuli or their effects, which moderates harm or exploits beneficial opportunities" (2007).

In non-managed natural systems, adaptation is not planned but occurs when forced to do so (e.g., as the climate warms, tree and animal species migrate to the north in order to live in suitable climatic conditions and habitats). In human society, much of the adaptation is planned and implemented by the government, public agencies, and private organizations. But for humans, adaptation is a risk-management process that has costs and is not foolproof. The estimated value of avoided damages against the costs of implementing the adaptation strategy should be taken into account.

Some examples of potential adaptations in various realms of human society are the following [25]:

- **Human health** (Urban tree planting, weather advisories, adjustment of clothing, and fluid intake)
- **Coastal areas and sea-level rise** (Shore protection, adaptive land-use measures, protection of water supplies from saltwater contamination)
- **Agriculture and forestry** (Controlling the planting dates, breeding novel, more-tolerant plant species and crops, controlling insect outbreaks)
- **Ecosystems and wildlife** (Protecting species-migration avenues, promoting management practices that provide reliance to the ecosystem)
- **Water and energy resources** (Improving water-use efficiency, conserving soil moisture, protecting fresh-water resources from saltwater intrusion, improving energy-use efficiency, diversifying power supply to face power-plant failures, protecting power facilities against extreme weather phenomena).

Other useful and easily accessible sources in which the adaptation of humans to climate change is discussed include [23, 25–28].

## 13.5 Adaptation of Immune and Social-Ecological Systems

The *immune system* aims to maintain the health of the body through protection from invasions of bacteria, viruses, fungi, and parasites. The immune system has the ability to detect and eliminate these harmful pathogens and remember successful responses to invasions so as to reuse these responses when similar pathogens invade in the future. The adaptability of immune systems is due to the distributed system of an extremely diverse set of modules (lymphocytes) that assure the detection of

pathogens by different modules. The principal advantage of the adaptive immune systems is their ability to match partial and temporal defense mechanisms to those of the pathogens' evolution. Actually, viruses and bacteria multiply quickly (with generational periods on the order of minutes or hours) which allows them to mutate and genetically change easily and quickly. Long-lived vertebrates cannot match the pace of pathogen evolution, but the adaptive immune system offers a suitable evolutionary adaptation to this mismatch in scale [29, 30]. This adaptation ability is achieved by a special class of white blood cells, the *lymphocytes*, which circulate throughout the body via the lymph system. The primary function of lymphocytes is to detect pathogens and help the organism to eliminate them. The immune system of vertebrates needs to face local disease ecology from the moment the organism is born. The initial capacity of the offspring comes from the antigen experience accumulated by the mother and is extended through lactation in mammals [31].

The immune adaptive systems function on several spatial and temporal scales, namely [32]:

1 Extremely small spatial and temporal scales at the molecular-level dynamics of the interacting antibodies.
2 Immunity level at which antibodies proliferate against a specific antigen. During this process, the antibody population becomes more specific and the concentration increases.

Analogously, *social-ecological systems* (**SESs**) work on several social-ecological scales that depend on the level of the social agent involved. For example, a community developing response mechanisms to deal with a particular disturbance represents a "*social agent*" that may vary from an individual to a state, learning, experimentation, and memory towards the adaptation goal. A social agent may involve several functional components (entrepreneurs, innovators, visionaries, experiment specialists, etc.) that contribute substantially to achieve the required adaptability.

It is noted that under certain circumstances an immune system or social-ecological system may lose adaptive capability by suppressing disturbances. For example, the suppressions of fire lead to an accumulation of tree biomass [33], and the lack of fire leads to suppression and elimination of fire-resistant species via competition from other species because there is a cost to being fire-resistant. Adaptation of SESs to reduce the risk of the system due to a small crisis (e.g., fires on an immune response) helps the system to prevent a large crisis (e.g., ecosystem conversion or bacterial resistance). The immune system has been extensively studied in the framework of complex adaptive systems. For example, in [34], a simulation method for the immune system is presented via a CAS model, and, in [35], the simulation of an immune system and HIV was considered using genetic algorithms, cellular automata, and classifier algorithms. A study of immune systems using genetic algorithms is presented in [36].

## 13.6 Stock Markets as Complex Adaptive Systems

*Classical stock-market theory* is based on a few basic assumptions, mainly "*primary efficient market*" and "*investor rationality.*" A short discussion of these assumptions follows [37]:

- **Stock-market efficiency** This assumption suggests that stock prices include all relevant information when this information is readily available and widely disseminated. Therefore, it is assumed that there is no systematic way to exploit trading opportunities and achieve better results. Market efficiency does not mean that stock prices are always correct, but it does imply that stock prices are not mispriced in a systematic or predictable way in any manner. The changes in prices come only as a result of the receipt of random (unexpected) information, a process modeled by a *random walk*. As a result, the efficient market assumption leads to modest trading activity and limited price fluctuations.
- **Investor rationality** Rational investors can rapidly and precisely evaluate and optimize risk/reward outcomes. They persistently seek profit opportunities, and their efforts lead to the market efficiency. Actually, rational investors try to obtain the highest return for a given risk level. Of course, investor rationality does not mean that *all* investors are rational profit seekers.

A classical stock market falls short in the following areas:

- **Stock-market returns are not normal** (as capital-market theory suggests). The return distributions show high "kurtosis," which implies that periods of relatively modest change are interspersed with higher than predicted changes (i.e., booms and crashes).
- **The random walk model is not supported by the data** Return series are frequently both persistent and trend-reinforced, i.e., financial-asset return can be predicted to some degree.
- **The relation of reward and risk is not linear** The *Capital-Asset Pricing Model* (CAPM) of rational investors is not valid always in practice.
- **Investors are not rational** This is due to several reasons: people make systematic errors in judgment, individuals risk preferences are primarily influenced by the way information is presented or "packaged," investors trade more than the theory suggests, and finally, people usually operate using inductive, not deductive, reasoning.

Despite the above drawbacks, classical theory has advanced very much our understanding of capital markets, but it appears to have approached its limits.

In [37], *Michael Mauboussin* develops a new challenging theory of *capital markets as complex adaptive systems*. This new theory was motivated by the observed fact that, as we add more players (agents) in the stock market game, something remarkably new occurs, which is the appearance of the well-known *self-organized criticality* (see Sect. 9.4). This takes place without any design or help from any outside agent, but it is the direct result of the dynamic interactions among

the agents of the system. Therefore, the system has the self-organizing feature and the other features of complex adaptive systems (aggregation, adaptive decision rules, negative- and positive-feedback loops, emergence, etc.).

The question is to see how the CAS framework resolves the above inconsistencies between the classical stock-market theory and actual practice. According to [37], the answer to this question is the following:

- The CAS model accounts for the high kurtosis ("fat rails") appearing in actual return distributions. This is because periods of stability punctuated by rapid change (attributable to criticality) is a feature of most complex adaptive systems.
- The trend persistence is a feature of most natural phenomena. Therefore, some degree of trend should be expected to occur in the stock market.
- The actual nonrationality of investors can be justified by the CAS model. Complex adaptive systems can explain market dynamics without the need to assume that investors have homogeneous expectations.
- A CAS model offers a better descriptive model of the market, the poor performance of actual portfolio managers is consistent with both the CAS model and the market-efficiency model.
- Researchers in the CAS area have developed the actual market process (e.g., [38]). These models employ agents with multiple "expectational models." Agents discard poorly performing rules in favor of more successful rules. The results of simulation show that, when the agents replace their expectational models slowly, the classical capital market-model predominates. The above CAS simulation of stock markets contributes to a better understanding of the behavior of capital markets.

To summarize, complex adaptive systems seem to offer a good description of how the capital market works. CAS predict stock-price changes more closely than what occurs in practice, while revealing why markets are so hard for investors to beat. Although the underlying assumptions of CAS are very simple, they are not so restrictive as investor rationality or lead-steer assumptions of classical market theory. Therefore, capital managers can go much less far astray with the CAS model compared to the classical stock-market efficiency model. A discussion on the electricity market as a complex adaptive system is provided in [39], where it is indicated that the electricity market is a CAS involving both the economic issues and the climatic/ecological issues.

In general, complex adaptive systems are all around us. They offer a model for thinking about our world but not for predicting what will happen. Complex adaptive systems are based on agents that contribute to the emergent operation of the system without knowing the system concept.

## 13.7 Society Is a Self-organizing System

The question whether human society is a self-organizing system has attracted the attention of sociologists and scientists for a long time [16, 40–46]. Sociologists agree that human societies have always been organized according to the primary means of subsistence, political and cultural traditions, beliefs, religion, and values.

*Kuhn* [47] has proposed studying the development of the sciences in terms of *paradigms* that historically are emerging from crises in communication. After its establishment, a paradigm starts to organize a science in terms of relevant communications and cognitions, and in terms of underlying communities. The concept of paradigms has offered sociology of science a good model for understanding self-organization as an agency at the supra-individual level. As we have seen in several places of this book, the concept of self-organization has been related to nonequilibrium thermodynamics [48]. Self-organization has also been considered in the framework of neurophysiology [49]. The use of the *self-organization* concept in sociology leads to the necessity to answer the question of the contingency of this theory as one more *paradigm* in science. One might ask what one gains or loses by using the self-organization hypothesis as a paradigm specific to the communications that they are allowed. *Loet Leydesdorff* [50–52] has argued that, in order to achieve sociological understanding with respect to the concept of paradigm and the incommensurability between paradigms, these concepts should be reformulated in terms of discourses, i.e., as communication systems. The paradigm concept refers to the possibility of self-organization in these communication systems. Leydesdorff has the opinion that there is no a priori reason to exclude sociological inference at the meta-theoretic level from this general mechanism. The inference leads to the hypothesis of self-organization as the general form of sociological scientific discourse, which is by default *chaotic*. Sociological theory itself is a reflexive scientific communication system. Self-organization teaches that the lower level variation is a necessity for a system to be able to organize itself. Only a reflexive understanding of the contingent history allows the further specification of the emerging system of reference. Both theory and methods can profit from the reflexive turn in relationships with one another. In sociology, the processes of *differentiation* and *institutionalization* are two basic processes that contribute to social self-organization. *Parsons* suggested that these two processes (i.e., internalization of cultural and social objects into the personality and the relations between the various components of society) can be understood using the same systemic relations among all stable systems of social interaction [45]. In his own words: “The phenomenon that cultural norms are internalized to personalities and institutionalized in collectivities is a case of the interpenetration of subsystems of action, in this case social system, cultural system and personality.… Hence the critical proposition is that institutionalized normative culture is an essential part of all stable systems of social interaction. Therefore, the social system and culture must be integrated in specific ways of their interpretations.” According to *Luhmann* [44], the social communication system cannot function without communicating individuals (actors), but only the message

(i.e., the action) is communicated, not the actor. Therefore, the action may have different meanings for the sending actor, the receiving actor, and the social system, because they have different systems of reference. The critical issue here is that Luhmann's theory does not include the actors in the social system, but the exchange information through interpenetration (i.e., via actions). Adding the *time dimension* to this theory, different frequencies may occur at the self-referential update within each subsystem. Social systems operate through actions by individuals (local nodes), although not all actors can participate in each update. For example, small economic transactions may have a strong impact on political processes, but these changes may be unnoticed (temporarily) for many of the actors involved.

According to *Leydesdorff* [52], the *reflexibility* needed for understanding the self-organization process (i.e., the differentiation between the instance of reflection and the reflected substance) requires a probabilistic interpretation of self-organizing social systems (e.g., Shannon's communication theory), but does not need a physical interpretation. Therefore, the reflexive analysis of societal self-organization involves the following analytical tasks [52]:

- Communication theory should be extended beyond nonequilibrium thermodynamics, i.e., to other nonequilibrium systems that do not rely on a physico-chemical interpretation.
- Substantive knowledge about what the systems communicate, how this is selected, and eventually how this stabilizes and self-organizes the social system over time has to be elaborated in each particular case.
- The reflexive nature of sociological systems has helped to understand what it means "to apply" mathematical communication theory to the social system as a special case. Actually, only sociology can understand itself reflexively as a special communication system. The natural sciences assume that data variance is provided by nature, whereas biology assumes that selection is a feature of nature. Psychology, although it shares with sociology a radical understanding of the reconstructive nature of knowledge, does not have the above self-reflexive understanding feature.

