

W.F. Lawless · Ranjeev Mittu  
Donald Sofge · Stephen Russell *Editors*

# Autonomy and Artificial Intelligence: A Threat or Savior?

 Springer

# Autonomy and Artificial Intelligence: A Threat or Savior?

W.F. Lawless • Ranjeev Mittu • Donald Sofge  
Stephen Russell  
Editors

# Autonomy and Artificial Intelligence: A Threat or Savior?

 Springer

*Editors*

W.F. Lawless  
Paine College  
Augusta, GA, USA

Ranjeev Mittu  
Naval Research Laboratory  
Washington, DC, USA

Donald Sofge  
Naval Research Laboratory  
Washington, DC, USA

Stephen Russell  
U.S. Army Research Laboratory  
Adelphi, MD, USA

ISBN 978-3-319-59718-8      ISBN 978-3-319-59719-5 (eBook)  
DOI 10.1007/978-3-319-59719-5

Library of Congress Control Number: 2017947297

© Springer International Publishing AG 2017

Chapters 1, 4, 5, 12, and 13 were created within the capacity of an US governmental employment. US copyright protection does not apply.

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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

This book derives from two Association for the Advancement of Artificial Intelligence (AAAI) symposia; the first symposium on “Foundations of Autonomy and Its (Cyber) Threats—From Individuals to Interdependence” was held at Stanford University from March 23 to 25, 2015, and the second symposium on “AI and the Mitigation of Human Error—Anomalies, Team Metrics and Thermodynamics” was held again at Stanford University from March 21 to 23, 2016. This book, titled *Autonomy and Artificial Intelligence: A Threat or Savior?*, combines and extends the themes of both symposia. Our goal for this book is to deal with the current state of the art in autonomy and artificial intelligence by examining the gaps in the existing research that must be addressed to better integrate autonomous and human systems. The research we present in this book will help to advance the next generation of systems that are already planned ranging from autonomous platforms and machines to teams of autonomous systems to provide better support to human operators, decision-makers, and the society.

This book explores how artificial intelligence (AI), by leading to an increase in the autonomy of machines and robots, is offering opportunities for an expanded but uncertain impact on society by humans, machines, and robots. To help readers better understand the relationships between AI, autonomy, humans, and machines that will help society reduce human errors in the use of advanced technologies (e.g., airplanes, trains, cars), this edited volume presents a wide selection of the underlying theories, computational models, experimental methods, and field applications. While other books deal with these topics individually, this book is unique in that it unifies the fields of autonomy and AI and frames them in the broader context of effective integration for human-autonomous machine and robotic systems.

The **introduction** in this volume begins by describing the current state of the art for research in AI, autonomy, and cyber-threats presented at Stanford University in the spring of 2015 (copies of the technical articles are available from AAAI at <http://www.aaai.org/Symposia/Spring/sss15symposia.php#ss03>; a link to the agenda for the symposium in 2015 along with contact information for the invited speakers and regular participants is at <https://sites.google.com/site/foundationsofautonomy-aaais2015/>) and for research in AI, autonomy, and error mitigation presented at the

same university in the spring of 2016 (copies of the technical articles are available from AAAI at <http://www.aaai.org/Symposia/Spring/sss16symposia.php#ss01>; a link to the agenda and contact information for the invited speakers and regular participants is at <https://sites.google.com/site/aiandthemitigationofhumanerror/>).

After introducing the themes in this book and the contributions from world-class researchers and scientists, individual chapters follow where they elaborate on key research topics at the heart of effective human-machine-robot-systems integration. These topics include computational support for intelligence analyses; the challenge of verifying today's and future autonomous systems; comparisons between today's machines and autism; implications of human-information interaction on artificial intelligence and errors; systems that reason; the autonomy of machines, robots, and buildings; and hybrid teams, where hybrid reflects arbitrary combinations of humans, machines, and robots.

The contributions to this volume are written by leading scientists across the field of autonomous systems research, ranging from industry and academia to government. Given the broad diversity of the research in this book, we strove to thoroughly examine the challenges and trends of systems that implement and exhibit AI; social implications of present and future systems made autonomous with AI; systems with AI seeking to develop trusted relationships among humans, machines, and robots; and effective human systems integration that must result for trust in these new systems and their applications to increase and to be sustained.

A brief summary of the AAAI symposia in the spring of 2015 and the spring of 2016 is presented below.

## **Spring 2015: Foundations of Autonomy and Its (Cyber) Threats—From Individuals to Interdependence**

### ***Spring 2015: Organizing Committee***

Ranjeev Mittu ([ranjeev.mittu@nrl.navy.mil](mailto:ranjeev.mittu@nrl.navy.mil)), Naval Research Laboratory

Gavin Taylor ([taylor@usna.edu](mailto:taylor@usna.edu)), US Naval Academy

Donald Sofge ([don.sofge@nrl.navy.mil](mailto:don.sofge@nrl.navy.mil)), Naval Research Laboratory, Navy Center for Applied Research in Artificial Intelligence

William F. Lawless ([wlawless@paine.edu](mailto:wlawless@paine.edu)), Paine College, Departments of Math and Psychology

### ***Spring 2015: Program Committee***

- David Atkinson ([datkinson@ihmc.us](mailto:datkinson@ihmc.us)), Senior Research Scientist, Florida Institute for Human and Machine Cognition

- Lashon B. Booker (booker@mitre.org), Ph.D., Senior Principal Scientist, The MITRE Corporation
- Jeffery Bradshaw (jbradshaw@ihmc.us), Senior Research Scientist, Florida Institute for Human and Machine Cognition
- Michael Floyd (michael.floyd@knexusresearch.com), Knexus Research
- Sharon Graves (sharon.s.graves@nasa.gov), NASA Deputy Project Manager, Safe Autonomous Systems Operations, Aeronautics Research Directorate
- Vladimir Gontar (galita@bgu.ac.il), Department of Industrial Engineering and Management, Ben-Gurion University of the Negev
- L. Magafas (lomagafas@otenet.gr), Director of Electronics and Signal Processing Lab., Eastern Macedonia and Thrace Institute of Technology, Kavala, GR
- Bolivar Rocha (bolivar.rocha@gmail.com), Brazil
- Satyandra K. Gupta (skgupta@umd.edu), Director, University of Maryland Robotics Center, Department of Mechanical Engineering and Institute for Systems Research
- Laurent Chaudron (laurent.chaudron@polytechnique.org), Director, ONERA Provence Research Center, French Air Force Academy
- Charles Howell (howell@mitre.org), Chief Engineer for Intelligence Programs and Integration, National Security Engineering Center, The MITRE Corporation
- Jennifer Burke (jennifer.l.burke2@boeing.com), Manager, Human-System Integrated Technologies, Boeing Research and Technology
- Tsuyoshi Murata (murata@cs.titech.ac.jp), Dept. of Computer Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology
- Julie Marble (julie.marble@navy.mil), Office of Naval Research, Program Officer for Hybrid Human-Computer Systems
- Doug Riecken (dougriecken@gmail.com), Columbia University Center for Computational Learning Systems
- Catherine Tessier (Catherine.Tessier@onera.fr), Senior Researcher, Dept. of Systems Control and Flight Dynamics, French Aerospace Lab, ONERA, Toulouse, France
- Simon Parsons (s.d.parsons@liverpool.ac.uk), Liverpool, Visiting Professor, Dept. of Computer Science, University of Liverpool; Dept. Graduate Deputy Chair and Co-Dir., Agents Lab, Brooklyn College
- Ciara Sibley (ciara.sibley@nrl.navy.mil), Engineering Research Psychologist, Naval Research Laboratory, Washington, DC

### ***Spring 2015: Invited Keynote Speakers***

- Gautam Trivedi (gautam.trivedi@nrl.navy.mil) and Brandon Enochs (brandon.enochs@nrl.navy.mil), Naval Research Laboratory, “Detecting, Analyzing and Locating Unauthorized Wireless Intrusions into Networks”
- Chris Berka (chris@b-alert.com), Advanced Brain Monitoring, “On the Road to Autonomy: Evaluating and Optimizing Hybrid Team Dynamics”

- Kristin E. Schaefer (kristin.e.schaefer2.ctr@mail.mil), US Army Research Lab (ARL), “Perspectives of Trust: Research at the US Army Research Laboratory”
- David R. Martinez (DMartinez@LL.mit.edu), Lincoln Laboratory, Massachusetts Institute of Technology, “Cyber Anomaly Detection with Machine Learning”
- Vladimir Gontar (vgontar@ucsd.edu), BioCircuits Institute, University of California San Diego (UCSD), Ben-Gurion University of the Negev, “Artificial Brain Systems Based on Neural Networks Discrete Chaotic Biochemical Reactions Dynamics and Its Application to Conscious and Creative Robots”

### *Spring 2015: Regular Speakers*

- Christopher A. Miller (cmiller@sift.net), Smart Information Flow Technologies, “Delegation, Intent, Cooperation and Their Failures”
- Ciara Sibley<sup>1</sup> ([ciara.sibley@nrl.navy.mil](mailto:ciara.sibley@nrl.navy.mil)), Joseph Coyne<sup>1</sup> ([joseph.coyne@nrl.navy.mil](mailto:joseph.coyne@nrl.navy.mil)), and Jeffery Morrison<sup>2</sup> ([jeffrey.morrison@nrl.navy.mil](mailto:jeffrey.morrison@nrl.navy.mil)), <sup>1</sup>Naval Research Laboratory, <sup>2</sup>Office of Naval Research, “Research Considerations for Managing Future Unmanned Systems”
- Gavin Taylor ([taylor@usna.edu](mailto:taylor@usna.edu)), Kawika Barabin, and Kent Sayre, Computer Science Department, US Naval Academy, Annapolis, MD 21402-5002, “An Application of Reinforcement Learning to Supervised Autonomy”
- David J. Atkinson ([datkinson@ihmc.us](mailto:datkinson@ihmc.us)), Florida Institute for Human and Machine Cognition, Ocala, FL, “Emerging Cyber-Security Issues of Autonomy and the Psychopathology of Intelligent Machines”
- Olivier Barthelemy<sup>1</sup> ([olivier.barteye@intradef.gouv.fr](mailto:olivier.barteye@intradef.gouv.fr)) and Laurent Chaudron<sup>2</sup> ([laurent.chaudron@polytechnique.org](mailto:laurent.chaudron@polytechnique.org)), CREC St-Cyr<sup>1</sup> and ONERA<sup>2</sup>, “Risk Management Systems Must Provide Automatic Decisions for Crisis Computable Algebras”
- William F. Lawless ([wlawless@paine.edu](mailto:wlawless@paine.edu)), Paine College, Augusta, GA, and Ira S. Moskowitz, Ranjeev Mittu, and Donald A. Sofge ([ira.moskowitz@nrl.navy.mil](mailto:ira.moskowitz@nrl.navy.mil); [ranjeev.mittu@nrl.navy.mil](mailto:ranjeev.mittu@nrl.navy.mil); [donald.sofge@nrl.navy.mil](mailto:donald.sofge@nrl.navy.mil)), Naval Research Laboratory, Washington, DC, “A Thermodynamics of Teams: Towards a Robust Computational Model of Autonomous Teams”
- Ranjeev Mittu<sup>1</sup> ([ranjeev.mittu@nrl.navy.mil](mailto:ranjeev.mittu@nrl.navy.mil)) and Julie Marble<sup>2</sup> ([julie.marble@nrl.navy.mil](mailto:julie.marble@nrl.navy.mil)), <sup>1</sup>Naval Research Laboratory, Information Technology Division, Washington, DC; <sup>2</sup> Office of Naval Research, VA 22203-1995 (changing to Johns Hopkins Applied Physics Lab, MD), “The Human Factor in Cybersecurity: Robust and Intelligent Defense”
- Myriam Abramson ([myriam.abramson@nrl.navy.mil](mailto:myriam.abramson@nrl.navy.mil)), Naval Research Laboratory, Washington, DC, “Cognitive Fingerprints”
- Ira S. Moskowitz<sup>1</sup> ([ira.moskowitz@nrl.navy.mil](mailto:ira.moskowitz@nrl.navy.mil)), William F. Lawless<sup>2</sup>, ([wlawless@paine.edu](mailto:wlawless@paine.edu)), Paul Hyden<sup>1</sup> ([paul.hyden@nrl.navy.mil](mailto:paul.hyden@nrl.navy.mil)), Ranjeev Mittu<sup>1</sup> ([ranjeev.mittu@nrl.navy.mil](mailto:ranjeev.mittu@nrl.navy.mil))



[jeev.mittu@nrl.navy.mil](mailto:jeev.mittu@nrl.navy.mil)), and Stephen Russell<sup>1</sup> ([stephen.m.russell8.civ@mail.mil](mailto:stephen.m.russell8.civ@mail.mil)), <sup>1</sup>Information Management and Decision Architectures Branch, Naval Research Laboratory, Washington, DC; <sup>2</sup>Departments of Mathematics and Psychology, Paine College, Augusta, GA, “A Network Science Approach to Entropy and Training”

- Boris Galitsky ([bgalitsky@hotmail.com](mailto:bgalitsky@hotmail.com)), Knowledge Trail Inc., San Jose, CA, “Team Formation by Children with Autism”
- Olivier Bartheye<sup>1</sup> ([olivier.barteye@intradef.gouv.fr](mailto:olivier.barteye@intradef.gouv.fr)) and Laurent Chaudron<sup>2</sup> ([laurent.chaudron@polytechnique.org](mailto:laurent.chaudron@polytechnique.org)), CREC St-Cyr<sup>1</sup> and ONERA<sup>2</sup>, “Algebraic Models of the Self-Orientation Concept for Autonomous Systems”

## **Spring 2016: AI and the Mitigation of Human Error— Anomalies, Team Metrics and Thermodynamics**

### *Spring 2016: Organizing Committee*

Ranjeev Mittu ([ranjeev.mittu@nrl.navy.mil](mailto:ranjeev.mittu@nrl.navy.mil)), Naval Research Laboratory

Gavin Taylor ([taylor@usna.edu](mailto:taylor@usna.edu)), US Naval Academy

Donald Sofge ([don.sofge@nrl.navy.mil](mailto:don.sofge@nrl.navy.mil)), Naval Research Laboratory

William F. Lawless ([wlawless@paine.edu](mailto:wlawless@paine.edu)), Paine College, Departments of Math and Psychology

### *Spring 2016: Program Committee (duplicates the spring 2015 symposium)*

#### **Spring 2016: Invited Keynote Speakers**

- Julie Adams ([julie.a.adams@vanderbilt.edu](mailto:julie.a.adams@vanderbilt.edu)), Vanderbilt University, Associate Professor of Computer Science and Computer Engineering, Electrical Engineering and Computer Science Department, “AI and the Mitigation of Error”
- Stephen Russell ([stephen.m.russell8.civ@mail.mil](mailto:stephen.m.russell8.civ@mail.mil)), Chief, Battlefield Information Processing Branch, US Army Research Lab, MD, “Human Information Interaction, Artificial Intelligence, and Errors”
- James Llinas ([llinas@buffalo.edu](mailto:llinas@buffalo.edu)), SUNY at Buffalo, “An Argumentation-Based System Support Toolkit for Intelligence Analyses”
- Martin Voshell ([mvosshell@cra.com](mailto:mvosshell@cra.com)), Charles River Analytics, “Multi-Level Human-Autonomy Teams for Distributed Mission Management”

## Spring 2016: Regular Speakers

- Ira S. Moskowitz (ira.moskowitz@nrl.navy.mil), NRL; “Human-Caused Bifurcations in a Hybrid Team—A Position Paper”
- Paul Hyden (paul.hyden@nrl.navy.mil), NRL, “Fortification Through Topological Dominance: Using Hop Distance and Randomized Topology Strategies to Enhance Network Security”
- Olivier Bartheye (olivier.barteye@intradef.gouv.fr), CREC St-Cyr, and Laurent Chaudron (laurent.chaudron@polytechnique.org), ONERA, “Epistemological Qualification of Valid Action Plans for UGVs or UAVs in Urban Areas”
- William F. Lawless, (wlawless@paine.edu), Paine College, “AI and the Mitigation of Error: A Thermodynamics of Teams”

## Questions for Speakers and Attendees at AAI-2015 and AAI-2016 and for Readers of This Book

Our spring AAI-2015 and AAI-2016 symposia offered speakers opportunities with AI to address the intractable, fundamental questions about cybersecurity, machines and robots, autonomy and its management, the malleability of preferences and beliefs in social settings, or the application of autonomy for hybrids at the individual, group, and system levels.

A list of unanswered fundamental questions included:

- Why have we yet to determine from a theoretical perspective the principles underlying individual, team, and system behaviors?
- Can autonomous systems be controlled to solve the problems faced by teams while maintaining defenses against threats and minimizing mistakes in competitive environments (e.g., cyber attacks, human error, system failure)?
- Do individuals seek to self-organize into autonomous groups like teams in order to better defend against attacks (e.g., cyber, merger, resources) or for other reasons (e.g., least entropy production (LEP) and maximum entropy production (MEP))?
- What does an autonomous organization need to predict its path forward and govern itself? What are the AI tools available to help an organization be more adept and creative?
- What signifies adaptation? For AI, does adaptation at an earlier time prevent or moderate adaptive responses to newer environmental changes?
- Is the stability state of hybrid teams the single state that generates the MEP rate?
- If social order requires MEP, and if the bistable perspectives present in debate (courtrooms, politics, science) lead to stable decisions, is the chosen decision an LEP or MEP state?

- Considering the evolution of social systems (e.g., in general, Cuba, North Korea, and Palestine have not evolved), are the systems that adjust to MEP the most efficient?

In addition, new threats may emerge due to the nature of the technology of autonomy itself (as well as the breakdown in traditional verification and validation (V&V) and test and evaluation (T&E) due to the expanded development and application of AI). This nature of advanced technology leads to other key AI questions for consideration now and in the future:

### **Fault Modes**

- Are there new types of fault modes that can be exploited by outsiders?

### **Detection**

- How can we detect that an intelligent, autonomous system has been or is being subverted?

### **Isolation**

- What is a “fail-safe” or “fail-operational” mode for an autonomous system, and can it be implemented?
- Implication of cascading faults (AI, system, cyber)

### **Resilience and Repair**

- What are the underlying causes of the symptoms of faults (e.g., nature of the algorithms, patterns of data, etc.)?

### **Consequences of Cyber Vulnerabilities**

- Inducement of fault modes
- Deception (including false flags)
- Subversion
- The human/social element (reliance, trust, and performance)

We invited speakers and attendants at our two symposia to address the following more specific AI topics (as we invite readers of this book to consider):

- Computational models of autonomy (with real or virtual individuals, teams, or systems) and performance (e.g., metrics, MEP) with or without interdependence, uncertainty, and stability
- Computational models that address autonomy and trust (e.g., the trust by autonomous machines of human behavior or the trust by humans of autonomous machine behavior)
- Computational models that address threats to autonomy and trust (cyber attacks, competitive threats, deception) and the fundamental barriers to system survivability (e.g., decisions, mistakes, etc.)

- Computational models for the effective or efficient management of complex systems (e.g., the results of decision-making, operational performance, metrics of effectiveness, efficiency)
- Models of multi-agent systems (e.g., multi-UAVs, multi-UxVs, model verification and validation) that address autonomy (e.g., its performance, effectiveness, and efficiency).

For future research projects and symposia (e.g., our symposium in 2017 on “Computational Context: Why It’s Important, What It Means, and Can It Be Computed?”; see <http://www.aaai.org/Symposia/Spring/sss17symposia.php#ss03>), we invite readers to consider other questions or topics from individual (e.g., cognitive science, economics), machine learning (ANNs; GAs), or interdependent (e.g., team, firm, system) perspectives.

After the AAAI-spring symposia in 2015 and 2016 were completed, the symposia presentations and technical reports and the book took on separate lives. The following individuals were responsible for the proposal submitted to Springer after the symposia, for the divergence between the topics considered by the two, and for editing this book that has resulted:

Augusta, GA, USA  
Washington, DC, USA  
Adelphi, MD, USA  
Washington, DC, USA

W.F. Lawless  
Ranjeev Mittu  
Donald Sofge  
Stephen Russell

# Contents

<b>1 Introduction</b> . . . . .	1
W.F. Lawless, Ranjeev Mittu, Stephen Russell, and Donald Sofge	
<b>2 Reexamining Computational Support for Intelligence Analysis: A Functional Design for a Future Capability</b> . . . . .	13
James Llinas, Galina Rogova, Kevin Barry, Rachel Hingst, Peter Gerken, and Alicia Ruvinsky	
<b>3 Task Allocation Using Parallelized Clustering and Auctioning Algorithms for Heterogeneous Robotic Swarms Operating on a Cloud Network</b> . . . . .	47
Jonathan Lwowski, Patrick Benavidez, John J. Prevost, and Mo Jamshidi	
<b>4 Human Information Interaction, Artificial Intelligence, and Errors</b> . . . . .	71
Stephen Russell, Ira S. Moskowitz, and Adrienne Raglin	
<b>5 Verification Challenges for Autonomous Systems</b> . . . . .	103
Signe A. Redfield and Mae L. Seto	
<b>6 Conceptualizing Overtrust in Robots: Why Do People Trust a Robot That Previously Failed?</b> . . . . .	129
Paul Robinette, Ayanna Howard, and Alan R. Wagner	
<b>7 Research Considerations and Tools for Evaluating Human- Automation Interaction with Future Unmanned Systems</b> . . . . .	157
Ciara Sibley, Joseph Coyne, and Sarah Sherwood	
<b>8 Robots Autonomy: Some Technical Issues</b> . . . . .	179
Catherine Tessier	
<b>9 How Children with Autism and Machines Learn to Interact</b> . . . . .	195
Boris A. Galitsky and Anna Parnis	

- 10 Semantic Vector Spaces for Broadening Consideration of Consequences . . . . . 227**  
Douglas Summers-Stay
- 11 On the Road to Autonomy: Evaluating and Optimizing Hybrid Team Dynamics . . . . . 245**  
Chris Berka and Maja Stikic
- 12 Cybersecurity and Optimization in Smart “Autonomous” Buildings . . . . . 263**  
Michael Mylrea and Sri Nikhil Gupta Gouriseti
- 13 Evaluations: Autonomy and Artificial Intelligence: A Threat or Savior? . . . . . 295**  
W.F. Lawless and Donald A. Sofge

# Chapter 1

## Introduction

**W.F. Lawless, Ranjeev Mittu, Stephen Russell, and Donald Sofge**

Two *Association for the Advancement of Artificial Intelligence* (AAAI) symposia, organized and held at Stanford in 2015 and 2016, are reviewed separately. After the second of these two symposia was completed, the conference organizers solicited book chapters from those who participated, as well as more widely, but framed by these two symposia. In this introduction, we review briefly the two symposia and then individually introduce the contributing chapters that follow.

### 1.1 Background of the 2015 Symposium

Our symposium at Stanford in 2015, titled the “Foundations of autonomy and its (Cyber) threats: From individuals to interdependence”, was organized by Ranjeev Mittu, Branch Head, Information Management and Decision Architectures Branch, Information Technology Division, US Naval Research Laboratory; Gavin Taylor, Computer Science, US Naval Academy; Donald Sofge, Computer Scientist, Distributed Autonomous Systems Group, Navy Center for Applied Research in Artificial Intelligence, US Naval Research Laboratory; and W.F. Lawless, Paine College, Departments of Mathematics and Psychology.

---

W.F. Lawless (✉)  
Paine College, 1235 15th Street, Augusta, GA 30901, USA  
e-mail: [w.lawless@icloud.com](mailto:w.lawless@icloud.com)

R. Mittu • D. Sofge  
US Naval Research Laboratory, 4555 Overlook Ave SW, Washington, DC 20375, USA  
e-mail: [ranjeev.mittu@nrl.navy.mil](mailto:ranjeev.mittu@nrl.navy.mil); [don.sofge@nrl.navy.mil](mailto:don.sofge@nrl.navy.mil)

S. Russell  
US Army Research Laboratory, 2800 Powder Mill Road, Adelphi, MD 20783-1197, USA  
e-mail: [stephen.m.russell8.civ@mail.mil](mailto:stephen.m.russell8.civ@mail.mil)

The 2015 symposium on the foundations of autonomy addressed the increasing use of artificial intelligence (AI) to manage and reduce the cyber threats to complex systems composed of individual machines and robots; and, in a new shift, teams, including hybrid teams composed arbitrarily of humans, machines, and robots. Already, AI has been useful in modeling the cyber defenses of individuals, organizations, and institutions, as well as the management of social systems. However, foundational problems remain for the continuing development of AI with autonomy for individual agents and teams, especially with objective measures able to optimize their function, performance and composition.

AI approaches often attempt to address autonomy by modeling aspects of human decision-making or behavior. Behavioral theory is either based on modeling the individual, such as through cognitive architectures or, more rarely, through group dynamics and interdependence theory. Approaches focusing on the individual assume that individuals are more stable than the social interactions in which they engage. Interdependence theory assumes the opposite, that a state of mutual dependence among participants in an interaction affects the individual and group beliefs and behaviors of participants whether these behaviors are perceived or not. The latter is conceptually more complex, but both approaches must satisfy the demand for manageable outcomes as autonomous agents, teams or systems grow in importance and number. Prediction in social systems is presently considered a human skill that can be enhanced (Tetlock and Gardner 2015). But the skill of prediction in social affairs has been found to be wanting, whether in political polling, economics, or government policies (reviewed in Lawless 2016).

Despite its theoretical complexity, including the inherent uncertainty and nonlinearity wrought by social interdependence, we argue that complex autonomous systems must consider multi-agent interactions in order to develop manageable, effective and efficient individual agents and hybrid teams. Important examples include cases of supervised autonomy, where a human oversees several interdependent autonomous systems; where an autonomous agent is working with a team of humans, such as the cyber defense of a network; or where the agent is intended to replace effective, but traditionally worker-intensive team tasks, such as warehousing and shipping. Autonomous agents that seek to fill these roles, but do not consider the interplay between the participating entities, will likely disappoint.

This symposium offered opportunities with AI to address these and other fundamental issues about autonomy and cyber threats, including applications to hybrids at the individual, group, and system levels.

## 1.2 Background of the 2016 Symposium

Our symposium at Stanford in 2015, titled “AI and the mitigation of human error: Anomalies, team metrics and thermodynamics”, was organized by the same four individuals as for the 2015 symposium.

AI has the potential to mitigate human error by reducing car accidents, airplane accidents, and other mistakes made mindfully or inadvertently by individual humans



or by teams. One worry about this bright future is that jobs may be lost. Another is from the perceived and actual loss of human control. For example, despite the loss of all aboard several commercial airliners in recent years, commercial airline pilots reject being replaced by AI (e.g., Markoff 2015).

An even greater, existential threat posed by AI is to the existence of humanity, raised by physicist Stephen Hawking, entrepreneur Elon Musk and computer billionaire Bill Gates. While recognizing what these leaders have said, Etzioni (2016), CEO of the Allen Institute for Artificial Intelligence and Professor of Computer Science at the University of Washington, provides his disagreement along with supporting comments made by others.

Across a wide range of occupations and industries, human error and human performance is a primary cause of accidents (Hollnagel 2009, p. 137). In general aviation, the FAA attributed accidents primarily to skill-based errors and poor decisions (e.g., Wiegmann et al. 2005, Fig. 3, p. 7; Table 1, p. 11).

Exacerbating the sources of human error, safety is one area an organization often skimps to save money. The diminution of safety coupled with human error led to the explosion in 2010 that destroyed the Deepwater Horizon in the Gulf of Mexico (USDC 2012, p. 21). Human error emerges as a top safety risk in the management of civilian air traffic control (Moon et al. 2011). Human error was the cause attributed to the recent sinking of Taiwan's Ocean Researcher V in the fall of 2014 (Showstack 2014). Human behavior is a leading cause of cyber breaches (Howarth 2014).

Humans cause accidents by lacking situational awareness, by a convergence to incomplete beliefs, or by emotional decision-making (for example, the Iranian Airbus flight erroneously downed by the USS Vincennes in 1988; in Fisher 2013). Other factors contributing to human error include poor problem diagnoses; poor planning, communication and execution; and poor organizational functioning.

In this symposium, the participants explored the humans' roles in the cause of accidents and the use of AI in mitigating human error; in reducing problems with teams, like suicide (for example, the German copilot, Libutz, who killed 150 aboard his Germanwings commercial aircraft; in Levs et al. 2015); and in reducing mistakes by military commanders (for example, the 2001 sinking of the Japanese tour boat by the USS Greeneville; in NTSB 2001).

This symposium provided a rigorous view of AI and its possible application to mitigate human error, to find anomalies in human operations, and to discover, when, for example, teams have gone awry, whether and how AI should intercede in the affairs of humans.

### 1.3 Contributed Chapters

Chapter 2, 'Reexamining Computational Support for Intelligence Analysis: A Functional Design for a Future Capability', explores the technological bases for exploiting argumentation-based methods coupled with information fusion techniques for improved intelligence analysis. It was written by James Llinas and Galina

Rogova at the Center for Multisource Information Fusion (CMIF), State University of New York at Buffalo, NY; and Kevin Barry, Rachel Hingst, Peter Gerken and Alicia Ruvinsky at the Lockheed Advanced Technology Laboratories (ATL) in Cherry Hill, NJ. Llinas and colleagues reviewed various Analysis Tool Suites (ATSS) framed by several examples of modern intelligence analyses. These tool-suites address entities in different environments of interest. But these tools do not support computational inter-entity associations for attribute/relation fusion. Most tools, if not all, are single-sourced for entity streams, with tools automating link analyses between bounded entity-pairs and “data fusion” with limited rigor. Most tools assume correct results for pre-processed extractions from entities. But while these tools serve to identify and visualize intuitive associations among entities, they seldom address uncertainty. Their primary function is to discover relational links among entities (like single-hop or limited-hop associations), achieved with limits on uncertainty and inter-entity associability, leaving the complex relations to be deciphered by human analysts. These deficiencies result in considerable cognitive overload on human analysts who need to mentally and largely manually assemble the desired situational interpretations in a narrative form. With the goal of providing a much more automated approach to complex hypothesis integration, the chapter reviews extensively the works in the domain of computational support for argumentation and, as one important element of an integrated functional design, nominates a unique, belief- and story-based hybrid argumentation subsystem design as one part of a combined approach. To deal with the largely textual data foundation of these intelligence analysis tasks, the chapter describes how a previously, author-developed, ‘hard plus soft’ information fusion system (that combines sensor/hard and textual/soft information) could be integrated into this overall design. The functional design described in the chapter combines these two unique capabilities into a scheme that arguably would overcome many of the deficiencies cited in the chapter to provide considerable improvement in efficiency and effectiveness for intelligence analyses.

One theme of this book is to use AI to minimize the errors made by humans. An extension could be the recovery from human errors, considered in Chap. 3. In this chapter, ‘Task Allocation Using Parallelized Clustering and Auctioning Algorithms for Heterogeneous Robotic Swarms Operating on a Cloud Network’, was addressed by Jonathan Lwowski, Patrick Benavidez, John J. Prevost and Mo Jamshidi at the Autonomous Control Engineering (ACE) Laboratory, University of Texas, San Antonio, TX. The authors present a new, centralized approach to the control of robot swarms devised to control a swarm of heterogeneous unmanned vehicles across land, in the water and in the air. The vehicles controlled by the authors’ research team consisted of autonomous surface vehicles and micro-aerial vehicles equipped with cameras and Global Positioning Systems (GPS). This equipment allowed the swarm to operate outdoors. By manipulations with the control program, the swarm was able to demonstrate that the individual robots could be controlled to complete a group task cooperatively and efficiently. The authors first demonstrated how air-based robots could construct a digital map of the local environment with key features (e.g., the locations of targets, such as the survivors of a shipwreck).

The map was uploaded into the cloud on a remote network where clustering algorithms were performed on the map to calculate optimal clusters of designated targets geographically. Afterwards, while still in the cloud, an auctioning algorithm based on factors like relative position on the map and robot capacities led to the assignment of the clusters to surface-based robots (viz., where survivors might be located and matched to recovery vehicles). Next, simulated surface robots traveled to their assigned clusters to complete the tasks allocated to them. Finally, the authors presented the results of their simulations for the cooperative swarm of robots with both software and hardware, demonstrating the effectiveness of their proposed algorithm to control a swarm of robots that might one day be used for the recovery of humans from a shipwreck.

Chapter 4, ‘Human Information Interaction, Artificial Intelligence, and Errors’, was written by Stephen Russell, Chief, Battlefield Information Processing Branch, US Army Research Laboratory, Adelphi, MD; Ira S. Moskowitz, Mathematician, Information Management and Decision Architectures Branch, Information Technology Division, US Naval Research Laboratory, Washington, DC; and Adrienne Raglin, Electrical Engineer, also in the Battlefield Information Processing Branch, US Army Research Laboratory, Adelphi, MD. In a time of pervasive and increasingly transparent computing devices, for humans, the importance of interaction with information itself will become more significant than the devices that provide information services and functionality today. From the perspective of the authors, Artificial Intelligence (AI) is a proxy for humans’ information interactions, not only providing assistance in the interactions themselves, but also providing guidance and automation in the applications of that information. Given this perspective, new opportunities for AI technologies will arise. But because of mismatches in human intent and goals and proxy-AI functionality, variability in information interaction will create opportunities for error. The trend towards AI-augmented human information interaction (HII) will cause an increased emphasis on cognitive-oriented information science research plus many new ways of thinking about errors, the way we humans manage errors and how we address the consequences of errors. With this focus on errors, the authors review the intersection of HII and AI in this chapter.