Figure 13.3a shows an example of self-organizing social systems, namely a self-organizing electronic voting system which involves two top layers: (i) self-organizing process/event layer, and (ii) IVCS interface layer, and three lower level layers (web services layer, applications/application server layer, and persistence layer). The interaction/self-organization taking place in these layers is realized through the design and implementation of several functions and applications as shown in the figure. Figure 13.3b shows a generic architecture for human-interactive, adaptive hard and soft automation systems. The nature and structure of the components of this architecture depend on the particular application concerned.

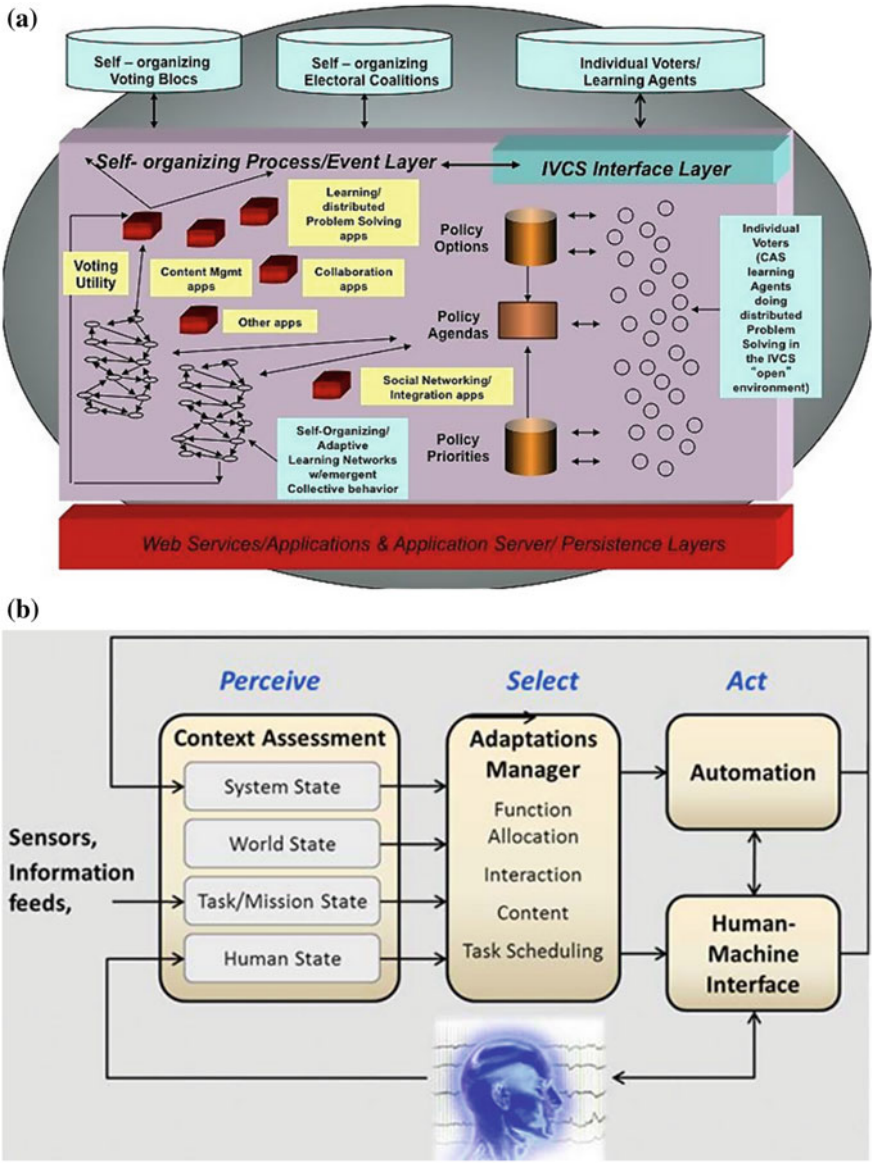



**Fig. 13.3** **a** Internet-based adaptive/self-organized voting system, **b** architecture of a generic adaptive automation system (**a**) (http://farm8.staticflickr.com/7130/6863744322_9c46f1558b_z.jpg, (**b**) (http://www.imse.iastate.edu/dorneich/files/2012/12/image1.png)

## 13.8 Knowledge Management in Self-organizing Social Systems

*Knowledge* is a process that involves three principal components:

- Cognition
- Communication
- Cooperation

The knowledge-management cycle in societal systems is shown in Fig. 13.4. It involves four stages, namely:

(i) Capturing stage
(ii) Organization stage
(iii) Assessment stage
(iv) Dissemination stage

Figure 13.5 shows a pictorial representation of three interdependent branches (pillars) of knowledge creation for system safety in the "knowledge-based society," namely: education, information, and science & technology, including the basic elements of them.

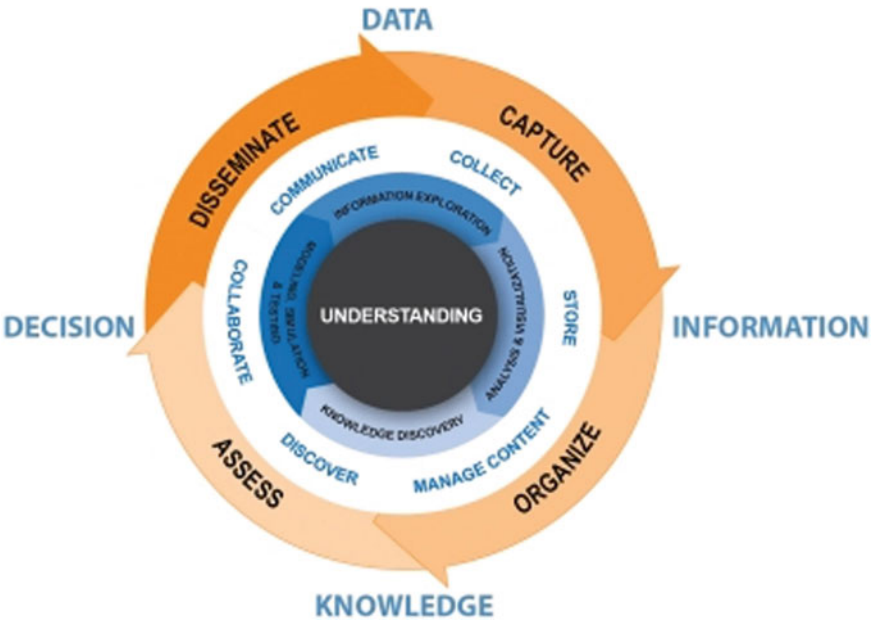



**Fig. 13.4** The knowledge-management cycle in social systems (http://www.caci.com/images/fcc/Knowledge_Management_Lifecycle.jpg)

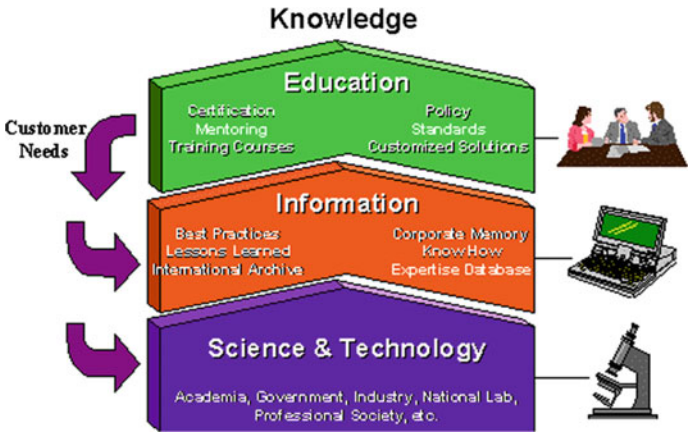



**Fig. 13.5** Pillars of knowledge in the "knowledge-based society" (http://www.system-safety.org/images/CreatingSafetyKnowledge.gif, http://www.system-safety.org/about/strategic.php)

"Education" involves the elements of certification, mentoring, training courses, policy, standards, and customized solutions.

"Information" involves the elements of best practices, lessons learned, international archives, corporate memory, know-how, and expertise database.

"Science and technology" involves the elements of academia, government, industry, national laboratories, professional society, etc. These pillars are appropriately applicable in other processes, aspects, and functions of "knowledge-based society." In particular, depending on the knowledge application, the "science and technology" pillar should be based on the knowledge of the particular scientific or technological area concerned, and/or be extended to humanistic, economic, managerial, ecological, medical, and other knowledge areas.

The knowledge involved in knowledge-based society and social systems needs to be properly managed and self-organized. Society is based on individuals. Therefore, social analysis has to start with individuals and then extended to groups, organizations, institutions, or networks. *Christian Fuchs* calls the self-organization of social systems *re-creation* [53]. Societal structures don't exist externally to, but only in and via, human agents. The interaction of human agents (actors) leads to new social qualities and structures that cannot be anticipated by merely analyzing the individual actors' performance. The internal structures influence individuals' thinking and actions (i.e., their constraints and enabling actions). This is actually a *top-down emergence*, in which new individual and group properties can emerge. The entire cycle is the fundamental process of societal self-organization which is called *re-creation* because a social system maintains and reproduces itself through permanent human agency and constraining/enabling processes. This concept was coined by *Hofkirchner* [54] and further elaborated by *Fuchs* [55–58]. Recreation means that individuals of a system persistently change their joint environment. This

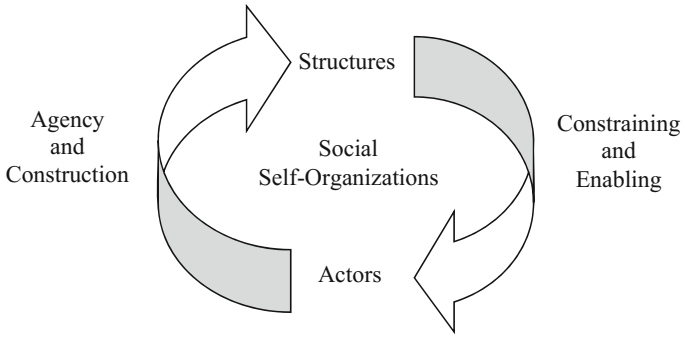



**Fig. 13.6** Structure of self-organization/re-creation

allows the system to change, maintain, adapt, and reproduce itself. The above self-organization concept of a social system is illustrated in Fig. 13.6.

*Giddens* in [59] states: "Human social activities, like some self-reproducing items in nature, are *recursive*. That is to say, they are not brought into being by social actors but continually *recreated* by them via the very means whereby they express themselves as actors." The information concept helps to explain the dynamics of self-organized units of matter: it is a relationship of reflection between a fluctuation that produces changes within a system and the structure of the system. In social systems, knowledge is the social manifestation of information. The units of organized matter are active individual or collective human actors [60].

In social systems, self-organization generates the so-called "*objective social knowledge,*" in which the social knowledge is produced in the course of the social interactions and relationships of several human actors. Objective social knowledge involves scientific/technological elements, life-support elements, and everything else that contributes to a society.

Therefore, one can classify objective social knowledge into the following types:

- Ecological knowledge
- Technological knowledge
- Economic knowledge
- Political knowledge
- Cultural knowledge

These types store existing knowledge about past social actions, and facilitate future social actions by exploiting the fundamental ways of acting socially and not needing to use exclusively new rules for each situation. On the informational level, the social interaction and production process involves the three aspects mentioned above, namely: *cognition*, *communication,* and *cooperation*. Part of subjective knowledge ("*cognition*") is communicated from one individual to the other and vice versa ("*communication*"). There is some degree of autonomy or "chance" in this process, but there is a possibility to produce new knowledge (*qualities*) as a result of synergies ("*cooperation*") between the individuals. The structural subjective

knowledge involved in the systems is coordinated, and something new emerges in a self-organization way.