In Chap. 5, Signe A. Redfield, Engineer, Evaluations of Autonomous Systems, U.S. Naval Research Laboratory; and Mae Seto, Defense Research and Development Canada (DRDC), Atlantic Research, Dalhousie University, write about ‘Verification Challenges for Autonomous Systems’. The authors have associated autonomy with robots coupled to sensory systems to move in physical reality and disassociated from the artificial intelligence (AI) they have associated with abstract problem solving; in their view, robots may use AI as a tool to navigate the physical environment, while AI may use a robot to implement a solution to a problem that AI has solved. The authors have identified a number of open research challenges in the area of verification of autonomous systems. They outline the existing tools available to identify associated gaps, to identify the challenges for the additional tools that do not currently exist but are needed, and to suggest new directions in which progress may probably be made. In their review, they note that existing research programs attempt to address these problems, but there are many more unexplored research challenges

than there are research programs underway to explore them, highlighting that this field of research is not yet mature. In this chapter, the authors attempt to enumerate the unexplored regions facing both autonomous robots and AI but also how to exploit the advantages already advanced by both systems.

Chapter 6, ‘Conceptualizing Overtrust in Robots: Why Do People Trust a Robot That Previously Failed?’, was written by Paul Robinette, Robotics Research Scientist, Massachusetts Institutes of Technology; Ayanna Howard, Electrical and Computer Engineering, Georgia Tech Research Institute (GTRI); and Alan R. Wagner, Director, Robot Ethics and Aerial Vehicles Laboratory (REAL), Aerospace Engineering, Penn State University. Most of the research by these authors had found that people tended to trust robots despite the errors made by robots; these findings have led the authors to develop a new research area on the repair of trust, especially under the conditions for which trust might be resuscitated. This chapter directly addresses how autonomous agents can help humans reduce or mitigate errors to increase trust in robotics, a problem when the robots are unreliable. The authors begin by presenting their research to suggest that humans tend to overly trust and forgive robots as guides during emergency situations, the focus of much of their research. Their experiments have shown that, at best, human participants in simulated emergencies focus on guidance provided by robots regardless of a robot’s prior performance or other guidance information, and, at worst, participants come to believe that the robot is more capable than other sources of information. Even when a robot-guide harms trust, a properly timed statement can convince a human participant to once again be assisted by a robot guide. Based on this evidence, they have conceptualized the overtrust of robots by using their previous framework of situational trust. They define two mechanisms where humans engage in overtrusting robots: misjudging the abilities or intentions of the robot and misjudging the risk they face in a scenario. The authors discuss their prior work in light of this reconceptualization in their attempt to explain their previous results, to encourage future work, and to help the public, robot designers and policy leaders to be guided appropriately by robots.

Chapter 7, ‘Research Considerations and Tools for Evaluating Human-Automation Interaction with Future Unmanned Systems’, was written by Ciara Sibley and Joseph Coyne, Engineering Research Psychologists in the Warfighter Human System Integration Laboratory Section, Information Management and Decision Architectures Branch, Information Technology Division, Naval Research Laboratory, Washington, DC; and by Sarah Sherwood, PhD candidate, Embry-Riddle Aeronautical University, Daytona Beach, FL. From their research perspective, the authors consider the approaches, methods and tools used to evaluate the interactions of humans and automation in present and future unmanned systems, specifically for unmanned aerial vehicles (UAVs). The authors discuss the levels of automation required to meet the objectives set by the Department of Defense to increase autonomy in robot-machine systems to allow a single human operator to supervise multiple UAVs, an inversion of what occurs today. This new paradigm, for which a lot of research across DoD is directed, requires significant improvements in automation reliability and capability based on a more fundamental understanding of

how human performance is and will be impacted when interacting with present-day and planned future systems. Research into interacting with automated systems has often focused on trust, reliability and automation levels. However, if the goal of automating systems is to minimize the need for the human oversight of robot-machine interactions with future systems, unfortunately, the majority of current research addressing greater autonomy for UAVs falls short of achieving DoD's objective for less human management. The authors discuss these limits and other challenges to assessing human interaction with automation by using traditional measures like speed and accuracy, but they also include in their review other measures of operator state such as workload and fatigue, situation awareness probes, and eye tracking. The authors close by discussing their new supervisory control testbed, which is integrated with multiple psychological sensors, and designed to assess human automation interaction across a broad range of mission contexts while also meeting DoD objectives.

In recent years, fueled by multi-media images of robots causing a threat to the human race, research on the autonomy of robots has raised significant concerns in society. These concerns are addressed in Chap. 8 'Robots Autonomy: Some Technical Issues' by Catherine Tessier, Senior Aeronautical Research Engineer and Subject Matter Expert (in Advanced Autonomy Robot Systems), ONERA (Le Centre Français de Recherche Aérospatiale; i.e., The French Aerospace Laboratory), Toulouse, France. From her perspective, society has become overly concerned about a future dominated by robots that would exhibit human-like features or intentions. Moreover, as Tessier points out, the concerns expressed by society hide the present "technical reality" of advancing far beyond the robots of today. In her chapter, she defines all of the terms in her review while discussing the present technical reality of robot autonomy along with common examples that make her chapter accessible to technical scientists and the lay public alike while at the same time delving into the challenge of assigning responsibility for the day when the control of machines is to be shared between robots and human operators (or another machine, robot or human). To allay the public's concern about the complexity of future robot systems, she also addresses the current real issue of ethics and morality for robots, namely, how robots should be designed to behave in specific situations where decisions involve conflicting moral values. At the end of her chapter, she addresses these issues directly with her review of the limits of human responsibility and by raising the question of whether and under what circumstances robots should be able to take control from humans. The key point of her chapter is that it focuses on objective, technically grounded considerations for robot autonomy and authority sharing between humans and robots.

Chapter 9, 'How Children with Autism and Machines Learn to Interact', was written by Boris A. Galitsky, Founder and Chief Scientist, Knowledge-Trail, Inc., San Jose, CA; and Anna Parnis, Department of Biology, Technion-Israel Institute of Technology, Haifa, Israel. These two authors explore how children with autism (CwA) interact with each other, teachers and others in society, and what kinds of difficulties they experience in the course of these interactions. They take a team approach. In their view, autistic reasoning is a means to explore team formation

and human reasoning in general because it is simple in comparison to the more sophisticated reasoning of controls and software systems usually considered by AI on the one hand; but on the other hand, their model allows them to explore human behavior in real-world environments that they believe may be generalizable to humans, machines and robots. From their perspective, they have discovered that reasoning about the mental world, impaired in various degrees in autistic patients, is the key limiting parameter for forming teams and cooperating among team members once teams have been formed. While teams of humans, robots and software agents have manifold other limitations when they attempt to form teams, including resources, conflicting desires, uncertainty and environmental constraints, based on their research, CwA have only a single limitation, expressed as reduced reasoning about their mental world. The authors correlate the complexity of the expressions for mental states that all children are capable of achieving with their ability to form teams. In the process, they describe a method to rehabilitate reasoning for CwA children, and they address its implications for the behavior of all children in a social world that entails interactions and cooperation in the formation of teams.

Chapter 10, 'Semantic Vector Spaces for Broadening Consideration of Consequences', was written by Douglas Summers Stay, Artificial Intelligence Researcher, Army Research Laboratory, Adelphi, MD. The author reviews three approaches for reasoning systems. He first proposes that reasoning systems combined with human intent on simple models of the world are unable to consider the potential negative side effects of their actions sufficiently well-enough to modify plans to avoid these adverse effects (e.g., reducing the potential for human error). After many years of research and effort dedicated to encoding the enormous and subtle body of social facts with the aim of converting common sense into a knowledge base, this approach has proved too difficult. As a second alternative, some scientists have encoded concepts and the relations between them in geometric structures, namely, distributed semantic vector spaces derived from large text corpora, to construct representations that capture subtle differences in the meaning of common-sense concepts while at the same time being able to perform analogical and associational reasoning that, unfortunately, limit knowledge bases. Encumbered by source materials, the second alternative is unreliable, poorly understood, and biased in the view it affords of the world. As a third alternative to both of the first two approaches, the author combines these two approaches to retain the best properties of each to lead to a richer understanding of the world and human intentions.

Chapter 11, 'On the Road to Autonomy: Evaluating and Optimizing Hybrid Team Dynamics', was written by Chris Berka, CEO and Co-Founder, Advanced Brain Monitoring (ABM); and Maja Stikic, Computer Scientist, ABM, both located in Carlsbad, CA. The authors explored the potential of traditional psychometric approaches for the study of teams (viz., engagement, workload, stress) supplemented with neuroscience methods for the measurement of teams in order to determine the dynamics of teams in real time (e.g., with electroencephalographic, or EEG, measurements). Their method adopted techniques that were designed to be inconspicuous to participants and observers so that the authors would be able to quantify the actual cognitive and emotional states of a team moment by moment.

The neuroscience approach allowed the authors to construct a new platform for the teams that they then used as a tool to study teams. With this new platform, the authors reviewed a number of the studies that they had conducted to provide a wide range of conditions and measurements for teams with this emerging technology (e.g., monitoring team neurodynamics; the emergence of team leadership; neural responses associated with storytelling narratives; neural processes associated with tutoring; and the training of surgical skills). The authors also discussed the implications of using this new technology in the study of teams and they closed their chapter with a review of the potential research that they are considering for the future.

Chapter 12, ‘Cyber-security and Optimization in Smart “Autonomous” Buildings’, was written by Michael Mylrea, Manager, Cybersecurity and Energy Technology, Pacific Northwest National Laboratory; also, a PhD candidate in the Executive Cybersecurity Doctoral Program at George Washington University; co-authored with Sri Nikhil Gupta Gourisetti, Research Engineer, Electricity Infrastructure, Pacific Northwest National Laboratory; a PhD candidate in Engineering Sciences and Systems Doctoral Program at the University of Arkansas at Little Rock. The authors note that significant resources have been invested in making buildings “smart” or more autonomous by digitizing, networking and automating key systems and operations. Smart autonomous buildings create new energy efficiency, economic and environmental opportunities. But as buildings become increasingly networked to the Internet, they can also become more vulnerable to various cyber threats. Automated, autonomous and Internet-connected buildings systems, equipment, controls, and sensors can significantly increase cyber and physical vulnerabilities that threaten the confidentiality, integrity, and availability of critical systems in organizations. Securing smart autonomous buildings presents a national security and economic challenge to the nation. Ignoring this challenge threatens business continuity and the availability of critical infrastructures that are enabled by smart buildings. In this chapter, the authors address these challenges and explore new opportunities in securing smart buildings that, on the path to autonomy, are enhanced by machine learning, cognitive sensing, artificial intelligence (AI) and smart-energy technologies. First they identify cyber-threats and challenges to smart autonomous buildings. Second they make recommendations on how AI enabled solutions can help smart buildings and facilities better protect, detect and respond to cyber-physical threats and vulnerabilities. Third they provide various case studies that examine how combining AI with innovative smart-energy technologies can increase both cybersecurity and energy efficiency savings in buildings. The authors conclude with their ideas for the future to continue to develop more resilience technology to counter threats to autonomous buildings and facilities.

The last of the contributed chapters, Chap. 13, ‘Evaluations: Autonomy and Artificial Intelligence: A Threat or Savior?’, was written by W.F. Lawless, Departments of Mathematics and Psychology, Paine College, Augusta, GA; and Donald A. Sofge, Computer Scientist, Distributed Autonomous Systems Group, Navy Center for Applied Research in Artificial Intelligence, Naval Research Laboratory, Washington, DC. In this chapter, the authors first review and evaluate their own research presented at AAI-2015 (computational autonomy) and AAI-2016 (reducing human errors). Then they evaluate each of the other contributed

chapters on their own terms that more or less mesh with these two parts of this book. The authors begin by discussing the remarkable recent successes with Artificial Intelligence (AI); e.g., machine learning. Yet, they note, these successes have been followed by extraordinary claims that autonomous robots in society may one day threaten human existence. They temper these claims by discussing how often many predictions about the future have missed the mark, including the 2016 Presidential election. Then the authors approach the field of AI with a theoretical perspective, beginning with how little is accepted about human-human interactions, an impediment to the advance of autonomous robot teams. At the heart of the failure by human experts to predict important social outcomes, like elections, is the phenomenon of interdependence (mutual information), the social aspect of the interaction that makes humans human, and the means by which human aggregation occurs in the social, political and cultural world (e.g., teams, political parties, juries). Little is accepted about what interdependence means and how to model it. But the authors believe that without a theoretical understanding and computational mastery of interdependence, while AI systems may be able to beat humans in games, AI systems will never be as innovative nor as capable of solving difficult problems as are humans, nor will humans have the confidence that AI may be able to help humans successfully reduce human errors. With the phenomenon of interdependence modeled mathematically, the authors evaluate the two themes and the chapters in this book to provide readers a path forward for further research with AI. In the first part of their chapter, the authors discuss the use of AI in the development of autonomy for individual machines, robots and humans in states of social interdependence, followed by its evaluation and then evaluations of the chapters with similar themes; in the second part, the authors discuss the use of AI and machines in reducing, preventing or mitigating human error in society, also followed by its evaluation and an evaluation of the remaining chapters.

## References

- Etzioni, O. (2016, 9/20), “No, the Experts Don’t Think Superintelligent AI is a Threat to Humanity. Ask the people who should really know”, MIT Technology Review, from <https://www.technologyreview.com/s/602410/no-the-experts-dont-think-superintelligent-ai-is-a-threat-to-humanity/>
- Fisher, M. (2013, 10/16), “The forgotten story of Iran Air Flight 655, Washington Post, from [https://www.washingtonpost.com/news/worldviews/wp/2013/10/16/the-forgotten-story-of-iran-air-flight-655/?utm\\_term=.cdf037448c6e](https://www.washingtonpost.com/news/worldviews/wp/2013/10/16/the-forgotten-story-of-iran-air-flight-655/?utm_term=.cdf037448c6e)
- Hollnagel, E. (2009), *The ETTO Principle: Efficiency-Thoroughness Trade-Off: Why Things That Go Right Sometimes Go Wrong*. Boca Raton, FL: CRC Press.
- Howarth, F. (2014, 9/2), “The Role of Human Error in Successful Security Attacks”, Security Intelligence, from <https://securityintelligence.com/the-role-of-human-error-in-successful-security-attacks/>
- Lawless, W.F. (published online December 2016; forthcoming), The entangled nature of interdependence. Bistability, irreproducibility and uncertainty, *Journal of Mathematical Psychology*, <http://dx.doi.org/10.1016/j.jmp.2016.11.001>
- Levs, J., Smith-Spark, L. and Yan, H. (2015, 3/26), “Germanwings Flight 9525 co-pilot deliberately crashed plane, officials say”, CNN, from <http://www.cnn.com/2015/03/26/europe/france-germanwings-plane-crash-main/>



- Markoff, J. (2015, 4/6), “Planes Without Pilots”, *New York Times*, from <https://www.nytimes.com/2015/04/07/science/planes-without-pilots.html>
- Moon, W., Yoo, K. and Y. Choi 2011, Air Traffic Volume and Air Traffic Control Human Errors, *Journal of Transportation Technologies*, 1(3): 47-53, doi: [10.4236/jtts.2011.13007](https://doi.org/10.4236/jtts.2011.13007).
- NTSB (2001, May), Marine Accident Brief DCA-01-MM-022 Collision 2/9/2001, National Transportation Safety Board, Washington, D.C. 20594, NTSB/MAB-05/01 from <https://www.ntsb.gov/investigations/AccidentReports/Reports/MAB0501.pdf>
- Showstack, R. (2014, 10/21), “Taiwan Shipwreck Is Major Loss for Ocean Research, Scientists Say. The 10 October shipwreck of Taiwan’s R/V Ocean Researcher V, which resulted in two deaths, is a major setback for ocean research in Taiwan, according to scientists”, EOS, from <https://eos.org/articles/taiwan-shipwreck-major-loss-ocean-research-scientists-say>
- Tetlock, P.E. and Gardner, D. (2015), *Superforecasting: The Art and Science of Prediction*, Crown.
- USDC (2012, 2/22) United States District Court Eastern District of Louisiana, MDL No. 2179 “Deepwater Horizon” in the Gulf. In re: Oil Spill by the Oil Rig of Mexico, on April 20, 2010. Applies to: 10-4536. Case 2:10-md-02179, Document 5809.
- Wiegmann, D. Faaborg, T., Boquet, A., Detwiler, C., Holcomb, K. and Shappell, S. (2005), “Human Error and General Aviation Accidents: A Comprehensive, Fine-Grained Analysis Using HFACS”, Office of Aerospace Medicine, DOT/FAA/AM-05/24, Washington, DC 20591, from [https://www.faa.gov/data\\_research/research/med\\_humanfacs/oamtechreports/2000s/media/0524.pdf](https://www.faa.gov/data_research/research/med_humanfacs/oamtechreports/2000s/media/0524.pdf)

# Chapter 2

## Reexamining Computational Support for Intelligence Analysis: A Functional Design for a Future Capability

James Llinas, Galina Rogova, Kevin Barry, Rachel Hingst, Peter Gerken, and Alicia Ruvinsky

### 2.1 Motivation

Analysis Tool Suites (ATS's) such as Analyst's Notebook<sup>1</sup> Analyst's Workspace (Andrews and North 2012), Sentinel Visualizer,<sup>2</sup> and Palantir Government,<sup>3</sup> Entity Workspace (Bier et al. 2006), and Jigsaw (Stasko et al. 2013), among others are examples of modern intelligence analysis frameworks. A major point for sensibly all these tool-suites is that they start by focusing on the entity level within the environments of interest. None overtly discuss computational support to inter-entity association and attribute/relation fusion. That is, most if not all are single-source-based as regards entity streams, with the tools doing varying degrees of automated link analysis among bounded entity-pairs toward realization of "data fusion" albeit with rather limited rigor. Further, most also assume that any preprocessing that provides entity extraction yields correct results. This framework of tool products provides the basis for identifying and visualizing relational connections between entities, but these connections are largely if not exclusively performed in the mind of the analyst. In most cases, nothing is done in the way of computational support to dealing with entity or relational uncertainties. The primary function of most of these ATS's is relational link discovery to discern inter-entity relations of bounded extent (in graph science terminology, usually single-hop or limited-hop relations),

---

<sup>1</sup><http://www03.ibm.com/software/products/en/analysts-notebook>.