We now discuss the *management of knowledge,* which appears to be of basic task in *knowledge-based society* (**KBS**). All social structures store knowledge about society: they involve a history of social relationships and enable future actions. Thus, all societies are actually knowledge-based societies. This knowledge-based character of society is enhanced continuously by the rising impact of scientific/technological advancements, knowledge-based/artificial-intelligence technologies, and accumulated expertise. A short list of basic features of knowledge follows [53]:

- Knowledge is a human and cultural product. It is a manifestation of information in the human-social environment. It does not exist in nature as such.
- Knowledge exists in both human brains and social structures and artifacts. It contains both subjective and objective elements that are mutually connected.
- Knowledge is intrinsically coupled to "not knowing" and is persistently updated and enhanced.
- Knowledge is a social not substantial public good that has a historical character.
- Knowledge production has a strongly cooperative and networked nature.
- Knowledge expands while it is used, it can be compressed, it can replace other economic resources, and it can be transmitted and transported.
- Purchasers of knowledge simply buy copies of the original data. The costs of reproducing knowledge are very low and become lower as the technology progress.
- In contrast to capital, knowledge appreciates by usage. Its marginal utility increases with its exploitation.

Social systems in the knowledge-based society possess the following characteristics:

- Complex knowledge patterns
- A networked nature
- A global (continuously enhanced) character
- Dynamic communication
- A high degree of complexity and flexibility

Thus, the question raised here is whether knowledge systems can be managed so as to secure efficiency of an organization and well-being of its members, and, if yes, what are the basic guiding rules of knowledge management. The first rule is that all kinds of human intervention should be minimized because intervention may be harmful and create problems.

*Haye* reveals the spontaneous nature of society and differentiates orders in self-forming (*spontaneous*) orders and deliberately arranged (*planned*) orders [61]. Spontaneous orders are called "*cosmos* (*world*)" and planned orders are called "*taxis* (*order*)." All cultural evolution, like natural evolution, is the result of adaptation to unexpected events and contingent situations. Social development is to a great extent "*unexpected*" and "*unavoidable.*" Self-forming orders cannot be

designed because they are produced permanently by people making many decisions independently of each other (to meet their own goals) in the complex knowledge environment. The market co-ordinates spontaneously the activities in a way that produces order (*order-out-of chaos*). The economic gain and competitive advantage that occurs for some actors (individuals or groups) is communicated to others through the market, which can then adapt to these changes. This enhances evolution. Therefore, evolution is not a humanly guided process, but a "self-forming" process. Activities of individuals could benefit other individuals not known to them. Unconscious, self-reforming order in society and markets is one side of the coin. Without successful actions of conscious co-ordination, society wouldn't be possible. Conscious, goal-directed production is a must for individuals and social beings. Humans must consciously identify their goals and find the means that lead to the achievement of these goals. This is simultaneously a conscious and a social process. Human society's existence is a purposeful existence, a conscious generation, and adaptation to natural and societal environment [55–58].

*Hayek, Luhmann,* and other scientists have argued that human intervention into self-organizing systems is neither possible nor desirable. Humans must rely on competition and adaptation to environmental and systemic effects. On the contrary, *Fuchs* considers *participatory* (coordinated/cooperative) systems design as a good alternative to such a systemic "*fatalism*" [53]. Design is an evolving process that steadily integrates new knowledge about the world, which is based on experiences in nature and society. The same approach was adopted by *Bernathy* [62] who stated that: "The notion of *empowering* people to make decisions that affect their lives and their systems is a core idea of true democracy. Much of this power today is delegated to others." Thus, for *Banathy,* the concept of *participatory system design* leads to a self-organizing and self-creating society.

*Fuchs* argued that *cooperation* in the strong sense is something much more than co-action that has the following characteristics [53]:

- Cooperating actors are mutually dependent and have many shared goals
- Cooperating actors can meet their goals faster and more efficiently than individually
- Cooperation exploits communication about common goals and about how they can be reached and involves mutual learning and common production of new realities
- Cooperation does not exclude conflicts that can be productive and constructive, if they are not escalated
- Cooperation involves interconnected and networked activity. Mutual interconnectivity and responsibility emerge

All self-organizing social systems involve mutually productive relationships of actors and social structures, and, according to the different degrees and qualities of participation and cooperation, are classified into five types, namely:

- Rigidly controlled systems
- Deterministic systems

- Purposive systems
- Heuristic systems
- Purpose-seeking systems

In conclusion, *participation* and *cooperation* are, according to *Christian Fuchs*, the two most effective (and democratic) methods for managing knowledge. The novel management principles refer to a new way of handling communication and social relationships, as well as their material effects in an organization. Social systems are self-organizing in the sense that order and knowledge emerge from "bottom-up" processes of cognition, communication, and cooperation. Self-organization is based on the creativity and activity of human being, and order emerges from decentralized "bottom-up" synergetic interactions. Managing knowledge is a basic task in KBS and can be performed in many different ways ranging between the two extremes of *hierarchical management* that is based on coercion control and steering and *social design* that is based on cooperation and participation.

## 13.9 Man-Made Self-organizing Controllers

### *13.9.1 A General Methodology*

In Sects. 13.7 and 13.8, we have discussed the role of self-organization in society and revealed the self-organization features possessed by human-social systems. We saw that participation and cooperation are the two basic ways of managing social knowledge. Technological and industrial control systems are man-made systems that are designed so as to exhibit suitable performance characteristics of accuracy, speed, reliability, and energy use. Therefore, it should be useful if man-made systems are designed so as to be self-organizing, i.e., so as to possess all the fundamental properties exhibited by natural (not man-made) systems, which assures movement from a disordered state to an ordered one. The most fundamental property is that the *structure* and *function* of the system "*emerge*" from free interactions between the elements. The purpose must not be explicitly designed, programmed, or controlled, but the system components should interact freely with each other and with the environment. The system operation should be self-adaptive for the system to go to a "*fit*" or "*preferable*" configuration (attractor), and the system's purpose to be generated is an "*emergent*" phenomenon (see Sect. 13.9.2).

A comprehensive study on the design and control of self-organizing systems is provided in [63]. A general methodology is presented and applied to design self-organizing traffic lights and self-organizing bureaucracies. This methodology receives the performance and operational requirements of a system and enables the designer to produce a system that satisfies the requirements. The methodology includes the following stages:

1. Representation
2. Modeling
3. Simulation
4. Application
5. Evaluation

These steps are not necessarily distinct and sequential because the stages merge with each other in both forward and backward directions. Backtracking takes place whenever the designer needs to go again to a previous stage for reconsideration before completing an iteration (cycle). A brief outline of the above steps is the following.

**Representation** In this step a specification (probably non-final) of the system components is selected. Actually, there may exist many different descriptions and one cannot say beforehand that a specific description is superior to another. Here the experience of the designer is crucial. The designer has to divide a system into elements (modules) with individual dynamics and goals and few interactions between elements. Dividing the system into modules implies division of the problem undergoing solution, which means that a complex task can be performed in parallel by different modules. If there exist only a few elements of interaction, then the more likely it is that the system will be understandable and predictable (i.e., the state space can be exhaustively analyzed), and the system complexity will be low. In this case, the use of traditional descriptive methods may be preferable. On the contrary, if the number of elements and interactions is high, or very high, the same is true for the system complexity, and the *representation* must be revised and improved before going to the modeling stage

**Modeling** The model should be as simple as possible and predict as much as possible. Simple models offer a better understanding of a process than complex models. Here a control mechanism must be specified that will secure that the system will do what it is required to do. In a self-organizing system, this control should be *internal* and *distributed*. The control should also be *adaptive* since a non-adaptive control mechanism would not be able to face the changes inside and outside the system. This can be done if the control mechanism is actively searching for solutions. The system must be equipped with a capability to *reduce friction* and increase synergy. Reduction of friction can be achieved by one or a mix of *courtesy, compromise, imposition* (forced courtesy), *eradication* (a special case of imposition), and *apoptosis* (programmed death as happens to cells when they are no longer needed for an organism). An increase of synergy can be achieved by the following actions: cooperation, altruism, and exploitation (forced altruism)

**Simulation** In this stage the model selected/developed in the modeling stage is simulated with appropriate computer programs. The aim of simulation is to test various scenarios and mediator strategies. The development of simulation must proceed from abstract to particular. An abstract scenario should be first tested and refined, before proceeding to a finer representation model. Simulation experiments should go from simple to extensive, i.e., proof of concept should be taken first, and

then extensive studies should be followed. Simulation should reach mature state before taking the implementation into the real world.

**Application** At this stage, the models(s) developed and tested in the previous stages are applied to a real system. This is a relatively easy task if the real system is a software system, but the application to a "material" (hardware) system will face many difficulties. Therefore the feasibility of application should be considered during all stages of design.

**Evaluation** The evaluation of the various aspects of system performance is a necessity during the application course, in order to measure the performance and compare it with the performance of the previous system(s). The system should not decide, once and for all, that its operation/solution is the best and should be able to adapt itself to the changing requirements.

Obviously, the above general methodology for designing self-organizing systems is not unique. Many other methodologies might be developed and proposed. One of them is described in [64], which has been applied to design a simple self-organizing industrial controller. This controller uses the *probability state variables* (**PSV**) for the parameter identification, which gives a signal corresponding to the parameter. **PSV** can only identify a single parameter and so, in the multi-parameter case, several identification units should be employed. The system uses a "*predictor*" (actually a proportional-plus-derivative (**PD**) controller) which receives the system error e(t) and transmits as its output the predictor error $e_p$ (t) = e (t) + Tė(t). The self-organizing controller minimizes the integral absolute predictor error (IAE), and a performance assessment unit is used that performs this minimization employing the (+1) for encouragement and (-1) as punishment according to the equation [65, 66]:

$$V = -\operatorname{sgn} e_p(t)\operatorname{sgn}\left(\dot{e}_p(t)\right)$$

The self-organizing feature of this controller is that no information about the *sign* of the controller output $e_p$ *(t)* is needed for the correct performance of the system. The controller is implemented in discrete-time (sample-data) form. The simulation results, obtained by strongly disturbing the system after reaching its steady state, showed that this controller produces very good results and can be used in practice. But obviously it is a simple controller which does not possess all the features of self-organization. A similar fuzzy controller is presented in [70, 71]. In the following, we will briefly describe the self-organizing traffic controller that was designed using the general methodology of Sect. 13.9.1 [63].

### *13.9.2 Self-organizing Traffic-Lights Control*

*Traffic congestion* is one of the major problems of modern highly populated cities. To lessen the consequences of congestion, suitable traffic-control systems have

been developed to regulate the flow of vehicles by not allowing them to go in any direction using traffic lights at street intersections. Of course, when car density saturates the streets, no traffic control is possible. Traffic systems are traditionally designed using mathematical, operational research, and computational methods to determine appropriate traffic policies (periods and phases of traffic lights) to optimize the overall system operation (time, energy consumption, driver patience, etc.). Unfortunately, despite the efforts to design intelligent traffic-light systems, many current traffic-light systems cannot cope with "*abnormal*" and "*extreme*" situations. *Carlos Gershenson* argued that traffic-light control is not so much an optimization problem, but rather an adaptation and self-organization problem. Optimization provides the best possible solution for a specific configuration and unchanged constraints. Therefore, an adaptive/self-organizing system is expected to provide a more efficient solution to the problem. The system proposed in [63] employs the general methodology outlined in Sect. 13.9.1. Specifically, the system considers the traffic lights as agents that want to "*get rid*" of cars as quickly as possible. To this end, they should avoid having green lights on empty streets and red lights on highly congestion streets. In the modeling phase, two classic methods were implemented, namely *marching* and *optim*, to compare their effectiveness with the *sotl-request* method. In the *march-step* model, all green lights are either *southbound* or *east-bound* and synchronized in time. The *optim* method is implemented trying to set the phases of traffic lights so that, as soon as a red light turns green, a car that was made to stop would find the following traffic light green. These two methods are non-adaptive because their operation is predetermined and does not take into account the actual state of the traffic. On the contrary, the *sotl-request* method enables the traffic lights to be sensitive to the current traffic condition, and, thus, respond to the needs of the oncoming vehicles. In the simulation, the *march-step* and *optim* methods were compared with the *sotl-request* method. The *sotl-request* method proved better for low-traffic densities, but very poor for high-traffic densities. For this reason, Gershenson has developed a new method, called *sotl-platoon,* which, before changing a red light to green, checks if a platoon is not crossing through, so as not to disperse it. In other words, a red light is not changed to green if, on the crossing street there is at least one car approaching at a given number of car-lengths from the intersection. The performance of the simulated model was measured using the following statistical figures:

- *Speed* (cruise speed is one patch/per time step, i.e., the speed at which cars proceed without obstruction)
- *Percentage of stopped cars*
- *Waiting time*.

The conclusions of the simulation are very briefly summarized in the following bullet points. The details can be found in [63]:

- The *marching* method is poor for low-traffic densities (with roughly less than three cars encountered between intersections) and has the best performance for very high densities (with more than eight cars between intersections)

- The *optim* method has an acceptable performance for low densities, but for high densities cars may enter a gridlock much more quickly than with the other methods.
- The *sotl-request* method has the best performance for low-traffic densities and for high traffic is very inefficient. This is because *platoons* (that are formed of observed sizes from 3 to 15 cars) can change red lights into green rapidly (in most cases, before actually arriving at the intersections), and there is a constant switching of lights, which reduces the speed of cars that are forced to stop on yellow lights and also breaks platoons (which has the result that the few cars passing have a higher probability of waiting longer at the next intersection).
- The *sotl-platoon* method can keep platoons together resulting in full synchronization for a wide range of density. The full synchronization shows how self-organizing traffic lights form platoons that modulate traffic lights, so that average car speed is maximized, waiting times are minimized, and the cars are stopped in a robust way. Moreover, the self-organizing traffic lights are efficient without *knowing* beforehand the locations and densities of the cars.

A distributed self-organizing system for urban traffic control based on swarm-self-organizing maps is contained in [67]. This system overcomes the requirement of other distributed systems to have available a special mechanism for synchronization between intersections. The proposed architecture (Fig. 13.7) involves one signal controller at each intersection in the traffic system. Communication between these controllers is essential. This architecture consists of three subsystems, namely: (1) the traffic-signal control system (TSCS), (2) the simulator, (3) the map converter (MP), (4) the vehicle sensor, and (5) the light control as the main output of TSCS. The simulator is used for verification and testing TSCS. The map converter obtains information and intersections from traffic networks, which is sent as input to the TSCS for the computation process and to the simulator for the verification and testing process. The vehicle sensor (VS) is an application that aims to detect and calculate the number of vehicles that pass through an intersection. The traffic lights are controlled via three parameters, namely: "cycle time" [the time needed for a full signal-phase cycle (red-yellow-green)], "green split" (the percentage of green assigned for each direction in a cycle), and "offset" (the difference time between starting times and green phases on successive signals). There must be a correlation between adjacent intersections with others as shown in Fig. 13.8.

Figure 13.8a, b shows a generic traffic light model used in the implementation of the control system.

This traffic controller was tested in an actual road scenario in Jakarta, and its performance was compared to the system used in the Jakarta Traffic Control System, giving favorable results.

Thirty tests were made using random input and simulation running times of 3 min. Another example of self-organizing traffic control systems that employs neural networks is described in [86].

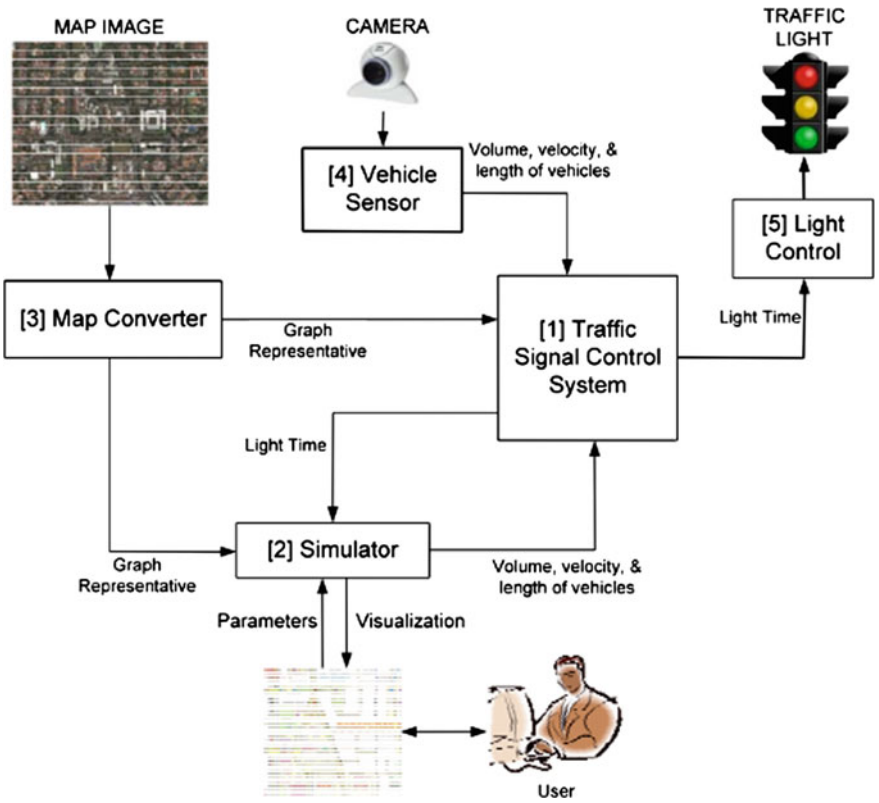



**Fig. 13.7** Architecture of self-organizing urban traffic-control system (http://www.s2is.org/Issues/v3/n3/papers/paper8.pdf)

## 13.10 Concluding Remarks

In this chapter, we have discussed a number of fundamental issues concerning the role of adaptation and self-organization in life and society. We started with a listing of adaptations of animals that enable them to live safely as much as possible and survive in their varying habitats. Then we discussed why the ecosystem falls within the framework of complex adaptive systems and the adaptation measures that are generally accepted as the minimum requirements to face the harmful impact of today's fast climate change. Next, we provided a short description of the adaptability functions of the immune systems that assure successful defense of an organism against the invasions of pathogenic bacteria, fungi, viruses, and parasites. It was then explained that social-ecological systems operate in a way similar to the immune system with the aid of the so-called "*social agents.*" Next, we outlined the basic aspects of the theory of capital markets as complex adaptive systems.

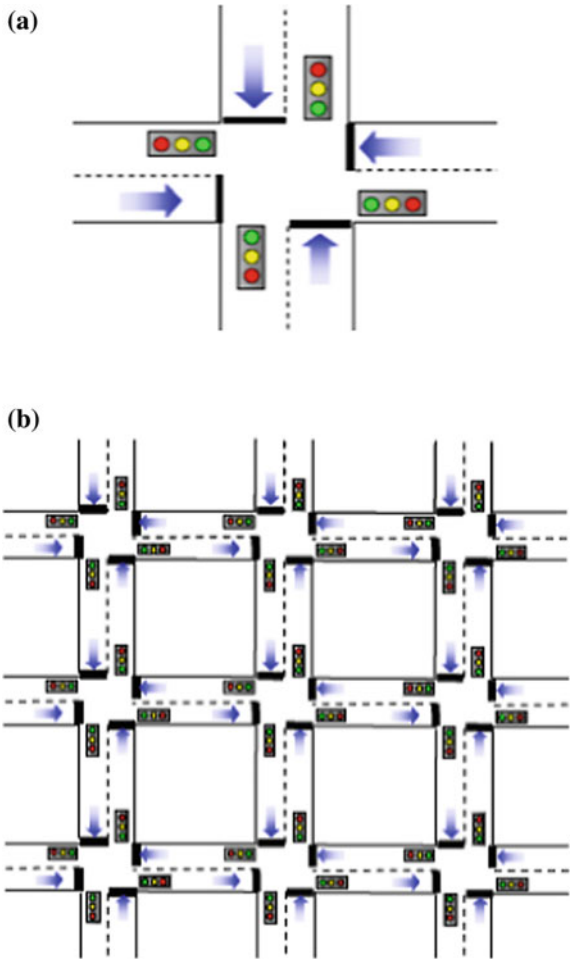



**Fig. 13.8** **a** Generic traffic light model, **b** traffic-light model with correlation between intersections

On the self-organizing side of the chapter, we first argued that human society is actually a self-organizing system based on the processes of differentiation, institutionalization, reflexibility, and communication between individuals and groups. Then we reviewed the ideas of *Christian Fuchs* about how knowledge in self-organizing (re-creation) social systems would be managed. The three processes involved in knowledge are cognition, communication, and cooperation, and, according to Fuchs, the two most effective methods for managing a knowledge-based society are participation and cooperation. The chapter ended with *Carlos Gershenson's* general methodology for designing man-made self-organizing systems and controllers and his case-study on the self-organizing control of traffic lights.

We close the chapter by mentioning three other very important areas of technological and societal application of adaptability and self-organization. These are:

- The World-Wide Web
- Bio-inspired self-organizing systems
- Self-organizing multi-agent robotic systems.

The WWW has evolved as a complex adaptive and self-organizing system characterized by scaling phenomena of the "fractal" type. The amount of information on the Web is overwhelming. Its wide distribution, openness, and high dynamics make it a really complex system. To find the information one wants is a big challenge. Therefore, the open field to develop and implement systematic integrating mechanisms of self-adaptation and self-organization is a very attractive perspective [14, 68]. In [67], it is demonstrated that the WWW possesses all the principal features of self-organizing systems, namely recursion, attractors, bifurcations, self-reference, self-similarity, self-repair, and autonomous agent performance.

Figure 13.9 shows an inclusive picture of modern interconnected societal and technological applications at both local and remote levels. The interconnection is implemented by several communication networks (WLAN, UMTS, Internet, Ad Hoc, PAN, etc.).

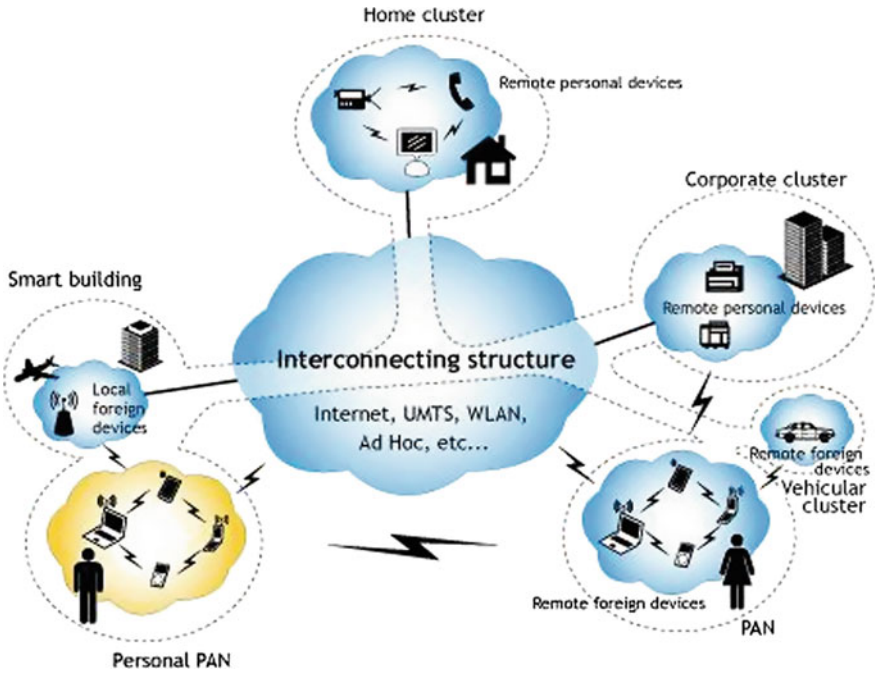



**Fig. 13.9** Representation of locally and remotely networked societal applications (http://4.bp.blogspot.com/-mR56Wqvbe4E/Ti2avgRqIPI/AAAAAAAAAlU/_C_pi6_pKcM/s1600/SONS-ANTS.jpg)

- **Bio-inspired man-made self-organizing systems** The principal task here is to find local behavior rules from which global properties emerge. Nature and biology offer many different examples of such rules of both the top-down (direct) and bottom-up (indirect) types. The visible differences between biological and technological solutions should be noted and exploited.