<sup>2</sup><http://www.fmsasg.com/>.

<sup>3</sup><http://www.palantir.com/>.

J. Llinas • G. Rogova

Center for Multisource Information Fusion (CMIF), State University of New York at Buffalo, Buffalo, NY, USA

e-mail: [llinas@buffalo.edu](mailto:llinas@buffalo.edu); [rogovagl@gmail.com](mailto:rogovagl@gmail.com)

K. Barry (✉) • R. Hingst • P. Gerken • A. Ruvinsky

Lockheed Advanced Technology Laboratories (ATL), Cherry Hill, NJ, USA

e-mail: [kevin.barry@lmco.com](mailto:kevin.barry@lmco.com); [rachel.hingst@lmco.com](mailto:rachel.hingst@lmco.com); [alicia.ruvinsky@lmco.com](mailto:alicia.ruvinsky@lmco.com)

achieved with quite limited analytical formality regarding issues of uncertainty, inter-data and/or inter-entity associability, and of relational complexities. Thus, deeper and broader analysis of entity and relational connectedness is left for the human analyst. This is especially true in regard to the assembly of typical final desired analysis products in the form of stories or narratives; said otherwise, there is very limited technical support for synthesis or fusion of hypotheses into the larger context of situational understanding. By and large, these tools try to support the Sensemaking (SM) or schema-development loop of SM (Pirolli and Card 2005; Klein et al. 2006), but either have no algorithmic or technological-process support or provide quite-limited automated support to these higher goals; these assessments are summarized in the review paper of Llinas (2014a, b).

Thus we perceive a need first for a processing/reasoning paradigm that can provide the framework for a more holistic, systemic based approach to intelligence analysis. As sensibly all critiques about intelligence analysis as well as the analysis requirements stated in field manuals describe that the main product that an analyst is driving toward is a narrative type description of some world condition/situation, we set this goal for our research presented in this chapter as well. So, primarily we are seeking to study ways that discrete, single-theme hypotheses can be synthesized or fused into a more holistic and semantic construct in the form of a story or narrative. Our approach incorporates methods of associating and fusing so-called hard (sensor) and soft (textual, semantic) information, as many intelligence analysis environments have such disparate data streams as input. (We note that virtually all the work in the areas we studied here only involve soft or textual type inputs.) We believe that the functional design produced here provides a basis for a next step involving research prototype development, and because of this we have also studied ways to test and evaluate such a prototype.

## 2.2 Goals and Requirements

In this research program, we sought to explore a number of possible computationally-aided enhancements in the ways that technologies can better support and improve the rigor and efficiency of intelligence analysis through the integration of new computationally-based methods and algorithms but also by exploring and nominating new ways in which improved human-machine symbiosis can be realized. Also, we were trying to strike the best balance between technologies and methods that are of the basic research variety while having plausibility in terms of potential for mid-term type operational deployment. Another main goal was toward providing support that can yield the type of “story” or narrative type product that many intelligence analysis environments require. These are those environments that allow for more contemplative methods, accommodating the formulation and evaluation of optional interpretations that have to be weighed and evaluated or argued for. This goal imputes a requirement for capabilities that support what we are calling “hypothesis synthesis” or “hypothesis fusion” as mentioned previously, where competing

hypotheses that evolve either: (a) directly from evidence or (b) developed from evidence or assumptions by disparate individual tools are traded off and synthesized into a defensible, integrated hypothesis at the narrative or situational level. In today's analysis environments, these synthesizing operations constitute and demand a high cognitive workload. A major goal is to develop a design whose overall rationale is traceable to and consistent with joint service and Intelligence Community future directions in methodological development balancing effectiveness, efficiency, and rigor; as a result, we have made efforts to garner real-world viewpoints on these directions.

## 2.3 Future Directions in Intelligence Analysis

### 2.3.1 *Reviews of Open Literature and Operational Environments*

The research described here was in fact partially inspired by our prior exploration of the nature of modern-day computational support for intelligence analysis in the open literature as summarized in Llinas (2014a, b). That work extensively examined much of the literature on such techniques with a focus on technology strategies and interfacing strategies in regard to methods to achieve some level of symbiosis. It should be noted that this survey also collected works from the field of criminal analysis and the related area of Artificial Intelligence and the Law. Our research team at the Center for Multisource Information Fusion has also addressed these topics under a large Army Research Office grant for Unified Research on Network-based Hard and Soft Information Fusion, see e.g., Llinas et al. (2010) and Date et al. (2013a, b) for the Counterinsurgency domain. In both of these surveys, what we primarily saw was a strategy for analytical tool suite design that resulted in collages of disparate tools of various descriptions. Each of these tools can be argued to be individually helpful, producing what we called "situational fragments", i.e. hypotheses, each of which are hypotheses about a particular slice of a situational condition. These problems, and the employment of modern technologies that allow evermore data and information to be available, are extraordinarily complex and it is natural to see "divide and conquer" solution, tool, and visualization strategies being applied. But the latent challenge for sensibly all human analysts involved in these situations is to connect the dots, evolve the most plausible story/narrative, or the most plausible argument in the face of inherent complexity and "big data" quantities and varieties of information. For that type of capability, we saw nothing at all in this survey, leading to our conclusion that there is a significant need for development of both a paradigm and associated technological support for hypothesis synthesis or fusion, aiding human analysts to assemble a more holistic picture (a narrative or story) much more efficiently.

In the Fall of 2015, a team visit to the Air Force National Air and Space Intelligence Center (NASIC) was carried out in order to assess our evolving perspectives regarding future analysis requirements. Because of our future-oriented perspective, our visit focused on the Advanced Analytics Cell (AAC) team, that similarly is studying such future requirements. In summary, this visit revealed that there was considerable commonality in the respective lines of thought across the activities of the AAC and our approach. It also broadly provided a level of confidence that the approach described here was sound and that it resonated with current advanced thinking at least in the Air Force as regards methods and needs of modern intelligence analysis.

### 2.3.2 Analytical Rigor in Intelligence Analysis/Argument Mapping

Another touchstone for the project as regards vetting our thinking and approach involved discussions with staff from the Army Intelligence Center at Ft. Huachuca, NM. Messrs Robert Sensenig and William Hedges (of Chenega Corp., advisors to the Army on intelligence matters) were our key points of contact. Two main topics were discussed: rigor in analysis, and the use of argument-based techniques of analysis. The Army is quite keen on the entire issue of improving rigor in analysis; this viewpoint certainly is consistent with our own thoughts regarding improvements in the intellectual aspects of analysis. Mr. Sensenig provided the charts of Figs. 2.1 and 2.2 below that depict the mapping/cross-correlation of analysis functions and levels

Low Rigor	Moderate Rigor	High Rigor
<p><b>Hypothesis Exploration</b></p> <p><i>"I have one hypothesis I like."</i></p> <ul style="list-style-type: none"> <li>• No consideration of alternatives.</li> <li>• Argues how data that does not fit or is new can fit favorite hypoth.</li> </ul>	<p><i>"I feel comfortable that one explanation accounts for majority data."</i></p> <ul style="list-style-type: none"> <li>• Unbalanced focus on ML COA.</li> <li>• Acknowledges other COA possible.</li> <li>• Considers risks of alternative COAs.</li> </ul>	<p><i>"I am confident of the best explanation and have seriously considered other possibilities."</i></p> <ul style="list-style-type: none"> <li>• Interactive debate from multiple perspectives on alternatives.</li> <li>• Actively considers and tracks data that does not fit ML or MD.</li> </ul>
<p><b>Information Search</b></p> <p><i>"I found something reasonably Comprehensive and believable."</i></p> <ul style="list-style-type: none"> <li>• Did not go beyond routine sources</li> <li>• Did not select multiple sources</li> <li>• Relied on second and third-hand sources, no direct comms with primary sources.</li> </ul>	<p><i>"I am seeing repeating patterns, and they all seem to agree or there seems to be two primary possibilities."</i></p> <ul style="list-style-type: none"> <li>• Actively seeks info that is not easily retrieved or collected.</li> <li>• Multiple data types and proximal sources considered for key findings</li> <li>• Read beyond specific tasking</li> </ul>	<p><i>"I am not learning anything new. I reached theoretical saturation."</i></p> <ul style="list-style-type: none"> <li>• Support from others to broaden sampled space.</li> <li>• Multiple data types and proximal sources considered for all inferences</li> <li>• More knowledgeable about subject area than most document authors.</li> </ul>
<p><b>Information Validation</b></p> <p><i>"I found one that sounds good"</i></p> <ul style="list-style-type: none"> <li>• Copies report with little re-interpretation, correlation</li> <li>• Does not display healthy skepticism.</li> <li>• No tracking of process, no knowledge of data pedigree</li> </ul>	<p><i>"I verified my key arguments and predictions are based on the most trustworthy source I have"</i></p> <ul style="list-style-type: none"> <li>• Attempts to verify arguments from multiple independent sources</li> <li>• Aware of how analysis could be wrong based on experience or feedback</li> <li>• Aware of corrupted data sources</li> </ul>	<p><i>"I feel confident that I validated, by reasonable means, the facts used to support key arguments."</i></p> <ul style="list-style-type: none"> <li>• Systematic, semi-formal processes employed to verify information</li> <li>• Clear distinction between facts, assumptions, inferences</li> <li>• Fully investigated "sourcing"</li> </ul>

Fig. 2.1 Mapping of analysis functions vs levels of rigor (Part 1) (Courtesy of Mr. Robert Sensenig, Chenega Corp.)

Low Rigor	Moderate Rigor	High Rigor
<p><b>Inference Resilience</b></p> <p><i>"My story/explanation/argument seems reasonable to me, independent of available supporting evidence."</i></p>	<p><i>"I feel that the evidence is reasonably solid for my primary explanation."</i></p> <ul style="list-style-type: none"> <li>• Considers whether being wrong about some inferences would influence or negate the best explanation.</li> <li>• Beware false precision!!</li> </ul>	<p><i>"I feel comfortable that the key inferences are resilient to inaccurate information."</i></p> <ul style="list-style-type: none"> <li>• Uses strategy to systematically consider strength of evidence if individual interpretations debunked.</li> <li>• Actively looked for reasons why a source might misinterpret or manipulate data/information.</li> </ul>
<p><b>SME Collaboration</b></p> <p><i>"I trust my supervisor to cover specialist content area or to be the SME."</i></p>	<p><i>"I have talked to SMEs, as time allowed, within my personal network."</i></p> <ul style="list-style-type: none"> <li>• Attempts to consult some of the right people.</li> </ul>	<p><i>"Leading expert in the key content area." (Beware Group Think!!)</i></p> <ul style="list-style-type: none"> <li>• Capital expended to gain access to leading experts in multiple fields related to the analysis.</li> </ul>
<p><b>Information Synthesis</b></p> <p><i>"I compiled the relevant info."</i></p> <ul style="list-style-type: none"> <li>• Numerical values or graphs disconnected from key arguments.</li> </ul>	<p><i>"I provided insight that goes beyond the source reporting &amp; key documents."</i></p> <ul style="list-style-type: none"> <li>• Validation of events in context.</li> <li>• Understanding depicted as an integrated view including tradeoff dimensions. (Frameworks, models).</li> </ul>	<p><i>"I considered diverse interpretations trying to identify new concepts"</i></p> <ul style="list-style-type: none"> <li>• Sensemaking metrics are high.</li> <li>• Collaborative cross checks applied to data synthesis processes</li> <li>• Collaborative use of diagrams to show relationships between evidence and hypothesis.</li> </ul>

Fig. 2.2 Mapping of analysis functions vs levels of rigor (Part 2) (Courtesy of Mr. Robert Sensenig, Chenega Corp.)

of rigor, notionally showing an analyst’s mind-set across these functions and levels, as well as thumbnails of analysis activities across the matrices. These charts are among the resources we used to direct our efforts.

Mr. Hedges recounted his experience in learning of argument-based methods of analysis and also shared segments of the Army’s training activities in the teaching of argument mapping for intelligence analysts. Figure 2.3 shows an excerpt of one of the training segments directed to teaching of argument mapping.

Overall, we believe it is quite clear that the thinking and approaches of this research program are very consistent with modern thoughts in both the Air Force and Army in regard to:

- The use of improved intellectual strategies and methods
- The need for an movements to improve analytical rigor
- The employment of argumentation-based methods and technologies as one framework to achieve these goals

## 2.4 Approaches to Computational Support

### 2.4.1 Paradigms and Methods

In today’s open-world environment, historical paradigms and methods that rely on deep analysis of an adversary’s Tactics, Techniques, and Procedures (TTP’s) as a basis for paradigms that can be broadly labeled as of a template-matching type are

**U.S. ARMY INTELLIGENCE CENTER AND FORT HUACHUCA  
Fort Huachuca, Arizona 85613-7002**

**LP Narrative & Teaching Plan: Argument Mapping  
24 April 2013  
PFN:xxxxxxx**

**Enabling Learning                      SLIDE 2: Objective**

<b>ACTION:</b>	Create an Argument Map to make analytic assumptions, intelligence gaps, or arguments more transparent.
<b>CONDITIONS:</b>	Given all class handouts to date, appropriate references, an operational framework scenario, and in-class discussion.
<b>STANDARDS:</b>	Create an argument map that incorporates critical and creative thinking and basic and diagnostic structured analytic techniques in order to provide clearer ACH understanding and validate the ACH.

**Fig. 2.3** Sample of curriculum at Army Intelligence School Training in argument mapping (Courtesy of Mr. William Hedges of Chenega Corp.)

considered unworkable. Modern-day adversaries and problem conditions demand more flexibility and accommodation of imperfections in analysis techniques. These environments, that we call “weak knowledge” problems, require a more flexible approach and one that allows for unknown states of affairs and degrees of ignorance while carrying out the best analysis possible. Such methods are usually labeled as defeasible and abductive<sup>4</sup> and are directed to the most rational hypotheses that can be defended in some way as “best”. In our exploration of alternatives, we narrowed our choices based on two factors: one was the commentaries on intelligence analysis and associated assertions about methodological requirements that balance evidence, arguments, and stories (i.e., nominated hypotheticals), and the other was a body of work we discovered that was centered in Europe that focused on methods of this type, with a deep basis on argumentation-based principles. One clear example of these remarks is shown in the writings of Schum (2005) who suggests that:

---

<sup>4</sup>We like Stanford’s definition here (<http://plato.stanford.edu/entries/reasoning-defeasible/>): “Reasoning is *defeasible* when the corresponding argument is rationally compelling but not deductively valid. The truth of the premises of a good defeasible argument provides support for the conclusion, even though it is possible for the premises to be true and the conclusion false. In other words, the relationship of support between premises and conclusion is a tentative one, potentially defeated by additional information.”

- “**Careful construction of arguments** in defense of the credibility and relevance of evidence **goes hand-in-hand with the construction of defensible and persuasive narratives.**”
- “In constructing a narrative account of a situation of interest we must be able to **anchor our story appropriately on the evidence** we have that is relevant to the conclusion we have reached. Careful argument construction provides the necessary anchors.”

These remarks, and the results of our surveys, suggest an exploration of methods that jointly exploit the union of evidence, arguments, and stories, in a synergistic dynamic that leads to “best” narratives that holistically convey the most rational explanation of the evidences and sub-stories. These source materials were the foundation of the evolution of our thinking to explore a paradigm of this nature.

### 2.4.2 *Argumentation Methods*

As we contend above, one main technological/theoretical theme that we pursue here is the examination of argumentation-based concepts, methods, and computationally-supported tools as one candidate paradigm supportive of intelligence analysis. Argumentation-based methods have a long history in the law and in the teaching of critical thinking, and in the last decade or so have found their way into supporting criminal and intelligence analysis. These extended applications have largely been a result of research and development in the construction of computational tools for “diagramming” or “mapping” arguments that enable and streamline the examination of the veracity of pro and contra arguments in various situations.<sup>5</sup> Before reviewing the state of the art in computational methods for argumentation based reasoning, we briefly review the different paradigms for argumentation itself; that is, there are different flavors or variations of methods that have the core notion of an argument as their foundation. This summary review is shown in Table 2.1 below.

The majority of argumentation-based methods utilize a deterministic formal logic and theorem proof approaches, and the notion of argument acceptance and attack, see, e.g., Simari and Rahwan (2009). There has been multiple argumentation schemes developed with each of them having advantages and drawbacks as methods useful for supporting decisions based on a highly uncertain environment. Most of them represent abstract argumentation, which determines an argument’s acceptability

---

<sup>5</sup>By the way, we see the (necessary) balancing of Pro and Contra arguments as another good feature of these argumentation methods; to some degree this is a built-in preventative to the human foible of confirmation bias.