Lower living organisms (e.g., protozoa) do not have learning mechanisms and capabilities to interchange behavior during their lifetime. New behavior is stored via the genes of the next generation. In higher living organisms, the physical body and capabilities grow at most times as part of the solution [69, 70]. In technological systems, the hardware (which is the analog of the body) must be fixed very early in the design. If the man-made system is built with only the software part of the biological self-organization, the result may be less effective than the biological prototype. Thus, it is a mistake to adopt biological solutions only because they are more elegant. Here the field of *artificial life* is offering the needed concepts, principles, and possibilities.

- **Self-organization in multi-agent robotics** This is a very active area of robotics initiated by *R. Brooks* [71, 72]. According to Brooks, the decomposition of intelligent systems is not meant to obtain independent information processing units that must interface with each other via representations. Actually, the system is decomposed into independent and parallel activity producers that all interface with the real world via perception and action. The performance of social insects is the result of collective intelligence, which was formalized and generalized by *Sulis* [73] as consisting of a large number of quasi-independent, stochastic agents, interacting locally both among themselves, as well as with an active environment, in the absence of hierarchical organization, and yet capable of adaptive behavior [74]. The local interactions in a self-organizing system may be based on *direct* communication among agents or on *indirect* communication via stimuli originating within the environment. In the 1950s, entomologist *Piere-Paul Grase* named this indirect interaction *stigmergy* [75]. Stigmergy combined with self-organization is called "*stigmergic self-organization.*" Stigmergy appears to be the basis of several collective behaviors in social insects. This concept has been studied by many researchers interested in multi-agent robotics; it has been integrated with evolution concepts, embodied agents concepts, and continuous-dynamic systems for designing "*collectively intelligent and self-organized multi-robotic systems*" [74]. This area of robotics will have many industrial and nonindustrial applications in modern society. Figure 13.10 shows three cases of experimental robot swarms studied within the European SYMBRION project. These intelligent, symbiotic multi-robotic systems are based on bio-inspired and modern computing paradigms. They can dock with each other, share energy and computational resources, and perform complex tasks such as autonomous navigation, perception of the environment, and grasping objects.

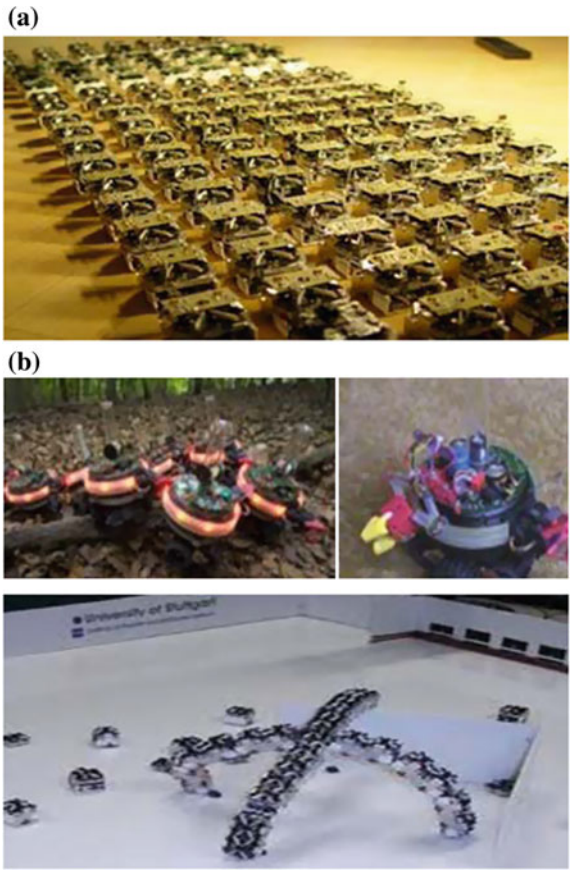



**Fig. 13.10** Three examples of robot swarms (http://www.ipvs.uni-stuttgart.de/abteilungen/bv/lehre/studentische_arbeiten/studienarbeiten/SO_industrial_micromontage/de)

Some other references in which the problem of designing man-made self-organizing systems has been treated via evolutionary methods, multidimensional Kiefer-Wolfowitz stochastic approximation, economic-political concepts, meta-data architecture techniques, and the ADELFE multi-agent technique are [76–81], respectively. Self-managing/self-organizing systems are defined and discussed in [82]. Three further references on fractals/chaos, self-organizing systems, and their relation to synergetics are [83–85].

## References

1. J.S. Huxley, *Evolution: The Modern Synthesis* (Harper and Row, New York, 1942)
2. R.M. Burian, *Adaptation: Historical Perspectives*, in ed. by E.F. Keller, E.A. Lloyd (Harvard University Press, Cambridge, MA, 1992)
3. R.C. Lewontin, Adaptation. *Sci. Am.* **239**(3), 213–230 (1978)
4. C.C. Williams, *Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought* (Princeton University Press, Princeton, NJ, 1966)

5. K. Schmidt-Nielsen, *Animal Physiology: Adaptation and Environment,* 5th edn. (Cambridge University Press, Cambridge, UK, 1997)
6. K. Noels, R. Yang, K. Saumure, Multiple Routes to Cross-Cultural Adaptation. in *Proceedings Annual Meeting of the International Communication Association,* Dresden, Germany, June 16, 2006. http://www.allacademic.com/meta/p92882_index.html (PDF 2009–05-25)
7. R. Kostelanetz, *The Edge of Adaptation: Man and the Emerging Society* (Better World Books, Mishawaka, IN, 2006)
8. D. Drbohlow, E. Janská, Immigrants and Their Adaptation Process in a New Host Society: On the Example of the Czech Republic. in *Proceedings 2nd EAPS Conference of the Working Group on International Migration Europe,* Rome, Italy, November 25–27, 2004
9. L. Shaw, Declining Human Fertility is Evolutionary Adaptatio. in *Proceedings 22*nd *Annual Conference of the European Society of Human Reproduction and Embryology,* 2004, www.eshre.com
10. H. Selye, *The Stress of Life* (Mc-Graw-Hill, New York, 1985). See Also: *Adaptation to Stress and Natural Therapies,* http://www.aapainmanage.org/literature/PainPrac/V10N3_Sandlow_AdaptationtoStress.pdf
11. R.M. Young, *Mind, Brain and Adaptation in the Nineteenth Century: Cerebral Localization and Its Biological Context from Gall to Ferrier* (Clarendon Press, Oxford, 1970; Reprinted: Oxford University Press, Oxford, 1990)
12. S.M. Saidam, On Route to an E-Society: Human Dependence on Technology and Adaptation Needs, http://www.comminit.com/en/node/242320/307
13. R.S. Lazarus, *Emotion and Adaptation* (Oxford University Press, Oxford, 2005)
14. M. Rupert, A. Rattrout, S. Hassas, The web from a complex adaptive systems perspective. J. *Comput. Syst. Sci.* **74**(2),133–145 2008
15. T.Y. Choi, K.J. Dooley, M. Rungtusanatham, *J. Oper. Manage.* **19**(3) 351–366 (2001)
16. L. Leydesdorff, Is society a self-organizing system? *J. Soc. Evol. Syst.* **16**, 331–349 (1993)
17. Animal Adaptations, http://www2.scholastic.com/browse/article.jsp?id=2840, http://teacher.scholastic.com/dirtrep/animal/index.htm, http://www.woodlands-junior.kent.sch.uk/Homework/adaptation.htm. Eco-Academy: Maimi MetaZoo and Zoological Society of Florida, Florida
18. S.A. Levin, Ecosystems and the biosphere as complex adaptive systems. *Ecosyst. Biomed. Life Sci. Earth Environ. Sci.* **1** 431–436, (1998)
19. S.A. Levin, Complex adaptive systems: Exploring the known, the unknown and the unknowable. *Am. Math. Soc.* **40**, 3–19 (2003)
20. E. Bonabeau, Social insect colonies as complex adaptive systems, *ecosystems. Biomed. Life Sci. Life Sci. Earth Environ. Sci.* **1**, 437–443 (1998)
21. G. Hartigsen, A. Kinzig, G. Peterson, Complex adaptive systems: Use and analysis of complex adaptive systems in ecosystem science: Overview of special section. *Ecosyst. Biomed. Life Sci. Earth Environ. Sci.* **1**, 427–430 (1998)
22. M. Janssen, Use of complex adaptive systems for modeling global change. *Ecosyst. Biomed. Life Sci. Earth Environ. Sci.* **1**, 457–463 (1998)
23. Environmental Change Institute (ECI). Oxford University, http://www.eci.ox.ac.uk/research/climate/adaptationsocieties.php
24. E.L. Schipper, M.P. Cigaran, M. McKenzie Hedger, Adaptation to Climate Change: The New Challenge for Development in the Developing World. in *Environmental and Energy Group UNDP*, (New York, July 2008)
25. Climate Change: Health and Environmental Effects EPA, New York, USA, http://www.epa.gov/climatechange/effects/adaptation.html
26. Joint Nature Conservation Committee, Living with Climate Change: Are there Limits to Adaptation? *Conference Proceedings,* http://www.jncc.gov.uk/page-4102, http://www.tyndall.ac.uk/research/programme3/adaptation2008/index.html
27. P. Berry, Adaptation Options on Natural Ecosystems *(UNFCC Report),* June 2007 http://unfcc.int/files/cooperation_and_support/financial_mechanism/application/pdf/berry.pdf

28. W. Easterling, B. Hurd, J. Smith, Coping with Global Climate Change: The Role of Adaptation in US. in *Pew Center on Global Climate Change*, (Arlington, VA, USA, 2004)
29. E.L. Cooper, Immune diversity throughout the animal kingdom. *Bioscience* **70**, 720–722 (1990)
30. G. Beck, G.S. Habicht, Immunity and the invertebrates. *Sci. Am.* **275**, 60–66 (1996)
31. T.W. McDale, C.M. Worthman, Evolutionary process and the ecology of human immune function. *Am. J. Human Biol.* **11**, 705–717 (1999)
32. M.A. Janssen, E.E. Osnas, Adaptive capacity of social-ecological systems: Lessons from immune systems. EcoHealth **2**, 1–10 (2005)
33. C.S. Holling, Resilience and stability of ecological systems. *Ann. Rev. Ecol. Syst.* **4**, 1–23 (1973)
34. E. Ahmed, A.H. Hashish, On modeling the immune system as a complex system theory, *Biosciences* **124**, 413–418 (2006)
35. J.C. Tay, A. Jhavar, CAFFISS: A Complex Adaptive Framework for Immune System Simulation. in *Proceedings of 2005 ACM Symposium on Applied Computing* (2005) pp 158–164
36. S. Forrest, A.S. Perelson, Genetic Algorithms and the Immune System. in *Proceedings of 1st Workshop on Parallel Problem Solving from Nature*, (Dortmund, Germany, 1991) pp 320–325
37. M.J. Mauboussin, Revisiting market efficiency: The stock market as a complex adaptive system, *J. App. Corp. Finan.* **14**(4), 47–55 (2002)
38. W.B. Arthur, Asset Pricing Under Endogenous Expectations in an Artificial Stock Market. in *The Economy of Evolving Complex Systems II,* ed. by W.B. Arthur, S.N. Durlaf, D.A. Lane (Addison-Wesley, Reading, MA, 1997)
39. Electricity Markets are Complex Adaptive Systems, http://epress.anu.edu.au/cs/chap11Grozev-final-2.jpg
40. T. Parsons, *The Structure of Social Action* (Free Press, New York, 1968)
41. F. Geyer, J. Van der Zouwen, Cybernetics and social science: Theories and research in sociocybernetics. *Kybernetes*, **20**(6) 81–92 (1991)
42. T. Imada, *Self-Organization and Society* (Springer, Berlin, 2009)
43. A. Giddens, *Central Problems of Sociology* (MacMillan, London, 1979)
44. N. Luhmann, Interpenetration-Zum Verhältnis Personaler und Sozialer Systeme, *Zeitschrift für Sociologie*, **6**, 62–76 (1978)
45. T. Parsons, *Interaction and Social Systems*, The International Encyclopedia of Social Sciences, vol. 7 (McGraw-Hill, New York, 1968) pp 429–441
46. *Global Oneness, Society-Organization of Society,* Encyclopedia II-Society, http://www.experiencefestival.com
47. T.S. Kuhn, *The Structure of Scientific Revolutions* (University of Chicago Press, Chicago, 1962)
48. I. Prigogine, I. Stengers, *Order out of Chaos* (Bantam, New York, 1984). (English translation of: 'La Nouvelle Alliance, Gallimard, Paris, 1979)
49. H.R. Maturana, *Biology of Language: The Epistemology of Reality.* ed. by G.A. Miller, E. Lenneberg, Psychology and Biology of Language and Thought (Academic Press, New York, 1978) pp. 27–63
50. L. Leydesdorff, The static and dynamic analysis of network data using information. Theory. *Soc. Netw.* **13**, 301–345 (1991)
51. L. Leydesdorff, 'Structure/ action' contingencies and the model of parallel distributed processing, *J. Theory Soc. Behav.* **23**, 47–77 (1993)
52. L. Leydesdorff, Is society a self-organizing system? *J. Soc. Evol. Syst.* **16**, 331–349 (1993)
53. C. Fuchs, Knowledge management in self-organizing social systems. *J. Know. Manage. Pract.* **5**, May 2004. http://www.tlajnc.com/articl61.htm http://fuchs.uti.at/papers/social-theory/

54. W. Hofkirchner, Emergence and the logic of explanation: An argument for the unity of science. Acta Polytechnica Scandinavica-Math. Comput. Manage. Eng. Series **91**, 23–30 (1998)
55. C. Fuchs, Social information and self-organization in the knowledge based society. *Triple C*, **1**(2), 105–169 (2003). http://triplec.uti.at
56. C. Fuchs, Structuration theory and self-organization. *Syst. Practice Action Res.* **16**(4), 133–167 (2003)
57. C. Fuchs, Co-operation and social self-organization. *Triple C,* **1**(1), 1–52 (2003)
58. C. Fuchs, A. Schlemm, The Self-Organization Society, INTAS-*Human Strategies in Complexity Research Report* (University of Salzburg, Austria, November 30, 2002)
59. A. Giddens, *The Constitution of Society* (University of California Press, Berkeley, 1984)
60. C. Fuchs, W. Hofkirchner, Knowledge self-organization and responsibility. *Kybernetes*, **34** (1–2), 244–260 (2005)
61. F.A.M. Hayek, *The Fatal Conceit: The Errors of Socialism*, in Collected Works, vol. 1 (Routledge, London, 1988)
62. B.A. Banathy, *Designing Social Systems in a Changing World* (Plenum Press, New York/ London 1996)
63. C. Gershenson, *Design and Control of Self-Organizing Systems* (CopIt ArXives, Mexicocity/ Boston, 2007)
64. S.M.H. Jamarani, M.R. Hashemi Golpayegani, design of self-organizing system controller for industrial application, *World Acad. Sci. Eng. Technol.* **12**, 14–18 (2005)
65. S.G. Tzafestas, G.G. Rigatos, Design and stability analysis of a new sliding mode fuzzy logic controller of reduced complexity. *Mach. Intell. Rob. Control.* **1**(1), 27–41 (1999)
66. S.G. Tzafestas, G.G. Rigatos, A simple robust sliding–mode fuzzy logic controller of the diagonal type. *J. Intell. Rob. Syst.* **26**(3–4), 353–388 (1999)
67. March W. Jatmiko, A. Azurat, et al., Self-organizing urban traffic control architecture with swarm-self-organizing map in Jakarta: Signal control system and simulator. *Int. J. Smart Intell. Syst.* **3**(3) (2010)
68. W. Willinger, R. Govindan, S. Jamin, V. Paxson, S. Shenker, Scaling Phenomena in the Internet: Critically Examining Criticality. *Proceedings of National-Academey Science USA*, **99**(1), pp 2573–2580 (2002). www.pnas.org/cgi/doi/10.1073/pnas012583099
69. D.L. Turcotte, J.B. Rundle, Self-organized complexity in the physical, biological and social sciences. *PNAS*, **99**(1), 2463–2465 (2002)
70. S. Camazine, J. Deneubourg, N. Franks, J. Sneyd, G. Theraulaz, E. Bonabeau, *Self-Organization in Biological Systems* (Princeton University Press, Princeton, 2001)
71. R.A. Brooks, Intelligence Without Reason, *AI Memo No. 1293*, MIT, AI Lab., MA, 1991; Also: Proceedings of IJCAI-91, Sydney, Australia (J. Mylopoulos and R. Reiter, eds.), pp 569–595 (Morgan Kaufmann San Mateo, CA, USA, 1991)
72. R.A. Brooks, Intelligence without representation, *Artificial Intell.* **47**(1–3), 139–159 (1991)
73. W. Sulis, Fundamental concepts of collective intelligence, *Nonlinear Dyn. Psychol. Life Sci.* **1** (1), 35–53 (1997)
74. E. Izquierdo-Torres, *Collective Intelligence in Multi-Agent Robotics: Stigmergy, Self-Organization and Evolution* (University of Sussex, Brighton U.K., Jan 2004)
75. P.P. Grassé, La Reconstruction du Nid et les Coordinations Interindividuelles Chez Bellicositermes Natalensis et Cubitermes Sp. La Theorie de la Stigmergie: Essai D' Interpretation du Comportoment des Termites Constucteur. *Insects Sociaux*, **6**, 41–83 (1959)
76. I. Fehérvári, W. Elemenreich, Design of Self-Organizing Systems Using Evolutionary Methods. in *Proceedings of Junior Scientist Conference (JSC'08),* (Vienna, Austria, 2008) pp 53–54
77. P.C. Badavas, G.N. Saridis, A performance-adaptive self-organizing control of a class of distributed systems, *IEEE Trans. Syst. Man Cybern.* **1**(1), 105–110 (1971)
78. D. Chistilin, Principles of Self-Organization and Sustainable Development of the World Economy are the Basis of Global Security *Institute World Economy and International*

*Relations*, Ukraine Academy of Sciences, www.necsi.edu/iccs6/viewpaper.php?id=280, www.aiecon.org/conference/aescs2009/articles/AESCS2009-17.pdf
79. G. Di. Marzo Serugendo, J. Fitzerald, A. Romanovsky, N. Guelfi, A Metadata-Based Architectural Model for Dynamical Resilient Systems. in *Proceedings of ACM Symposium on Applied Computing (SAC'07)*, (2007) pp 566–572
80. R. Frei, G. Di Marzo Serugendo, J. Barata, Designing Self-Organization for Evolvable Assembly Systems. in *Proceedings IEEE International Conference on Self-Adaptive and Self-Organizing Systems (SASO'08),* (*IEEE Computer Science*, New York, 2008)
81. G. Picard, M.-P. Gleizes, *The ADELE Methodology—Designing Adaptive Cooperative Multi-Agent Systems.* in *The Agent–Oriented Software Engineering: Methodologies and Software Engineering for Agent Systems* (Kluwer, Boston, 2004) pp 157–176
82. G. Mühl, M. Werner, M.A. Jaeger, K. Herrmann, H. Parzyjegla, On the Definitions of Self-Managing and Self-Organizing Systems. In *Proceedings of SAKS 2007: KiVS-2007 Workshop Selbstorga-nisierende, Adaptive Kontextsensitive Vezteilte Systeme* Bern, Switzerland (Springer Verlag, March 1, 2007)
83. D.L. Turcotte, *Fractals and Chaos in Geology and Geophysics* (Cambridge University Press, Cambridge, U.K., 1997)
84. H. Von Foester, On Self–Organizing Systems and Their Environments. in ed. by M.C. Yovitts S. Cameron, *Self-Organizing Systems* (Pergamon, Oxford, 1960) pp 31–50
85. H. Haken, Synergetics and the Problem of Self-Organization. in ed. by G. Roth H. Schwegler, *Self-Organizing Systems: An Interdisciplinary Approach* (Campus Verlag, Frankfurt/ New York, 1981) pp 9–13
86. T. Nataksuji, T. Kaka, Development of a self-organizing traffic control system using neural network models. *Transpor. Res. Board.* **1324**, 137–145 (1991)

# Index


**A**

Abiotic components, 499, 510, 524, 530, 553–556
Ackerman technique, 350
Action-perception cycle, 282
Adaptation, 2, 15, 31–34, 145, 171, 278, 364, 365, 376, 378, 380, 410–420, 445, 449, 454, 474, 483–485, 620, 628–630, 632–636, 646, 650, 652, 654
Adaptation measurement, 410
Adaptive-Neurofuzzy Inference System (ANFIS), 380
Adiabatic processes, 80, 99, 105, 119
Adjoint state variable, 77, 339
Aerobic respiration, 496, 497
Algebraic Riccati equation, 359, 362, 363, 373, 383
Algebraic stability criteria, 295, 314, 343
Alternative non-fossil fuels, 60
Amplitude Modulation (AM), 172–178, 185
Analog modulation–demodulation, 174, 185
Analog signal analysis and processing, 185, 205
Animal muscle energy, 506, 530
Annual energy review, 524
Antarctic food web, 501, 502
Arrow of time, 105, 132–137, 139, 140, 413
Artificial intelligence, 220, 222, 226, 233, 235, 240, 376, 446, 485, 610
Assimilation efficiency, 503
Asymptotic Bode plot, 307
ATP, 6, 118, 491, 492, 494–498, 530, 578
Automated learning, 236
Autotroph, 500, 503
Available energy, 45, 74, 97, 107, 112, 113, 126, 127, 490, 510, 516, 530

**B**

Bacteria, 5, 13, 17, 18, 41, 61, 118, 263, 492, 494, 546, 636, 637, 653
Basic control loop, 292
*Berlo model* of communication, 170
Bifurcation, 125, 445, 475, 476, 484, 597, 654
Biochemical cycling, 20, 118, 125, 490, 491, 499, 586
Biochemical pathway, 490, 499
Biocomputing/Biocomputation, 536–538, 546, 568
Biologically primary knowledge, 542
Biologically secondary knowledge, 542, 543
Biomass-based energy, 64, 506
Biomass pyramid, 503, 505
Biotic components, 499, 542
Block code, 200, 201, 203
Block diagram, 169, 177, 292, 312, 314, 340, 355, 390, 614
Bode plots, 278, 305–309, 317
Borrowing and reorganizing principle, 543–545
Branches of thermodynamics, 74, 116, 119, 126, 141

**C**

Calvin cycle, 491, 492
Canonical form, 187, 293, 298, 314, 341, 346, 374, 390
Capital, 30, 118, 122, 515, 519, 536, 560, 618, 637, 638, 645, 653
Carboxylation, 491, 492
Carnivores, 119, 499, 500, 503
Carnot cycle, 43, 105, 106, 124
Carrier molecules, 496
Catabolic process, 498




S. G. Tzafestas, *Energy, Information, Feedback, Adaptation, and Self-organization*, Intelligent Systems, Control and Automation: Science and Engineering 90, https://doi.org/10.1007/978-3-319-66999-1