**Table 2.1** Types of Argumentation-based Paradigms

Argumentation types	Methods	Prototypes <sup>a</sup>
<b>Abstract argumentation</b>	Involving formal logic, theorem proof, and based on the notion of argument acceptance and attack	CISpaces, Carneades Araucaria and various others
<b>Story-based argumentation</b>	Abduction-based reasoning about hypothetical stories explaining the evidence	Bex’s research on design; AVERS
<b>Hybrid methods</b>	Combination of logic and probability or belief	
<b>Assumption based probability/ belief based argumentation.</b> (A probabilistic extension of abstract argumentation.)	Conjunction of uncertain assumptions to define arguments and disjunction of arguments Assigning probabilities/beliefs to assumptions	ABEL
<b>Belief-story based argumentation</b>	Observations are explained by hypothetical stories Uncertain arguments based on evidence are combined to support alternative stories and select the most credible one (abductions)	This was the goal or the research described in this chapter

<sup>a</sup>See later discussion on Prototypes for citations

on the basis of its ability to counterattack all arguments attacking it. A more promising approach introduced in e.g. Bex (2013) is an abstract story-based argumentation, in which hypothetical “causal stories are hypothesized to explain the evidence, after which these stories can be supported and attacked using evidential arguments.” A combination of logic and belief theories for argumentation under uncertainty has been considered for assumption based argumentation, see e.g. Haenni (2001), but these models require a known and complete knowledge base, which does not exist in the context, which we are addressing here. We seek abductive reasoning methods that combine certain desirable capabilities:

- Allowance for open-world reasoning
- Allowance for assigning and combining beliefs in arguments and reliability of the source (i.e., a basis for assigning and combining/propagating uncertainty)
- Integration of human intelligence that enables hypothetical stories to be combined with hypotheses resulting from evidence-based arguments
- Method of evaluating and selecting the most credible stories

Abductive reasoning is often labeled as “backward reasoning” in that it explores/nominates plausible conclusions or assertions that can “explain” or rationalize the evidence available; the notion is that a rearward look is taken from the conclusion toward the available evidence. Abductive reasoning is also often described as reasoning to the best explanation. Our approach is also hybrid in bringing together the abductive reasoning over both the uncertain arguments and human-nominated storylines and rationalizing both lines with the also-uncertain evidence. To deal with

these uncertainties, we propose to incorporate the Transferable Belief Model (TBM) see, e.g. Smets (1994). Briefly the TBM is a two-level model, in which quantified beliefs in hypotheses about an object or state of the environment are represented and combined at the *credal* level<sup>6</sup> while decisions are made based on so-called *pignistic* probabilities obtained from the combined belief by the *pignistic*<sup>7</sup> transformation at the *pignistic level*. So taken together, our approach can be summarized as involving the explicit incorporation of uncertainty into hybrid story-based argumentation, depicted in Fig. 2.4.

The basic ideas of the story-based approach are presented in Fig. 2.5 that shows that:

- Arguments are derived from evidential foundations
- Stories are analyst-nominated (with computational support, e.g., prior case libraries) hypotheticals
- Together these lead to the assembly of sub-stories and, again with computational support (see Sect. 2.7 on our ideas), to the development of an integrated Narrative/Story

In the following, we provide our view of the state of the art in each of several functional areas necessary toward realization of a desired level of automated capability for a future semi-automated, computationally supported analysis prototype that realizes the hybrid capability described. We note, from the literature, a set of particular argumentation-related functional categories: Argument Detection-Construction-Invention-Mining-Accrual and, importantly (as it dominates the literature) Visualization that will serve as the basis for our review.

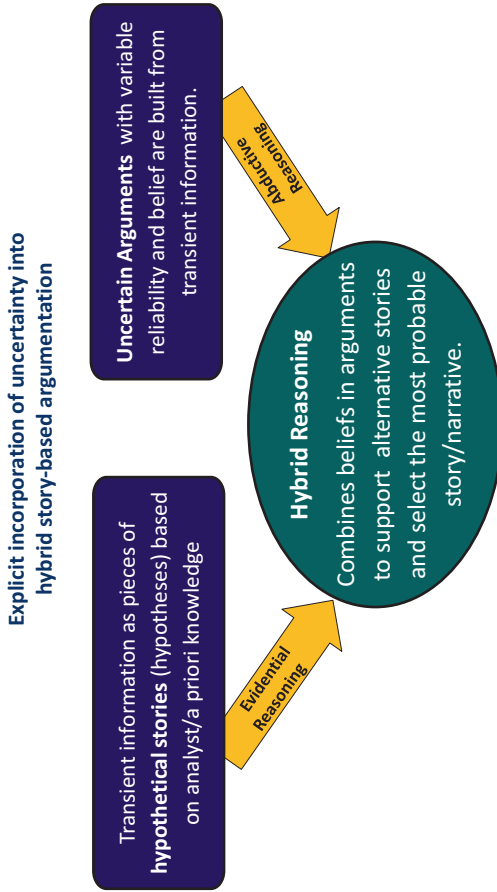
### 2.4.3 *Computational Support to Argumentation: The State of the Art*

It is realized that the input to any modern intelligence analysis system could be in a wide variety of formats and types in terms of media and modalities. As regards the role of these varying inputs toward supporting argument formation, however, it is considered that textual inputs provide the most likely format for somewhat-direct input-to-argument formulation. Most other input types would more likely represent evidential data (such as sensor data) and require a more complex structuring process to frame the data into argument forms. (Later it will be seen that we address sensor

---

<sup>6</sup>Credal will be seen to mean belief but in regard to conducting analysis this term is taken to mean a (human's) conviction of the truth of some statement or the reality of some being or phenomenon especially when based on examination of evidence.

<sup>7</sup>Pignistic is a term coined by Smets and is drawn from the Latin *pignus* for "bet", and can be taken to imply or relate to a probability that a rational person would assign to an option when required to make a decision.



**Fig. 2.4** Depiction of the proposed hybrid approach

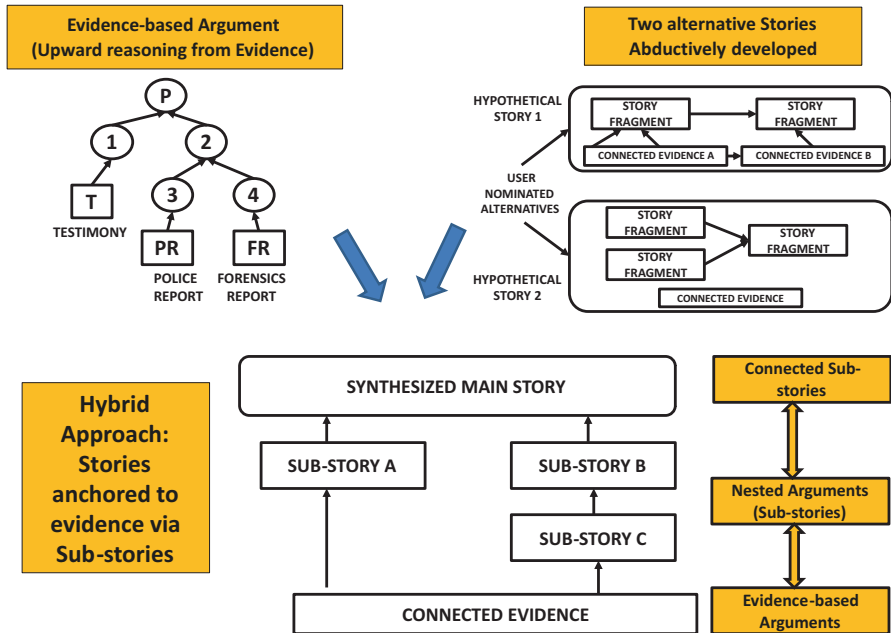


Fig. 2.5 Overview of Bex’s scheme for joint argument-story exploitation

data as an input stream of interest in proposing our design but we note that sensibly all current systems do not include such “hard” data as input.)

Our review of current prototype argument systems shows that the front-ends of these prototypes do not currently provide any automated support to the identification of either the basic linguistic form of an argument (based on lexical content and other factors) or types of argument structures based on argument taxonomies (usually called “schemes” in the argument literature) from textual report, prose-type input, whether structured or not. Thus, a significant human cognitive operation is needed in these prototypes for the formulation of the very basic constructs (arguments) upon which next analysis steps, some computationally-aided, depend. Moen et al. (2007) in discussing the Araucaria prototype<sup>8</sup> designed for argument visualization, say that “The manual structuring of an argumentative text into a graph visualization as is done in the Araucaria research is a very costly job.”

However, we will see that approaches to computational support for extracting parts of or entire argument schemes from text has been addressed but has not, for whatever reasons, been integrated into modern prototype systems. As noted above, this functional activity comes under different names, such as argument detection, argument construction, and argument mining—we simply use the term detection here but draw on works having these other labels to describe what is happening in

<sup>8</sup>An argument mapping tool developed at the University of Dundee; see <http://www.arg-tech.org/index.php/projects/>.

the research community. We will review some sample works in this area and also provide a broader summary view of the state of the art in the next section.<sup>9</sup>

### 2.4.3.1 Argument Detection

Moen et al. (2007) Automatic Detection of Arguments in Legal Texts

This paper describes the results of experiments on the detection of arguments in texts with a focus on legal texts. As will be seen in related works on detection, the detection operation is seen as a classification problem based on defined features of a postulated argument scheme. A classifier is developed in the paper and trained on a set of annotated arguments. Different feature sets are evaluated involving lexical, syntactic, semantic, and discourse properties of the texts, and each of their contributions to classifier accuracy is examined.

Strategies for detecting argument constructs clearly require some defining process for the nature of argument forms or schemes in a linguistic sense; said otherwise, an ontology of argument forms is required. Moen et al state that “The most prominent indicators of rhetorical structure are lexical cues (Allen 1995), most typically expressed by conjunctions and by certain kinds of adverbial groups.” Humans can do this well but one important factor exploited by humans to do so is the context of the textual phrases, and this is very hard to do automatically. The approach in Moens et al. (2007) is admitted to be a bounded first step toward automating this process, and they take an approach built on isolated sentences. They represent sentences as a vector of features and use annotated training data to train a classifier. (It will be seen that this problem is broadly treated as a classification problem in the literature.) We will not review the details of the features and methods but they use a multinomial Bayes classifier and a Maximum Entropy based classifier in this work. It is interesting to see that even simple feature sets yield reasonable (~70+% accuracy) results. The paper also reviews related works and remarks that this type of research on detection is very limited in the legal domain at least (as of the date of this publication, 2007).

Mochales-Palau and Moens (2007)

In a later work, these authors develop an approach to detect sentences that contain argument structures (i.e., apart from not discerning the existence of Walton-type schema; in Walton et al., 2008). A maximum-entropy-based classification is used to determine if input sentences are argumentative or not, and more specifically if they contain a premise, a conclusion or a non-argumentative sentence. These same

---

<sup>9</sup>For the Reader: our reviews in the next section are running commentaries about selected papers from the literature that address each reviewed topic; in various places any emphasis provided is our own. Some excerpts from the original papers are included without quotation.

authors also study and develop a context-free grammar for argument detection in Mochales and Moens (2008), but this was a very limited study across a ten document training set.

### Feng and Hirst (2011), Classifying Arguments by Scheme

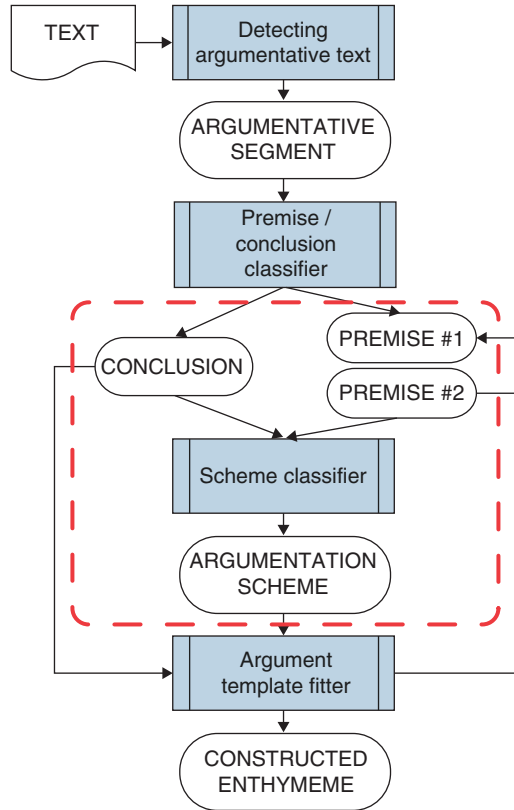
This work is oriented to a subtle issue in argumentation, the issue of enthymemes; as part of an approach to argument detection, in reasonably-frequent cases, there are implicit premises that are never present in the prose text, and these are called enthymemes. To address this issue however, the authors argue that by first identifying the particular argumentation scheme that an argument is using will help to bridge the gap between stated and unstated propositions in the argument, because each argumentation scheme is a relatively fixed “template” for arguing. The idea here is that the argument scheme classification system is a stage following argument detection and proposition classification; that is, a two-stage system involving two different classification systems.

This paper (and some others) relies on the notion of argument schemes or schemata; such schemes are structures or templates for forms of arguments. Walton’s set of 65 argumentation schemes is one of the most-cited scheme-sets in the argumentation literature. According to Feng and Hirst (2011), the five schemes defined in Table 2.2 copied below are the most commonly used ones, and they are the focus of the scheme classification system that is described in this paper. The functional

**Table 2.2** Five top argument schemata from Walton et al. (2008)

<b>Argument from example</b>
<i>Premise:</i> In this particular case, the individual $a$ has property $F$ and also property $G$
<i>Conclusion:</i> Therefore, generally, if $x$ has property $F$ , then it also has property $G$
<b>Argument from cause to effect</b>
<i>Major premise:</i> Generally, if $A$ occurs, then $B$ will (might) occur
<i>Minor premise:</i> In this case, $A$ occurs (might occur)
<i>Conclusion:</i> Therefore, in this case, $B$ will (might) occur
<b>Practical reasoning</b>
<i>Major premise:</i> I have a goal $G$
<i>Minor premise:</i> Carrying out action $A$ is a means to realize $G$
<i>Conclusion:</i> Therefore, I ought (practically speaking) to carry out this action $A$
<b>Argument from consequences</b>
<i>Premise:</i> If $A$ is (is not) brought about, good (bad) consequences will (will not) plausibly occur
<i>Conclusion:</i> Therefore, $A$ should (should not) be brought about
<b>Argument from verbal classification</b>
<i>Individual premise:</i> $a$ has a particular property $F$
<i>Classification premise:</i> For all $x$ , if $x$ has property $F$ , then $x$ can be classified as having property $G$
<i>Conclusion:</i> Therefore, $a$ has property $G$

**Fig. 2.6** Functional flow of argument scheme detection



approach is shown in Fig. 2.6, where it can be seen that argument detection from text precedes the argument *scheme* classification step.

The classifier approach is essentially entropy based. Performance is quite variable, since the various argument schemata vary significantly in the specificity of cue phrases; this is an issue to be dealt with in classifying argument schemata. Note that a training data set for either argument detection or scheme detection requires that the textual corpus be labeled with the “true” argument constructs. This study used the Araucaria data set available at the Araucaria research project website, <http://www.arg-tech.org/index.php/projects/>.

### 2.4.3.2 Argument Mining

Moens (2013), State of the Art in Argument Mining

Argumentation mining is defined by Moens as the (automated/automatic) detection of the argumentative discourse structure in text or speech and the recognition or functional classification of the components of the argumentation. It is clear from

this definition that various functional capabilities are required in mining to include detection of lexical units, identification of sentences containing arguments, and the fit of an argument sample to a predefined argument schema. This type of functionality falls into the domain of Information Retrieval systems, to provide the end user with instructive visualizations and summaries of an argumentative structure. Moens dates argument mining as having started in 2007. The notion of argument “zoning” is mentioned as an area of some study, where a document or corpus is examined to localize sections possibly containing argument-based content. Moens reviews some works that perform these types of functions as typical of the current state of the art; typical Precision/Recall/F measures are in the high 60 to low/mid 70% range, which is just fair performance.<sup>10</sup>

This paper also describes some capability goals for argument mining systems. While discussing the use of machine learning methods, the goal of detecting or recognizing a “full argumentation tree” is mentioned. Cited papers use either a set of piecewise classifiers or a single set-wise or tree-wise classifier, but these are cited only as methodological examples, i.e., these works do not apply such methods to the argument mining problem. Another important argumentation mining issue stated by Moens is the correct identification of the relationships between text segments (e.g., the relationship of being a premise for a certain conclusion) and defining appropriate features that indicate this relationship. Moens suggests that textual entailment in natural language processing, which focuses on detecting directional relations between text fragments may be useful.