Catabolism, 489, 497, 498
Cellular mobile communication system, 252, 253
Channel coding, 158, 198, 199, 205, 209
Chaos, 124, 144, 410, 422, 424, 425, 428, 432, 437, 438, 440, 445, 467, 468, 646
Charcoal, 506, 507
Classical control period, 285–287
Classical control theory, 290
Clean energy, 514
Climate change, 27, 30, 64, 522, 530, 629, 632–635, 653
Closed-loop characteristic polynomial, 303, 351
Closed-loop control system, 249, 303
Closed-loop poles, 299, 303, 318, 329
$CO_2$emission, 528, 529, 634
Coal mines, 509
Coding theory, 198
Combined source and channel coding, 198
Communication systems, 168, 170, 173, 174, 190, 193, 219, 250, 286, 287, 387, 539, 548, 639
Compensator design, 278, 317, 319, 320, 324
Complete dependence, 474
Complex Adaptive System (CAS), 409, 410, 421, 423–425, 441, 445, 447, 448, 454, 455, 462, 463, 466, 474, 481, 483, 628, 629, 631, 632, 637–639, 652
Complex plane s, 299
Complex plane z, 312
Computation of the transition matrix, 342, 344
Computer architectures, 238, 239, 242
Computer engineering, 220, 225, 238
Computer-integrated manufacturing, 551, 556, 559
Computer integrated manufacturing, 606
Computer languages, 232
Computer networks, 158, 159, 165, 168, 214, 220, 250, 253, 254, 256, 260, 262, 421
Computer programming, 226, 231
Computer science, 74, 158, 159, 161, 166, 220, 225–227, 238, 246
Computer vision, 30, 140
Concepts of thermodynamics, 75
Constant damping line, 313
Consumption efficiency, 503
Controllable canonical form, 350, 353, 372
Control via relays, 371
Convolutional codes, 200–204
Cuticle cells, 494
Cyanobacteria/chloroxybacteria, 16, 17, 492, 494
Cybernetics, 220, 278, 410, 454, 462, 470, 471, 577

**D**
Damping factor ζ, 301
Dark reactions, 491
Data acquisition systems, 248
Database types, 230
Data structures, 228, 229, 231
Decentralized optimal controller, 288, 359, 369, 383
Decentralized system, 477
Decouplability matrix, 352
Decoupling controller, 352, 386
Defuzzification, 377
Demand management, 490, 519, 520
Demodulation, 158, 173, 176, 183, 185, 188, 214
Density operator ρ, 95, 96, 99
Describing function, 278, 290, 324–326, 329, 333
Deterministic (non-stochastic) uncertainty, 369
Detritivores, 499, 500
Development of information systems, 268
Differential entropy, 90, 196, 197
Differential equation, 103, 228, 285, 289, 290, 310, 330, 340, 341, 343, 358, 374, 381, 383, 594, 616
Digital signal analysis and processing, 172, 241
Direct definition of entropy, 191, 193
Direct expert controller, 377
Direct neuro-controller, 378
Discrete-event system, 381, 389, 391
Discrete-Event System (DES) control, 391
Discrete-time system, 278, 312–316, 343, 345–347, 352, 355, 356, 371, 383
Dissipation (entropy export), 130, 139, 141, 466, 484, 489
Dissipative nonequilibrium system, 470
Distributed-parameter system, 338, 381, 383–385
DNA, 2, 3, 5–7, 9–11, 13, 16, 20, 33, 34, 124, 227, 412, 418, 446, 537–539, 544–547, 568
Dominant second-order system, 299
Duality theorem (principle), 348

**E**
Earth's energy resources, 513
Ecological system, 113, 119, 463, 478, 479, 539, 629, 636, 653
Economic cycle, 516
Economic systems, 122, 281, 394, 454, 516, 536, 601, 617, 619

Ecosystem, 29, 64, 113, 119, 214, 278, 421, 440, 468, 490, 498–502, 504, 530, 536, 539, 629, 631, 632, 635, 652
Eigenvalue, 94, 95, 228, 341, 351, 352, 354, 371, 372, 432
Eigenvalue placement controller, 349, 350
Eigenvectors, 82, 96, 203, 364, 385
Electric motor, 61, 283, 509, 513
Electric power distribution system, 556
Electronic commerce, 214, 537, 561
Embedded systems, 220, 238, 248, 249, 391
Emergence, 22, 125, 161, 410, 422, 441–445, 455, 462, 472, 474, 632, 643
Energy, 2, 3, 11, 12, 15, 17, 23, 30, 31, 34, 39–51, 53–56, 58, 59, 61, 63, 66, 70, 76, 78, 82, 84–86, 88, 91, 94, 97, 98, 103–105, 107, 109, 111, 112, 116, 118, 120, 121, 124, 126–128, 131, 141, 144, 210, 211, 250, 251, 326, 343, 344, 383, 420, 431, 466, 472, 489, 492, 494, 497–501, 505, 506, 510, 512, 513, 516, 518–520, 522, 524, 529, 530, 616, 635, 656
Energy conservation, 118, 518, 520
Energy consumption, 62, 506, 520, 522, 616, 650
Energy conversion efficiency, 503, 504
Energy flow, 112, 118, 119, 121, 490, 501, 505, 506, 515, 530, 632
Energy-handling, 2, 489, 490, 517, 647
Energy in society, 506
Energy management, 518
Energy movement, 41, 70, 73, 118, 499
Energy price, 518
Energy pyramid, 501, 503, 504
Energy resources, 54, 62, 63, 490, 505, 506, 508, 512, 524, 635
Energy saving, 518, 519, 530
Energy sources, 12, 32, 40, 50, 51, 61, 70, 518, 524, 529
Energy types, 40, 45, 70, 506, 530
Energy value, 82, 85, 87, 99, 111, 139, 168, 179, 181, 186–188, 192–194, 199, 212
Entropy as information content of a measurement, 191
Entropy concept, 43, 82, 89, 94, 97, 98, 126, 132, 144
Entropy export, 462, 483
Entropy interpretations, 74, 126
Entropy quotes, 143
Environmental organizing and linking principle, 545
Enzymatic reactions, 491, 492
Epidermis cells, 494
Equations of motion, 83
Equilibrium state, 77, 80, 83, 89, 98, 99, 103, 104, 107, 343, 467, 577, 620
Estimation criteria, 361
Evolution of life, 2, 15, 16, 34, 410
Exergy, 88, 103, 107–109, 114, 117, 122–124, 127, 137, 141, 489, 501, 516
Exhaustible energy sources, 52
Expectation operator, 362
Expert system, 234, 236–238, 269, 270, 377, 446

**F**

Feedback, 2, 12, 31–34, 159, 171, 278–282, 286, 288, 292, 295, 314, 317, 324, 338, 349, 352–354, 372, 393, 412, 415, 416, 467, 471, 576, 585, 589, 593, 595, 608, 618, 620, 632, 638
Feedback compensator, 317
Feedback control system, 278, 281, 283, 324, 326, 327, 333, 601
Finite canonical model (form), 340, 341, 351
Finite state automata, 381, 389, 393
Finite-state automata/machines control, 338
Finite-State Machines (FSM), 393
First law of thermodynamics, 44, 84
Fixed wireless communications, 251
Flow of energy, 489, 504
Food chain, 119, 489, 490, 499, 501, 503, 504, 530, 632
Food web, 501, 502, 530, 631
Forward transfer function, 292, 293
Fourth law of thermodynamics, 112, 114
Frequency demodulation, 180
Frequency domain methods, 206, 295, 308, 311, 319, 324, 616
Frequency modulation, 172, 178, 181, 183
Full-order state observer, 355
Functions of information systems, 265
Fundamental theorems of information theory, 191, 205
Fuzzy-Inference Mechanism (FIM), 377
Fuzzy (linguistic) rules, 377
Fuzzy Logic Controller (FLC), 377
Fuzzy relation, 378, 384, 390, 414, 444, 463, 490, 506, 537, 618
Fuzzy Rule Base (FRB), 377
Fuzzy set, 377, 380
Fuzzy set operations, 123, 138, 188, 241, 443, 558, 606, 611, 615
Fuzzy system, 377

**G**
Gain Margin (GM), 287, 296, 301, 304, 308, 310
Gain Scheduling Control (GSC), 363, 367
Gas turbine, 509, 514
Gauss-Markov Model (GMM), 361, 362
General linear state-space model, 340
Global warming, 61, 62, 64, 522, 529
Glycolysis, 496, 498
Glycose, 490, 494
Grammian controllability matrix, 346, 348
Green house emissions, 54, 529

**H**
Hamiltonian, 96, 358, 373
Hamiltonian function, 94, 100, 119, 208, 212, 268, 278, 292, 295, 301, 305, 318, 326, 390, 634, 640
Hamilton–Jacobi-Bellman (H-J-B) equation, 358, 373
Heated liquid, 463, 480
Herbivores, 119, 499, 500, 503
Heterotroph, 17, 500
Hierarchical optimization, 382
Hierarchy of knowledge levels, 235
Historical landmarks of information, 164, 165
History of automatic control, 283
History of energy, 42
Human muscle energy, 506, 530
Human resources, 23, 515, 604, 614
Human Society, 15, 21, 24, 31, 32, 34, 54, 165, 265, 409, 410, 421, 455, 465, 490, 505, 524, 530, 568, 628, 635, 639, 646, 653
Hurwitz stability criterion, 285, 286, 333
Hybrid neuro-fuzzy controller, 385
Hydroelectric power, 56, 57, 524

**I**
ICT in business, 123
ICT in education, 281
ICT in medicine, 54, 563
ICT in office automation, 536, 551, 552
ICT in transportation, 564
Impulse response, 203, 204, 294, 295
Indirect expert controller, 377
Indirect neurocontroller, 378
Information, 2, 3, 5–7, 10, 13, 15, 21, 31–34, 124, 126, 131, 134–136, 145, 158, 160, 168, 171, 186, 187, 190, 192–194, 207, 209, 220, 223, 229, 242, 266, 363, 381, 394, 424, 440, 453, 463, 477, 536–541, 543, 544, 546, 548–550, 561, 563, 568, 616, 633, 644, 654
Information and Communication Technology (ICT), 536, 549, 550, 560, 561, 568, 569
Information and life, 537
Information and society, 548
Information architecture, 222, 224, 267
Information science, 158, 159, 161, 168, 220, 221, 224, 271, 481
Information store principle, 543
Information system design, 271
Information system development cycle, 271
Information systems, 157–159, 161, 168, 220, 225, 246, 247, 262, 264, 268–271, 551, 561
Information technology, 157–159, 161, 166, 215, 220, 222, 225, 238, 250, 264, 266, 268, 271, 517, 548, 549, 556, 560, 561, 563, 568, 569, 604, 606
Information theory, 157–159, 161, 164, 165, 168, 169, 190, 193, 198, 208, 214, 215, 250, 445, 540–542, 549
Input-Fuzzification Unit (IFU), 377
Integrodifferential equation, 322, 383, 616
Intelligent control, 338, 375, 376, 380, 386, 485, 603
Intelligent Transportation System (ITS), 565, 566
Intensive and extensive properties, 75, 76
Interconnected system, 260, 382, 383
Internal combustion engines, 509, 512
Internet, 22, 158, 165, 200, 214, 222, 249, 250, 253, 256, 260, 262, 264, 536, 552, 561, 562, 568, 569, 654
Isentropic processes, 80
Isobaric processes, 80
Isochoric processes, 80
Isolated system (universe), 76, 85, 99, 102, 104–106, 108, 110, 126, 130, 134, 143
Isothermal processes, 80, 105

**J**
Jayne's maximum entropy principle, 211
Jordan block, 342
*Jordan* (or *modal*) *canonical form*, 340, 342, 346, 347

**K**
Kalman–Bucy filter, 361–363
Kalman decomposition, 348, 349
Knowledge-based systems, 226, 266

**L**
Lactic acid, 496, 498
Landau–Pollack bandwidth signal dimensionality theorem, 191, 205, 210

Land-food web, 501
Laplace transform, 291–293, 295, 312, 333
Large-scale lumped-parameter system, 381
L-asymptotic stability, 344
Leverier algorithm, 343
Life, 1, 2, 11, 15, 18, 21, 27, 31, 34, 60, 70, 74, 116, 122, 124, 125, 133, 141, 142, 144, 145, 158, 214, 219, 224, 235, 250, 270, 277, 359, 394, 411, 413, 419, 443, 447, 461, 485, 489, 494, 499, 510, 520, 529, 530, 536, 541, 550, 556, 568, 569, 576, 583, 601, 619, 629, 644, 652, 654
Life and human thermodynamics quotes, 144
Likelihood function, 361
Limit cycle, 329, 331, 332, 475
Linear difference equation, 311, 312
Linear quadratic control, 597
Living cell, 3, 13, 34, 91, 145, 494, 497, 536, 578
$l_1$ robust control, 290, 338, 363, 369, 371, 485
L-stability, 344, 345
Lyapunov direct method, 344
Lyapunov function, 343–345, 364, 374, 383
Lyapunov stability method, 285, 343

**M**
Magnetization, 463, 466, 478–480
Magnitude Bode plot, 306
Magnitude condition, 300, 308
Marketed energy use, 40, 522, 524, 529, 647
Markov chain, 391, 454
Mass-spring system, 330
Matter flow, 30, 34, 40, 41, 74, 76, 85, 112, 115, 121, 135, 141, 161, 410, 412, 434, 468, 512, 644
Maximum uncertainty, 463
Maxwell's demon, 74, 130–132
Mealy machine, 390
Mean square error, 361
Membership function $\mu_A(x)$, 382
Mesophyll, 494
Metabolic rate, 489, 490, 498, 578
Metabolism, 123, 489, 490, 497, 498, 503, 506, 507, 581, 582
Metabolism processes, 2, 446, 497, 498
Mitochondria, 496, 497
Mobile communications, 251, 256, 564
Model-Based Predictive Control (MBPC), 364, 368
Model matching controller, 349, 353
Model-Reference Adaptive Control (MRAC), 364, 394
Modern control period, 288
Modulation, 158, 169, 172, 173, 175, 185, 187, 189, 587
*Moore's law*, 536
Multi-Input/Multi-Output (MIMO) system, 290, 338, 345, 349, 352, 369, 370
Multilayer perceptron, 378
Multimedia Web Information Systems (MWIS), 262, 263
Multi-model adaptive filter, 363
Multiple-Model Adaptive Control (MMAC), 363

**N**
NADPH, 491, 492
Narrow limits of change principle, 543
Natural self-organizing system, 473
Negative feedback, 182, 278, 280, 333, 466, 467, 478, 483, 576–578, 581, 583, 602, 619, 620
Negative feedback system, 283
Network types, 254
Neural networks, 234, 378, 443, 471, 482, 483, 592
Neuro-control with non-supervisory learning, 380
Neuro-control with reinforcement learning, 380
Neuro-control with supervisory learning, 380
Nichols plot, 278, 310, 311, 316, 317, 320, 333
Non interacting (decoupling) controller, 349
Nonlinear control, 278, 286, 324, 338, 371, 386, 394
Nonlinear control via calculus of variations, 355, 394
Nonlinear distributed-parameter system, 385
Nonlinear system root-locus, 278, 301, 320, 333
Nonlinear systems, 122, 288, 290, 324, 326–330, 333, 340, 372, 373, 428
Nonstatistical theory of physics, 74, 82
Nuclear reactor, 50, 118, 394
Nucleotides, 5, 6, 20, 34, 537
Nyquist frequency, 206, 211
Nyquist method, 278, 301
Nyquist path, 302, 303, 316
Nyquist path on the z-plane, 313–316
Nyquist plot, 303–305, 310, 317, 319, 327, 333
Nyquist plot of describing function, 328
Nyquist–Shannon sampling theorem, 191, 205, 210, 211
Nyquist stability criterion, 278, 310

**O**
Observability theory, 346
*Observable canonical model* (*form*), 340
Office automation, 536, 551, 552
Omnivores, 499, 500
Onsager's reciprocal relations, 114, 115, 121, 141
Open-loop gain open-loop poles, 299
Operating systems, 164, 220, 238, 240, 247
Optimal control, 287, 288, 355, 356, 362, 363, 373, 385, 390, 597
Optimal cost function, 356, 358
Optimal nonlinear control, 373
Organelles, 4, 492
Organismal metabolism, 498
Output controllability, 346
Output Defuzzification Unit (ODU), 377
Output measurement, 361, 616

**P**
Palisade parenchyma cells, 494
Parallel computing, 220, 238, 243, 245
Partial differential equation, 228, 383, 435
Petri-Nets (PNs) theory, 391
Phase Bode plot, 309
Phase condition, 300
Phase demodulation, 183
Phase-lag circuit, 176, 319, 325, 474, 556
Phase-lag compensator, 318, 320
Phase-lead circuit, 320
Phase-lead compensator, 319
Phase lead-lag circuit, 324
Phase lead-lag compensator, 319
PhaseLocked–Loop (PLL) detector, 181
Phase margin, 301, 304, 308–310
Phase modulation, 172, 183
Phase-plane, 330, 332, 333
Phase-plane trajectory, 331
Photosynthesis, 18, 118, 489–494, 498, 499, 501, 506, 530
Photosynthetic pigments, 491, 494
Phytoplankton, 499, 501
PI control, 322
PID control, 321, 322, 377
PID controller tuning, 278, 320, 322–324, 365, 387
Pillars of democracy, 27, 28
Pillars of education, 34, 643
Pillars of fulfilled Living, 27, 29
Pillars of instructional technology, 30
Pillars of life and society, 620
Pillars of sustainable development, 27, 29
PN graph, 392
Polar plots, 302
Pole-placement control, 367
Pole-zero diagram, 353
Pontryagin (maximum) principle, 359, 373
Positive feedback, 68, 278, 280, 281, 467, 576, 583, 585, 586, 602, 620
Power density, 506
Power generation and distribution, 536, 551, 555
Preclassical control period, 283
Prehistoric and early control period, 284
Pre-industrial inanimate prime mover, 506
Primary consumer, 500, 501, 504
Primary energy consumption, 520, 521
Primary sector, 516, 517
Principle of optimality, 355–357
Product detector, 177
Production economy, 24–26, 30, 118, 421, 490, 530, 619
Production efficiency, 503
Production process, 515, 644
Production waste, 516
Proportional control, 285, 318, 322
Pulse Amplitude Modulation (PAM), 186
Pulse Frequency Modulation (PFM), 186, 187
Pulse modulation–demodulation, 185
Pulse transfer function, 312
Pulse Width Modulation (PWM), 186–188
Pyramid of population numbers, 503
Pyruvic acid, 496, 498

**Q**
Quadratic cost function, 358
Quadrature detector, 181–183
Quantum-mechanics entropy, 74, 94, 95, 97
Quantum mechanics entropy, 242, 467
Quaternary sector, 516, 517
Queuing theory, 391

**R**
Radial basis function network, 382, 383
Randomness as genesis principle, 543
Reachability, 349
Reachable state, 349
Reachable system, 349
Renewable-energy sources, 54
Respiration, 3, 490, 494, 496, 497, 499, 583
Reversible process, 45, 75, 79, 80, 83, 89, 94, 98, 100, 107
Riccati equation, 359, 383
RNA, 2, 3, 5, 6, 8, 20, 34, 539, 544
Robust control, 290, 363, 485
Robust controller design, 382, 386
Robust performance, 369
Robust sliding-mode control, 371

Robust stability, 369
Root locus, 277, 278, 287, 290, 293, 299–301, 314, 315, 317–320, 329, 333
Routh stability criterion, 285, 314

**S**
Sampled data (pulse) transfer function, 312
Sampled-data system, 288, 289, 311
Scarce resources, 514
*Schramm* model of communication, 170
Scientific computing, 220, 226–228, 242
Secondary sector, 516–518
Second law of thermodynamics, 88, 103, 108, 115, 121, 126, 127, 130, 131, 134, 135, 516
Second-order systems, 278, 296, 297
Sectors of modern economic systems, 516
Selective retention, 462, 468, 483, 484
Selective variety, 474, 475, 483
Self-organization, 2, 32, 34, 114, 122, 240, 422, 438, 440, 461–468, 470–472, 474–476, 479–485, 539, 628, 629, 639, 640, 643–645, 647, 649, 650, 654, 655
Self-organization mechanisms, 114, 144, 422, 461, 462, 464, 466, 467, 475, 485, 628, 639, 654
Self-organized criticality, 445, 462, 468–470, 484, 638
Self-organizing map, 463, 482, 651
Self-similarity, 422, 469, 472, 474, 484, 485, 654
Self-Tuning Control (STC), 363
Separation principle, 363
Sequence detector, 389
Series compensator, 317, 324, 354
Shannon's channel capacity theorem, 191
Shannon's source coding theorem, 191, 205, 207
Shannon–Weaver communication model, 169
Signal and image processing, 225
Signal uncertainty, 369
Similarity state transformation, 349
Sliding condition, 374, 375
Sliding-mode robust controller, 374
Smith predictor, 386–388
Software engineering, 220, 238, 246, 247
Software systems, 246, 391, 549
Source coding, 158, 198, 199, 207
Space chaos (fractals), 228, 410, 422, 425, 426, 428, 429, 434, 445
Stability, 26, 104, 119, 125, 278, 285, 286, 290, 295, 296, 314, 315, 327, 333, 343, 386, 394, 476, 480, 638
Stable limit cycle, 331, 332
State controllability, 346
State feedback control, 353
State-feedback linearization, 349, 371
State-observer design, 354
State-space equations, 338, 360
State-space model, 338, 342, 388
State-transition diagram, 389, 390
Static DES control, 393
Statistical concept of entropy, 92, 126
Steady-state error, 293, 295, 318, 320
Steady state performance, 295, 317, 318, 320
Steam turbines, 509
Step response, 287, 290, 296, 298, 368
Stochastic optimal control, 360
Stochastic signal, 360
Stochastic system, 287, 361
Stoma, 494
Structure adaptation, 411, 628
Structure of information systems, 222
Sub-critical bifurcation, 476
Substantiative role of information, 538, 539
Sun's radiant energy, 489
Supercritical bifurcation, 475
Sustainability, 29, 30, 52, 62, 113, 518, 520, 529, 530
Synergetics, 466, 472
System state, 77, 82, 339, 354, 364, 470
System with time delays, 381, 386

**T**
Telecommunications, 167, 220, 225, 249–251, 260, 548, 560
Tertiary sector, 516–518
Theoretical computer science, 225–227
Thermodynamic equilibrium, 74, 75, 77–79, 89, 92, 93, 107, 108, 121, 125, 128–130, 132
Thermodynamic properties, 77, 92
Thermodynamics, 39, 40, 42, 44, 45, 70, 73–76, 82–85, 88, 92, 97, 100, 103, 107, 108, 111, 112, 114, 116, 118–128, 130–132, 135, 138, 141, 142, 144, 145, 165, 413, 489, 516, 639
Thermodynamics general quotes, 141
Thermoeconomics, 118, 122, 516
Third law of thermodynamics, 111
Time chaos (strange attractors), 410
Time-Delay Systems *(TDS)* control, 338
Time response, 301, 331, 332, 343
Time series expansion, 588
Time-varying systems, 338, 340, 343
Trajectory-tracking control, 374

Trans-critical bifurcation, 475, 476
Transducer, 389, 390
Transfer function, 285, 286, 289, 290, 292, 293, 295, 297, 299, 301, 305, 311, 312, 318, 320, 322, 323, 341, 353, 354, 369, 370, 379, 387, 592, 616
Transient response performance, 296
Transition matrix, 342, 344
Transmission of genetic information, 541
Transmission sense of information in biology, 536, 540
Trophic levels, 500, 502, 631, 632
Types of information Systems, 265, 266

**U**
Unit circle, 310, 313–316
Unit ramp response, 322
Universe, 20, 21, 41, 45, 74–76, 79, 85, 108, 120, 121, 125, 135, 136, 139, 161, 471
Unstable limit cycle, 331, 332

**V**
Van der Pol limit cycle, 332, 333
Violent manifestations of Earth's energy, 65

**W**
Web-Based Information Systems (WIS), 262
Web browser, 228, 261, 263
Wind power, 524
Wireless telecommunications, 251
Wood biomass, 507, 509
World Wide Web (WWW), 224, 250, 260, 264, 569

**Z**
Zeroth law of thermodynamics, 84
Ziegler-Nichols PID controller tuning, 317
Zooplankton, 499
Z-transform, 312