### 2.4.3.3 Argument Invention

Walton and Gordon (2012), the Carneades Model of Argument Invention

This paper seems a bit off-topic for our purposes but one aspect that may be of interest is that the mechanics involved in argument invention may hint at how stories (in a knowledge base) and arguments achieve some symbiosis. Argument invention is a method used by ancient Greek philosophers and rhetoricians that can be used to help an arguer find arguments that could be used to prove a claim he needs to defend. The Carneades Argumentation System (named after the Greek skeptical philosopher Carneades) is said by Walton and Gordon to be the first argument mapping tool with an integrated inference engine for constructing arguments from knowledge-bases, designed to support argument invention. It can be said that the notion of invention revolves around the notion of how arguments are evaluated or defended; the idea is to provide automated support to improve the acceptability of an argument. This tool is intended for rhetorical-type applications but conceptually could have applicability in analysis frameworks.

---

<sup>10</sup>The F measure is the harmonic mean of precision and recall, and can be viewed as a compromise between recall and precision. It is high only when both recall and precision are high.



We offer an aside regarding argument evaluation, drawn from Walton and Gordon (2012), as follows: one approach to argument evaluation revolves around the idea of “critical questions” to evaluate an argument. Walton and Gordon (2012, p. 1) suggest: “Critical questions were first introduced by Arthur Hastings (1963) as part of his analysis of presumptive argumentation schemes. The critical questions attached to an argumentation scheme enumerate ways of challenging arguments created using the scheme. The current method of evaluating an argument that fits a scheme, like that for an argument from expert opinion, is by a shifting of the burden of proof from one side to the other in a dialog. When the respondent asks one of the critical questions matching the scheme, the burden of proof shifts back to the proponent’s side, defeating or undercutting the argument until the critical question has been answered successfully. At least this has been the general approach of argumentation theory.” Thus, the presence of critical questions could serve as a mechanism to assure that pro and contra sides of an argument receive attention.

The Carneades design approach provides a number of “assistants” for helping users with various argumentation tasks, including a “find arguments” assistant for inventing arguments from argumentation schemes and facts in a knowledge base, an “instantiate scheme” assistant for constructing or reconstructing arguments by using argumentation schemes, and a “find positions” assistant for helping users to find minimal, consistent sets of statements which would make a goal statement acceptable. The schemes representing knowledge of the domain in the knowledge base must be programmed manually by an expert. A distinctive contribution of the Carneades system is the integration of an inference engine in an argument mapping tool. Although the paper does not emphasize application in the legal domain, it seems clear that this system is oriented to either legal applications or in rhetorical applications as mentioned previously.

#### **2.4.3.4 Argument Visualization (a.k.a. Mapping, Diagramming)**

Argument visualization is often claimed to be a powerful method to analyze and evaluate arguments by providing a capability to perceive dependencies among argument components of evidential components, premises, and conclusions, focusing on the logical, evidential or inferential relationships among propositions. Argument visualization and theoretical modeling play important roles to cope with working memory limitations for problem solving, providing some relief to the cognitive workload that these analyses impute. Since the task of constructing such visualizations (also described in the literature as argument mapping or diagramming) is laborious, researchers have turned to the development of software tools that support the construction and visualization of arguments in various representation formats that have included graphs and matrices among other forms. To say that there have been a number of prototype systems developed that support argument diagramming is rather an understatement—a website provided by Carnegie-Mellon University ([http://www.phil.cmu.edu/projects/argument\\_mapping/](http://www.phil.cmu.edu/projects/argument_mapping/)) shows, just on the first page, the following subset of tools shown in Table 2.3; the complete table goes on

**Table 2.3** Sampling of computer-supported argument diagramming tools (see [http://www.phil.cmu.edu/projects/argument\\_mapping/](http://www.phil.cmu.edu/projects/argument_mapping/))

Tool	Description	Representation	Audience
Athena	Argument mapper from Blekinge Institute of Technology and CERTEC, Sweden	Simplified Toulmin	Education
ArgMAP	Argument mapper	Simplified Toulmin	Research
ArguMed	Argument mapper based on DEFLog	DEFLog (Toulmin extension)	Research
Argutect	Argument mapping-like “thought-processor” from Knosis, Pittsburgh	Thought tree (tree of questions and answers, can be used as simplified Toulmin)	Productivity, education
Araucaria	Argument mapper from University of Dundee, UK	Simplified Toulmin	Education
Belvedere	Collaborative concept mapper and evidence matrix originally developed by D. Suthers at LRDC, Pittsburgh, now at LLIT, University of Hawai’i at Manoa	Inquiry/evidence maps and matrices (links between claims and supporting data)	Education
Causality Lab	Allows students to solve social science problems by building hypotheses, collecting data and making causal inferences	Causal diagram and data charts	Education
Carneades (.pdf)	Toulmin based mathematical model for legal argumentation	Toulmin	Law
ClaimMaker/ ClaimFinder/ ClaimMapper	Concept mapping of knowledge claims from S. Buckingham Shum’s Scholarly Ontologies Project, KMI, Open University, UK	Concept map with semiformal ontology for argumentation	Research

(continued)

**Table 2.3** (continued)

Tool	Description	Representation	Audience
Compendium	IBIS mapping tool originally developed by Verizon Research Labs and associated with CogNexus Institute and KMI, Open University	Dialogue map (concept map with ontology: nodes can represent issues, ideas, pro, con, and notes)	Ill-structured problems
Convince Me	Creates diagrammatic representations of hypothesis and evidence	Evidence map	Education
Deatabase	Deatabase is the world's most useful resource for student debaters. Inside you will find arguments for and against hundreds of debating topics, written by expert debaters, judges and coaches	Communal, simplified Toulmin	Education

for 2–1/2 pages. Note also the range of representational forms, in part dependent on the argument-model used in the application.

The effectiveness of such diagramming or mapping tools is reviewed in (van den Braack et al. 2006). Among the tools that were experimentally tested for their effectiveness were Belvedere, Convince Me, Questmap, and Reason!Able, which are a sampling of tools from Table 2.3.<sup>11</sup> While there are many issues regarding such evaluations discussed by van den Braack including criticisms about statistical testing methodology, the paper concludes that (p. 7) “most results indicated that the tools have a positive effect on argumentation skills and make the users better reasoners. However, most experiments did not yield (statistically) significant effects.” Another study (Twardy 2004) showed that (manual) argument mapping generally helped in understanding arguments and also enhanced critical thinking; the study also showed that the benefits were greater with computer-based argument mapping.

In Mani and Klein (2005), they review structured argumentation as an analysis framework for “open-ended” (i.e., in operational cases where absolute truth is unknown) intelligence analysis. The paper is a short, opinion-type paper and asserts that structured arguments are a means not just of representing and reusing reasoning (one useful benefit), but also a means of communicating and sharing the argument, as analysis is often collaborative. They suggest that one way of assessing the quality of the associated reasoning is to determine how easy the argument is to follow and

<sup>11</sup> See the website listed at Table 2.3 for further details on these systems.

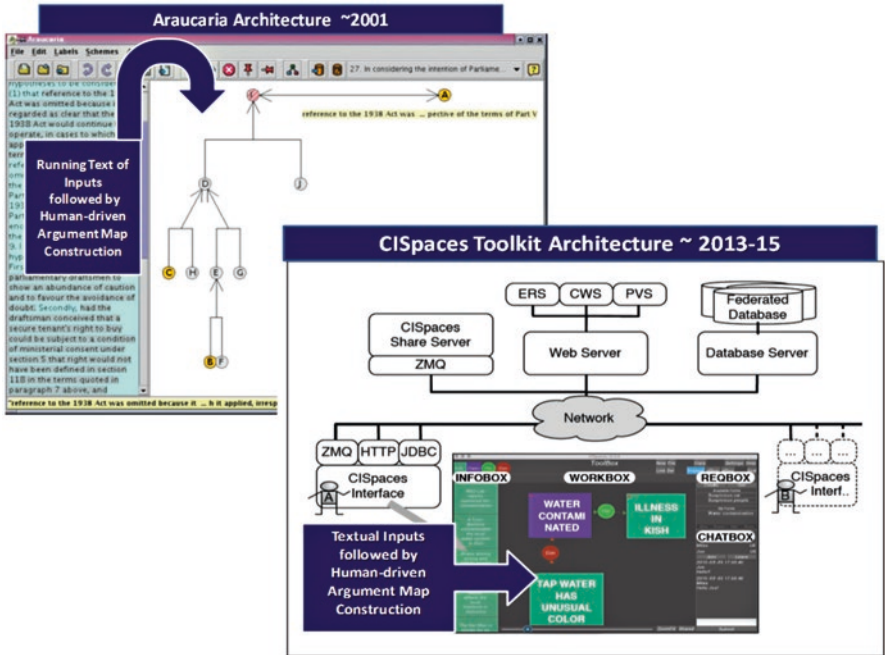


Fig. 2.7 Sampling of argument visualization prototypes

understand. If arguments are constructed in agreeable ways (e.g., based on argument models/schema) and correspondingly visualized, presumably they can be more easily communicated and discussed with others.

To allow an appreciation for what such visualizations look like, we show some examples in Fig. 2.7; these are drawn from Gordon’s presentation in (1996); we use his format as it typically provides a screenshot with some remarks on associated features. Araucaria is very frequently cited as an exemplar of relatively recent prototypes for argument visualization (see for example Suthers et al. 1995; Reed and Rowe 2004). The most recent prototype we are aware of is CISpaces, developed under joint US-UK efforts and led by Norman at the University of Dundee. It can be seen that Araucaria, while having many attractive features, still imputes a high cognitive load onto human analysts is working with streaming text and manually developing the diagrammatic argument constructs. CISpaces incorporates various additional features such as chat for collaborative analysis but still imputes similarly high cognitive workloads for argument mapping; see additional comments below.

## 2.5 Current-Day Computational Support to Argumentation

One other remark that we will offer here is that the greater proportion of research along the lines of computational support schemes for analysis has been carried out in Europe or at least outside of the USA. Among the leading centers of such research are:

- ARG-Tech, at the University of Dundee in Scotland (<http://www.arg.dundee.ac.uk/>)
- Centre for Research in Reasoning, Argumentation and Rhetoric, University of Windsor, Canada (<http://www1.uwindsor.ca/crrar/>)
- Intelligent Systems Group, University of Utrecht, Holland (<http://www.cs.uu.nl/groups/IS/>)
- Intelligent Systems Group, University College London (<http://is.cs.ucl.ac.uk/introduction/>)

To the extent that there is belief that computationally-supported argumentation methods can be helpful to intelligence analysis, this situation should be of concern to the US academic and industrial research communities.

### 2.5.1 *AVERS and CISpaces as Leading Relevant Prototypes*

The research program described in this paper was largely initiated by an early review of a dissertation in Holland having to do with “Sensemaking software for crime analysis” (van den Braack et al. 2007) by Susan van den Braack. That dissertation provided the spark of thinking, as was first explored in that work, for a hybrid, story and argumentation based approach to intelligence analysis since intelligence and criminal analysis requirements have quite similar requirements. This dissertation described AVERS as a prototype developed within the dissertation effort that was designed to explore alternative “scenarios” (stories in effect) based on evidentially-supported arguments. A prototype was developed in the university framework but unfortunately the code for that prototype was not subsequently maintained (we had contacted Dr. van den Braak to explore this). Nevertheless, as described in van den Braack (2010), it is clear that the thinking related to the design and realization of AVERS was very synergistic to our line of research. Formalisms for combining stories and arguments in this hybrid environment were put forward in Bex et al. (2007a, b).

During our program, largely because of our close relations to researchers at the Army Research Laboratory, we learned that, under the “International Technology Alliance (ITA)” program (a US-UK cooperative research program) that a team at the University of Aberdeen (at ARG-Tech as noted above) was carrying out the development of a prototype called “CISpaces”, with goals also similar to ours.

CISpaces was conceptualized as an initial set of tools for collaborative analysis of arguments and debate, providing a uniform way of constructing and exchanging

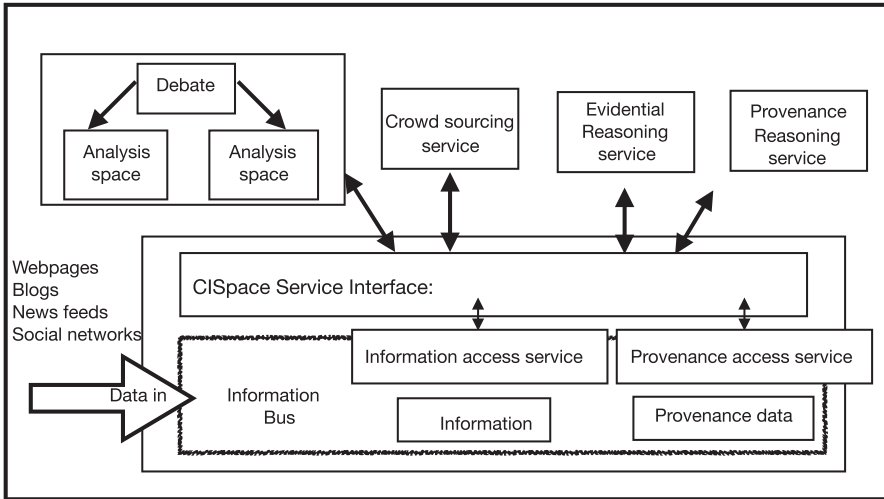


Fig. 2.8 CISpaces functional architecture

arguments based upon argumentation schemes. The top-level functional design is shown in Fig. 2.8 below (Toniolo et al. 2014) and comprises three main services in a service-based architecture:

- the evidential reasoning service, supporting collaboration between users in drawing inferences and forming opinions structured by argumentation schemes;
- the crowd-sourcing service, enabling users to post requests for aggregated opinions from samples of a population;
- the provenance reasoning service, facilitating the storage and retrieval of provenance data including provenance of information and analysis.

The core components of CISpaces, as it is highly oriented to a collaborative, multi-analysts environment, are the WorkBox, the ChatBox and the ReqBox. As described by Toniolo, the WorkBox permits users to elaborate information by adding new claims or by manually importing information and conclusions from different locations; e.g., social networks, blogs. Different forms of argumentation-based dialogue are supported through the ChatBox: collaborative debate, information retrieval through crowd-sourcing, and reasoning about provenance. The list of active debates is intended to be maintained in the ReqBox.

While the development of a real software prototype of this type should be applauded for its forward-thinking approach and for moving the bar of computational support to argumentation to a new level, our thoughts on prototype design addressed other, additional issues:

- Inclusion of both Hard/sensor data as well as Soft/textual/linguistic data as input
  - This is a major change as sensibly all existing argumentation support prototypes are strictly text-input-based

- Major reduction in analyst cognitive workload
  - We see this as involving an aggressive inclusion of front-end, automated processing to aid in argument detection and construction, a major cognitive workload factor of all current prototypes, to include CISpaces.
  - Another aspect is in automated support to final analysis product development, seen as a narrative or story descriptive of a situational estimate of interest (none of the computational systems described here address this at all)
- Major concern for managing information quality along various lines, including automated support for relevance-checking and tracking and assessing provenance of input sources.

Because of our concern for these information quality factors, we established a research thrust along these lines. A later section also addresses our ideas, largely from our Lockheed teammates, on computational support to narrative development.

## 2.6 Computational Support for Narrative Development

As described earlier, for a broad range of intelligence analysis requirements, the desired final output of analysis is a situational picture of some type. In most cases these situations are best communicated as a story or narrative description. However, none of the system concepts and prototypes reviewed here addresses the issue of providing computational support to narrative development. In this next section, we describe our team’s approach and some actual prototyping (done by Lockheed in conjunction with Virginia Tech in a separate effort).

### 2.6.1 *Using Topic Modeling to Assess Story Relevance and Narrative Formation*

As was remarked in particular for Sect. 2.4.3, here too we note that some elements of this section were extracted closely from the conference paper that reported the original work on Topic Modeling carried out in part by Lockheed ATL<sup>12</sup>; see Schlacter et al. (2015) for the original paper.

Storytelling as a data-mining concept was introduced by Kumar et al. (2008). Storytelling (or “connecting the dots”) aims to relate seemingly disjoint objects by uncovering hidden or latent connections and finding a coherent intermediate chain of objects. This problem has been studied in a variety of contexts, such as entity networks (Hossain et al. 2012a, b, c), social networks (Faloutsos et al. 2004),

---

<sup>12</sup>Lockheed’s Advanced Technology Laboratories; see <http://www.lockheedmartin.com/us/atl.html>.

cellular networks (Hossain et al. 2012a), and document collections (Hossain et al. 2012b; Shahaf and Guestrin 2010; Shahaf et al. 2012, 2013). The unsupervised learning technique for storytelling called Story Chaining links related documents in a corpus to build a story or narrative arc. The story chaining approach uses a real-time, flexible storytelling approach that can be used for streaming (online) data as well as for offline data. Because it is fully unsupervised, this approach does not carry the costs of competing approaches such as the need for configuration with domain knowledge or labeling of training data. As such, Story Chaining is ideal for new and frequently evolving domains. Figure 2.9 presents an example of a story chain generated from a corpus of news stories published in Brazil in 2013. The story chains generated from this approach can potentially tell a story about what is happening over time and across news articles by focusing on how the same people, organizations, and locations occur between documents. For this reason, story chains may be considered to be a narrative structure.

Because story chaining is an unsupervised, automated process that generates many results, there is a need to identify the story chains that contain the clearest narratives. One technique uses context overlap as a measure to produce stories that stick to one context by extracting context sentences from a document using a Naive Bayes classifier. Others, for assessing quality, also use dispersion plots and dispersion coefficient to evaluate the overlap of contents of the documents in a chain and thereby quality. Shahaf et al. (2013), as referenced above, define concepts of chain coherence, coverage, and connectivity that offer more insights into the storytelling process. Our approach differs in that it learns a topic model over the corpus and tries to associate certain types of topic change across a story chain as an indicator of how clear of a narrative structure is contained within a story chain.

Topic models are probabilistic models for uncovering the underlying semantic structure of a document collection based on a hierarchical Bayesian analysis of the original texts (Blei et al. 2003). They have been applied to a wide range of text to discover patterns of word use, or topics, across a corpus and to connect documents that share similar structure. In this way, topic models provide a way to create a structure from unstructured text in an unsupervised manner. We leverage them in our work primarily for this reason.

In our research, we have investigated the use of a topic model based analytics to evaluate the clarity of the story chain narrative structure. This work proposes two different kinds of measures of assessment, representativeness and quality.

Firstly, we considered a measure of representativeness that captures how well a story chain represents the corpus from which it was generated. For example, the story chain in Fig. 2.9 was generated from a corpus of thousands of documents published in Brazil in 2013 and it tells a clear story about the Pope visiting Brazil. The stories in the chain take place over a period of 11 days and fit well with the dominant theme of the corpus during that time period which focuses on social issues and protests. Our measure of representativeness is assessed by comparing the similarity of topics found over time in a story chain against those expressed in the corpus during the same time period. This measure assumes the corpus contains dominant topics that are desirable to understand. Our hypothesis for investigating



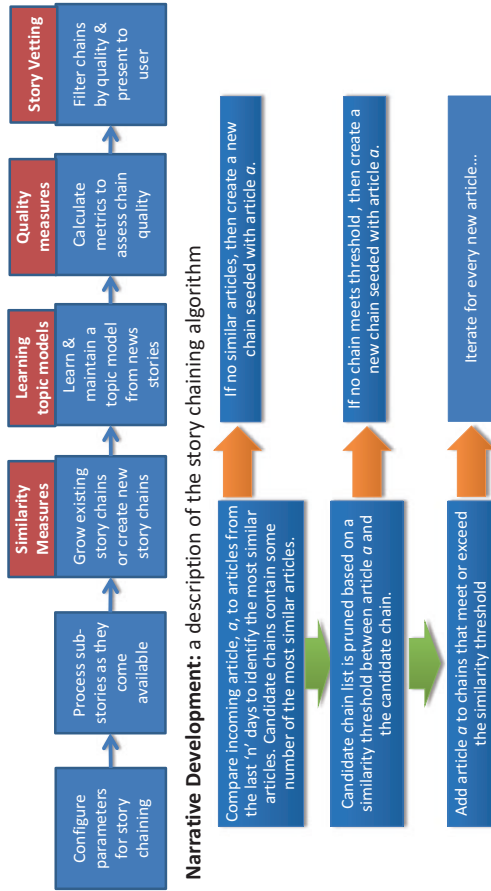


Fig. 2.9 Overview of topic modeling strategy for narrative development

representativeness was the idea that story chains with similar topic expressions to the corpus will convey narratives that are central to the corpus.

Secondly, we considered a measure of quality in which higher quality story chains exhibit a characteristic of focusing on a small number of stable topics, rather than many interleaved or shifting topics. To evaluate this form of quality, we decomposed the measure into two contributing measures, topic persistence and topic consistency.

Topic persistence was designed to capture volatility in topic focus within a story chain. In other words, how often does the topic of a chain shift across each link in the story chain? For example, consider a story chain that has 11 articles such that there are 10 transitions in the story chain connecting one article to the next article in the chain. Topic Persistence (TP) will indicate how well topics persist between links. If most of those ten transitions represent a change in the main topic of the article, then that story chain would have a lower TP score than a chain in which most of those ten transitions represented no change in the main topic. In this way, if a story chain has a high TP score, then most of the links in the chain represent connections between two articles that are discussing the same main topic, and hence, the narrative structure is exhibiting more stable structure for a, hypothetically, better quality chain.

Topic consistency (TC) is a relative assessment of the stability of the main topic of the story chain. More specifically, TC assesses how regularly the main topic of the story chain appears as a main topic of an article within the story chain. For example, if a story chain is made up of ten articles and has a main topic of political unrest, TC will indicate how stable that main topic of political unrest is by looking at each of the ten contributing articles and seeing if political unrest appears as the primary topic within those ten articles. If only three of those ten articles are focused on political unrest for a  $TC = 3/10$  or 30%, that means that most of the articles in the chain are focused on (1) different topics, and (2) a variety of different topics such that consensus did not exceed three. Compare this to a scenario in which the story chain had seven articles focusing on political unrest where  $TC = 7/10$  or 70%. In this case, the topic is much more consistent throughout the chain (not necessarily consecutively) and hence, the narrative structure more centered on political unrest and, hypothetically, of better quality.

Our results indicate that using topic model based analytics to predict the quality of a narrative structure is a promising avenue of research. We found correlations between all of our analytics and the human scoring of our story chains, with particularly strong correlation to the relevance metric.

The need to build situational awareness from increasingly large sets of textual data requires automatic methods to construct narrative structures from text without regard to domain factors such as actors, event types, etc. The metrics presented in this paper provide a means to assess these narrative structures so that only the most useful narrative structures are transformed into narratives. In this work, we define three metrics of relevance, topic persistence and topic consistency to assess narrative structure. We specify and implement these measures with respect to a narrative structure of story chains generated by an unsupervised narrative generation

technique presented in Hossain et al. (2012b). This data is processed to provide analytical evidence for the usefulness of these metrics for identifying high quality story chains.

## 2.7 Developing a Functional Design for an Advanced-Capability Prototype

An effective approach to architecting our proposed decision-support concept requires that we assert our views of the overall reasoning process from evidence to decision-making and decision enablement. Most traditional characterizations describe decision-making (DM) as contemplative, analytic, involving nomination and evaluation of options that are weighed in some context, eventually leading to a choice of a “course of action (COA)”. This model, often labeled as the “System 2” model, can be seen in most descriptions of the “Military Decision-Making Process” or MDMP as for example in published military Field Manuals such as in HQ, Dept of Army (2010). The literature also identifies a “System 1” or largely intuitive decision-making paradigm (IDM) that operates in conjunction with System 2 processes in what is argued to be an improved DM process model, often called the “Dual-Process Model”. Most research in decision support however has focused on System 2 DM ideas since this model is quantitative and can be mathematically studied using notions of utility theory and other frameworks for mensuration. We intend however to factor the Dual-Process Model concept into our systemic design approach; the basis of this rationale cannot be elaborated here but we offer our references for the interested reader, e.g., Croskerry (2009) and Djulbegovic et al. (2012).

Furthermore, in our view of the System Support context for DM, we see what today are called Sensemaking processes, as lying between automated System Support capabilities, such as Data Fusion processes and DM processes, in a stage wherein “final” situation assessments and understandings (in the human mind) are developed. Thus, our view of this meta-process is as a three-stage operation: System Support (SS) as an automated process that nominates algorithmically-formed situational hypotheses (such as from the combined operations of data fusion and argumentation), followed by human-computer, mixed-initiative processes for Sensemaking and symbiosis, whose narrative-type products provide the vetted situational assessments needed for decision-making. There is a substantive literature on Sensemaking, such as those previously cited (Llinas 2014a, b; Gross et al. 2014). Our key thoughts on and rationale for the meta-architecture for System Support described briefly here have been summarized in (Llinas 2014a, b). Finally, in the face of significant production pressures and rapidly proliferating data availability—and the resulting data overload deluging the professional analyst—it is increasingly easy for analysts and decision-makers to be trapped by shallow, low-rigor analysis; improvements in rigor have been previously discussed and are part of our proposed design. At the highest level, and consistent with the System Support/Fusion—

Sensemaking–Decision-making interdependent processes concept, we see our initial prototype as embedded in the Sensemaking dynamic (note that this is an initial, design-in-process), as shown in Fig. 2.10.

Building on these ideas, we formed our initial functional design as shown in Fig. 2.11. Included in this design are the specifics of the Hard-Soft data association operations that would be part of the Fusion/System Support segment in an eventual final design. The figure can be examined by starting at the bottom where notional Use Cases are also shown—these include current service-specific mission operations, Joint service operations, and a technological type thrust that examines the proposed methods as having disruptive properties:

- **Army: Operations in Megacities, Syrian Civil War**
  - **Megacity operations are an evolving new Army interest**
- **Navy: Piracy (NATO), Autonomous ISR Systems**
  - **Piracy is a continuing NATO interest, ONR has considerable interest in UAV/UXV operations**
- **Joint: Expeditionary Operations (Anti-Access Area Denial, A2AD),**
  - **Joint operations dealing with A2AD issues are an evolving widespread interest**
- **Assess Hybrid Argumentation Technology as Disruptive**
  - **And of course these proposed methods can be studied from the technological point of view as a new and disruptive capability**

For any Use Case, we envision that there would be the opportunity or need to enable both Hard and Soft data stream inputs of various types as peculiar to each of the Use Cases. Based on our own research in computational support techniques for Relevance filtering and Provenance accounting, we show those two functional blocks first, operating on both data streams. (Note that there may be some preprocessing required for the Hard Data stream to frame the results into Entity-Attribute sets.) These filters ideally provide relevant and qualified data to two processes: a Natural Language Processor (NLP) and Argument Detection and Nomination (ADM) process. The functions of each of these operations are:

- NLP: extract Named Entities and associated features and attributes of those Named Entities
- ADM: detect and construct argument phrases with labeled Schemas as possible

Metadata is also considered for both processing operations. The outputs of both NLP and ADM (and possible Hard Data preprocessing) are inputs to the Hard/Soft Data Association process that correlates the Entity-Attribute sets and forms the associated and reconciled fused Entity/Attribute results, i.e., the associated, fused Entity/Enriched Attribute evidential data set as shown on Fig. 2.11. This output provides a feedback to the Argument Detection processing (that contains labeled

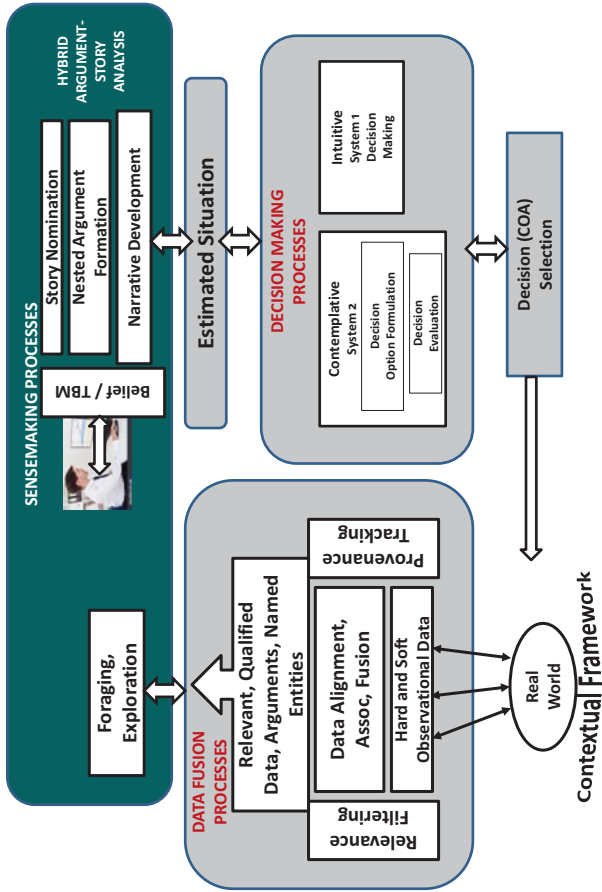


Fig. 2.10 The hybrid scheme in the context of a meta-architecture involving fusion-sensemaking-decision-making (Llinas 2014a, b)

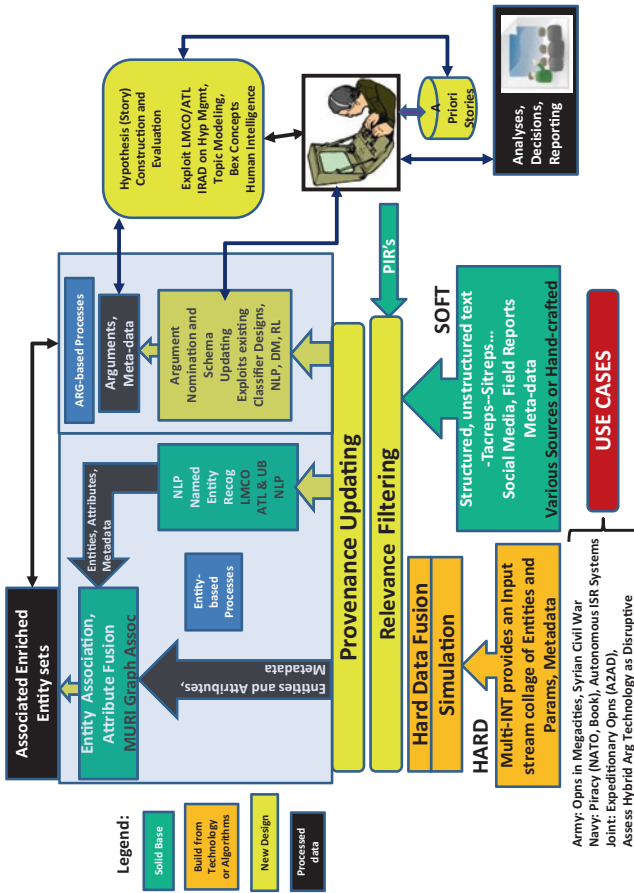


Fig. 2.11 The final functional design

Entities) so that these identified Entities can be enriched with the associated/fused Attributes. Note that there can be possible outlier Entities here, since the ADM process is only Soft-data-based; this is a reconciliation issue yet to be determined. One idea is to engage the human analyst in the process of integrating and managing these outlier Entities. At this point, this front-end processing has *automatically* produced nominated arguments with associated and enriched/fused Entity/Attribute pairs—this capability is a high-priority goal of our approach as this capability has the potential to greatly reduce human cognition workload in terms of argument construction, a major issue even in the most modern prototypes we have reviewed. These nominated arguments then are vetted with analyst intervention and once vetted can provide draft input to our proposed Topic Modeling/Narrative Construction software that aids in a mixed-initiative, human-machine symbiotic process of hybrid argument/story combination. This approach takes into account the uncertainty inherent into the environment as well as the results of argument detection and nomination. These operations will likely involve the management of competing hypotheses for which Lockheed Internal Research and Development (IRAD) software may also provide automated support. These operations would take advantage of Bex's theories and methods for hybrid correlation of the evidentially-grounded arguments and stories emanating both from the analyst and from the Topic Modeling story-nomination process.

This is of course an ambitious vision but is one that sets a new milestone we think for automated support to intelligence analysis. A number of details have to be worked out but the considerably advanced capabilities that a system like this can provide will move the bar forward in terms of revolutionary, disruptive automated support to intelligence analysis.

### ***2.7.1 Looking Ahead: Possible Test and Evaluation Schemes***

Given that our end-goal of this project was to develop initial thoughts on a functional design, it was considered necessary to explore possible strategies for Test and Evaluation (T&E) as well as possible metrics for evaluation, since the quality of any possible prototype would be measured by some appropriate T&E approach.

There are various important functions in the proposed top-level design of Fig. 2.11. As the multisource Data Association process is considered key in any Information Fusion process, one critical aspect of a T&E approach would suggest a scheme for evaluating Hard-Soft Data Association. Here, we would suggest the approach of the MURI program that the Center for Multisource Information Fusion at the University at Buffalo developed as at least a starting approach (this is well-documented in Gross et al. 2014; Date et al. 2013a, b); this technique was explored and tested with good success on that program.

Testing of Natural Language Processing (NLP) methods is a very broad topic but one focus for the proposed design is on Named Entity extraction, a key capability for good performance in the proposed scheme. Here too the methods employed on

the prior MURI program could be applied to evaluate performance in any Use Case application; these techniques are discussed in Shapiro (2012).

There is not much literature on specific evaluation techniques for the various front-end argument detection/construction methods we would intend to explore, but most of these rely on some type of classification framework, and evaluation of such text extraction methods. The cited literature of Sect. 2.6, along with various survey papers on classifier evaluation form an adequate starting point for developing an evaluation approach.

Evaluating the quality of argument constructs is an area where there is considerable literature. There are various websites on this topic (e.g., <http://www.csuchico.edu/~egampel/students/evaluating.html>) and a wide variety of papers that address this topic, e.g., Corner and Hahn (2009). Much of the literature discusses notions of argument strength, different for deductive, inductive, and abductive arguments and introduces related ideas on validity of premises and other issues. This literature is helpful toward test planning but we prefer Dahl's ideas on the notion of argument persuasiveness that in turn relates to ideas on "explanatory coherence" as a technique for evaluating the persuasiveness of arguments; see Dahl (2013), Thagard (2000), and Ng and Mooney (1990).

Of course, the best evaluation approach would reveal the impacts of these combined technologies on mission-based analysis effectiveness; however, since the proposed design and suggested methods are, in our opinion, still at the formative stage, much testing and evaluation would have to be done to first establish technological credibility before mission effectiveness assessments could (or should) be carried out.

**Acknowledgement** This publication results from research supported by the Naval Postgraduate School Assistance Grant No. **N00244-15-1-0051** awarded by the NAVSUP Fleet Logistics Center San Diego (NAVSUP FLC San Diego). The views expressed in written materials or publications, and/or made by speakers, moderators, and presenters, do not necessarily reflect the official policies of the Naval Postgraduate School nor does mention of trade names, commercial practices, or organizations imply endorsement by the U.S. Government.

## References

- Schlacter, J., et al. (2015), Leveraging Topic Models to Develop Metrics for Evaluating the Quality of Narrative Threads Extracted from News Stories, 6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences, AHFE 2015 Procedia Manufacturing, Volume 3
- Allen, J. (1995), *Natural Language Understanding* (2nd ed.). Benjamin-Cummings Publishing Co., Inc., Redwood City, CA, USA
- Andrews, C. and North, C. (2012), "Analyst's Workspace: An Embodied Sensemaking Environment For Large, High-Resolution Displays", *Proc. 2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, Seattle, WA.
- Bex, F., S. van den Braak, H. van Oostendorp, H. Prakken, B. Verheij, and G. Vreeswijk (2007a), "Sense-making software for crime investigation: how to combine stories and arguments?," *Law, Probability and Risk*, vol. 6, iss. 1-4, pp. 145-168.



- Bex, F. et al. (2007b), Sense-making software for crime investigation: how to combine stories and arguments?, *Law, Probability and Risk*.
- Bex, F. (2013) Abductive Argumentation with Stories. ICAIL-2013, in: *Workshop on Formal Aspects of Evidential Inference*, 2013
- Bier, E.A., Ishak, E.W., and Chi, E. (2006), "Entity Workspace: An Evidence File That Aids Memory, Inference, and Reading", In *ISI*, San Diego, CA, 2006, pp. 466-472.
- Blei, D.M., Ng, A. Y., and Jordan, M. I. (2003), "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022.
- Corner, A. and Hahn, U. (2009)., Evaluating Science Arguments: Evidence, Uncertainty, and Argument Strength, *Journal of Experimental Psychology Applied* 15(3):199-212.
- Croskerry, P. (2009), *A Universal Model of Diagnostic Reasoning*, Academic Medicine, Vol 84, No 8, pp1022–8.
- Dahl, E. (2013), *Intelligence and Surprise Attack: Failure and Success from Pearl Harbor to 9/11 and Beyond*, Georgetown University Press.
- Date, K., Gross, G. A., Khopkar, S, Nagi, R. and K. Sambhoos (2013a), "Data association and graph analytical processing of hard and soft intelligence data", *Proceedings of the 16th International Conference on Information Fusion (Fusion 2013)*, Istanbul, Turkey
- Date, K., GA Gross, and Nagi R. (2013b), "Test and Evaluation of Data Association Algorithms in Hard+Soft Data Fusion," *Proc.of the 17th International Conference on Information Fusion*, Salamanca, Spain
- Djulbegovic, B., et al. (2012), Dual processing model of medical decision-making, *BMC Medical Informatics and Decision Making*, Vol. 12
- Faloutsos, C. KS McCurley, and Tomkins A. (2004), "Fast discovery of connection subgraphs." *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* 22: 118-127.
- Feng, V.W. and Hirst, G.. (2011), Classifying Arguments by Scheme, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 987–996, Portland, Oregon.
- Gordon, T.F. (1996), Computational Dialectics, In Hoschka, P., editor, *Computers as Assistants - A New Generation of Support Systems*, pages 186–203., Lawrence Erlbaum Associates.
- Gross, G., et al. (2014), Systemic Test and Evaluation of a Hard+Soft Information Fusion Framework; Challenges and Current Approaches, in: "*Fusion2014*," *International conference on Information Fusion*,
- Haenni, R. (2001) Cost-bounded argumentation, *International Journal of Approximate Reasoning*, 26(2):101–127.
- Hastings, A.C. (1963), *A Reformulation of the Modes of Reasoning in Argumentation*, Ph.D. dissertation, Northwestern University, Evanston, Illinois.
- Headquarters, Dept of Army (2010), Army Field Manual 5-0, The Operations Process
- Hossain, MS, M Akbar, and Nicholas F Polys (2012a). "Narratives in the network: interactive methods for mining cell signaling networks." *Journal of Computational Biology* 19.9:1043-1059.
- Hossain, MS., et al. (2012b), Connecting the dots between PubMed abstracts PloS one 7.1
- Hossain, M. S., Butler, P., Boedihardjo, A. P., and Ramakrishnan, N. (2012c). Storytelling in entity networks to support intelligence analysts, *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*.
- Klein, G., et al. (2006), Making Sense of Sensemaking 2: A Macrocognitive Model, *IEEE Intelligent Systems*, Volume:21, Issue: 5.
- Kumar, D., Ramakrishnan, N., Helm, R. F., and Potts, M. (2008), Algorithms for storytelling, *IEEE Transactions on Knowledge and Data Engineering*, 20(6), 736-751.
- Llinas, J. (2014a), Reexamining Information Fusion--Decision Making Inter-dependencies, in *Proc. of the IEEE Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*, San Antonio, TX.
- Llinas, J, Nagi, R., Hall, D.L., and Lavery, J. (2010), "A Multi-Disciplinary University Research Initiative in Hard and Soft Information Fusion: Overview, Research Strategies and Initial Results", *Proc. of the International Conference on Information Fusion*, Edinburgh, UK.

- Llinas, J. (2014b), A Survey of Automated Methods for Sensemaking Support, *Proc of the SPIE Conf on Next-Generation Analyst*, Baltimore, MD
- Mani, I. and Klein, G.L. (2005), Evaluating Intelligence Analysis Arguments in Open-ended Situations, *Proc of the Intl Conf on Intelligence Analysis*, McLean Va.
- Mochales, R. and Moens, M. (2008), Study on the Structure of Argumentation in Case Law, *Proceedings of the 2008 Twenty-First Annual Conference on Legal Knowledge and Information Systems: JURIX 2008*
- Mochales-Palau, R. and Moens, M. (2007), Study on Sentence Relations in the Automatic Detection of Argumentation in Legal Cases, *Proceedings of the 2007 Twentieth Annual Conference on Legal Knowledge and Information Systems: JURIX 2007*
- Moens, M., et al (2007), Automatic Detection of Arguments in Legal Texts, *Proceedings of the 11th international conference on Artificial intelligence and Law*.
- Moens, M. (2013), Argumentation Mining: Where are we now, where do we want to be and how do we get there?, *FIRE '13 Proceedings of the 5th 2013 Forum on Information Retrieval Evaluation*.
- Ng, H.T., and Mooney, R.J. (1990), On the Role of Coherence in Abductive Explanation, in *Proceedings of the Eighth National Conference on Artificial Intelligence (AAAI-90)*
- Pirolli, P. and Card, S. (2005), The Sensemaking Process and Leverage Points for Analyst Technology as Identified Through Cognitive Task Analysis, In *Proceedings of 2005 International Conference on Intelligence Analysis* (McLean, VA, USA, May, 2005). pp.337-342, Boston, MA.
- Reed, C. and Rowe, G. (2004), ARAUCARIA: Software for Argument Analysis, Diagramming and Representation, *International Journal on AI Tools* 13 (4) 961–980.
- Schum, D. (2005), Narratives in Intelligence Analysis: Necessary but Often Dangerous, *University College London Studies in Evidence Science*.
- Shahaf, D., and Guestrin, C., (2010), Connecting the dots between news articles, *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*: 623-632.
- Shahaf, D, Guestrin, C, and Horvitz, E. (2012) Trains of thought: Generating information maps.” *Proceedings of the 21st international conference on World Wide Web*: 899-908.
- Shahaf, D. et al (2013), Information cartography: creating zoomable, large-scale maps of information. *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*: 1097-1105.
- Shapiro, S. (2012), Tractor: Toward Deep Understanding of Short Intelligence Messages, University Seminar, available at: <http://studylib.net/doc/10515245/tractor-toward-deep-understanding-of-short-intelligence-m>, 2012
- Simari, G and Rahwan, I (2009) *Argumentation in artificial intelligence*, Springer.
- Smets, P. (1994), The transferable belief model, *Artificial Intelligence*, Volume 66, Issue 2, Pages 191-23.
- Stasko, J., Gorg, C., Liu, Z., and Singhal, K. (2013), “Jigsaw: Supporting Investigative Analysis through Interactive Visualization”, *Proc. 2007 IEEE Conference on Visual Analytics Science and Technology (VAST)*, Sacramento, CA.
- Suthers, D. et al, (1995), ‘Belvedere: Engaging students in critical discussion of science and public policy issues’, in *AI-Ed 95, the 7th World Conference on Artificial Intelligence in Education*, pp. 266–273, (1995).
- Thagard, P. (2000), *Probabilistic Networks and Explanatory Coherence*, *Cognitive Science Quarterly* 1, 93-116
- Toniolo, A., Ouyang RW, Dropps T, Allen JA, Johnson DP, de Mel G, Norman TJ, (2014), Argumentation-based collaborative intelligence analysis in CIspace, in *Frontiers in Artificial Intelligence and Applications*; Vol. 266, IOS Press
- Twardy, C. (2004): Argument maps improve critical thinking. *Teaching Philosophy* 27 (2):95--116
- van den Braack, S. W. et al, (2007), AVERS: an argument visualization tool for representing stories about evidence, *Proceedings of the 11th international conference on Artificial intelligence and law*, Stanford, CA.

- van den Braack, S. W. (2010), *Sensemaking software for crime analysis*, Dissertation, Univ of Utrecht, Holland.
- van den Braack, S.W., et al (2006), A critical review of argument visualization tools: do users become better reasoners?, *ECAI-06 CMNA Workshop*, 2006
- Walton, D., Reed, C., and Macagno.F. (2008), *Argumentation Schemes*, Cambridge University Press.
- Walton, D. and Gordon, T.F. (2012), The Carneades Model of Argument Invention, *Pragmatics & Cognition*, 20(1), Web 2.0, Amsterdam, The Netherlands

# Chapter 3

## Task Allocation Using Parallelized Clustering and Auctioning Algorithms for Heterogeneous Robotic Swarms Operating on a Cloud Network

Jonathan Lwowski, Patrick Benavidez, John J. Prevost, and Mo Jamshidi

### 3.1 Introduction

In recent years, robotic swarms have become increasingly popular in both civilian and military applications. These applications include search and rescue, land surveying and surveillance. The popularity increase is due to the fact that the robotic swarm can perform more complex tasks than a single robot. For example, when an autonomous surface vehicle (ASV) is traveling through an obstacle filled environment, it is difficult for it to plan a long-term path that includes obstacle avoidance. When paired with an unmanned air vehicle (UAV), such as a micro-aerial vehicle (MAV), the MAV provides a different perspective of the environment to the ASV allowing a long-term path plan (Lwowski et al. 2016). Ray et al. (2009) also developed a robotic swarm to complete a complex task. They used a multi-agent rover network to control a group of rovers to remain in a desired formation. Each agent in their swarm can estimate the behavior of the other agents, which reduces the necessary communications between the agents. This swarm can be used for a variety of tasks such as search and rescue. Gallardo et al. (2016) also developed a robotic swarm to maintain a desired formation that could be used for search and rescue. This swarm utilizes a leader-follower approach where the swarm follows a virtual leader position at a point detected by a MAV's bottom facing camera, relative to the agents in the swarm.

As the size of the robotic swarms increase, harder problems can also be solved, but this comes with increased complexity. One of the important problems to solve when dealing with larger robotic swarms is task allocation, which has resulted in increasing interest in the research community (Gerkey and Mataric 2004). There are

---

J. Lwowski (✉) • P. Benavidez • J.J. Prevost • M. Jamshidi  
Autonomous Control Engineering Lab, The University of Texas at San Antonio,  
San Antonio, TX, USA  
e-mail: [Jonathan.Lwowski@gmail.com](mailto:Jonathan.Lwowski@gmail.com); [Patrick.Benavidez@utsa.edu](mailto:Patrick.Benavidez@utsa.edu); [Jeff.Prevost@utsa.edu](mailto:Jeff.Prevost@utsa.edu);  
[mo.jamshidi@utsa.edu](mailto:mo.jamshidi@utsa.edu)

several different factors in the literature that have been studied relating to task allocation for robotic swarms. One of the factors is team organization. Team organization is the hierarchical system the robotic swarm uses to make decisions. There are two main types of team organizations, centralized and distributed. Centralized robotic swarms have a leader that is giving orders, or making plans for the other agents in the system (Shia 2011). For example, Liu and Kroll (2012) developed a centralized task allocation and path planning algorithm using A\* for inspecting industrial plants for gas leaks. In distributed robotic swarms, all of the agents are mainly governing themselves, and therefore do not have a central leader (Shia 2011). For example, Giordani et al. (2014) developed a decentralized (distributed) algorithm for multi-robot task allocation. Their algorithm employs each agent as a decision maker in the system, which uses a distributed version of the Hungarian algorithm (Kuhn 1955).

Another factor that has been studied relating to task allocation for robotic swarms is communication losses or delays. Communication delays and failures can have huge impacts on the task allocation method used for the swarm. For example, in a centralized swarm, if communication to a leader is lost, then all of the other agents will fail because they need the input from the leader. In a distributed swarm, if an agent is designed to complete a certain task, but has no communication to the rest of the swarm, then it will not be assigned to that or any other tasks. Several groups have done research on these topics. For example, Sujit and Sousa (2012) developed a behavior-based coordination algorithm for a multi-agent system that experiences partial and full communication failures. These algorithms modify the behavior of faulty and non-faulty agents depending on the types of failure they encounter. In another paper, Dutta et al. (2016) designed a nonlinear controller that uses communication connectivity to control a swarm of MAVs to organize into a desired formation while maintaining strong connectivity. This is important because not only does the swarm stay in its desired formation, but the MAVs will not lose communication to each other.

In this paper, a heterogeneous centralized robotic swarm, consisting of ASVs and MAVs (equipped with cameras and GPS), is presented. This system can be used for a variety of scenarios such as the extinguishing of hot spots after a forest fire, and autonomously detecting and watering dry spots on a farm. For the remainder of this paper, the algorithms will be discussed with the scenario of a search and rescue of people floating in the ocean after a cruise-ship disaster in mind. The heterogeneous robotic swarm works cooperatively with a cloud network to show how the advantages of each agent can perform task allocation in an organized and efficient manner. For example, the aerial vantage point of the MAVs are used to detect and localize the people from the ship floating in the ocean, which would be hard to do using just ASVs. On the other hand, the ASVs are used to rescue the people floating because it would be impractical to have the MAVs perform the rescue operations. This symbiotic relationship allows the system to utilize the advantages of each system to save people in an efficient and fast manner.

The rest of the paper is organized as follows. Section 3.2 will discuss the robotic swarm system and all of the algorithms used. The simulation and hardware emulation results are discussed in Sect. 3.3, and concluding remarks will be in the final section.

**Table 3.1** Specifications of cruise-ships and coast guard ships

Ship	Capacity (people)	Speed (km/h)	Length (m)
Harmony of the Seas (large cruiseliner) (France 2012)	6360	46	362
Queen Mary 2 (small cruiseliner) (Cunard 2002)	3090	56	345
45-foot Motor Lifeboat (large coast guard boat) (US Coast Guard 2007)	34	46	15
Defender-class Boat (small coast guard boat) (US Coast Guard 2006)	10	85	9

## 3.2 Robotic Swarm Methodology

### 3.2.1 Scenario

In recent news, several cruise-ship crashes have occurred, such as the Costa Concordia crash where 32 people died on January 13, 2012 (BBC News 2015). In this crash, caused by human error, approximately 4250 people were on-board the cruise-liner. Considering how many people were on the Costa Concordia, many more people could have been injured. These cruise-ship disasters inspired the development of our system, but this system can easily be used for a variety of other scenarios. To perform the simulations at a proper scale, background research was conducted to determine the relevant parameters of cruise-ships and coast guard ships. This information can be seen in Table 3.1. The information in Table 3.1 was used during the simulations.

### 3.2.2 System Overview

The system works by using six main tasks; (a) localization of the people to be rescued, (b) building the map, (c) clustering the victims, (d) meta-clustering the clusters, (e) auctioning the meta-clusters, and (f) the shortest path solver. These tasks are all used sequentially, beginning when an alert about the location of a ship crash has been received. Once the ship crash's location has been received, a group of MAVs equipped with bottom-facing stereo cameras fly over the scene. Using the stereo cameras, the MAVs localize the people floating in the ocean. The locations of the victims are sent to the cloud, and the cloud begins to build a map of the environment. It is assumed that the mothership has an on-board cloud network that is available to all of the agents in the swarm. Once the map has been built, the cloud network will simplify the map by clustering the people into groups. Information about these groups such as location and size are stored on the cloud. After the people have been clustered into groups, the cloud then clusters the groups into meta-clusters. The reason to perform this second round of clustering is to reduce the number of clusters

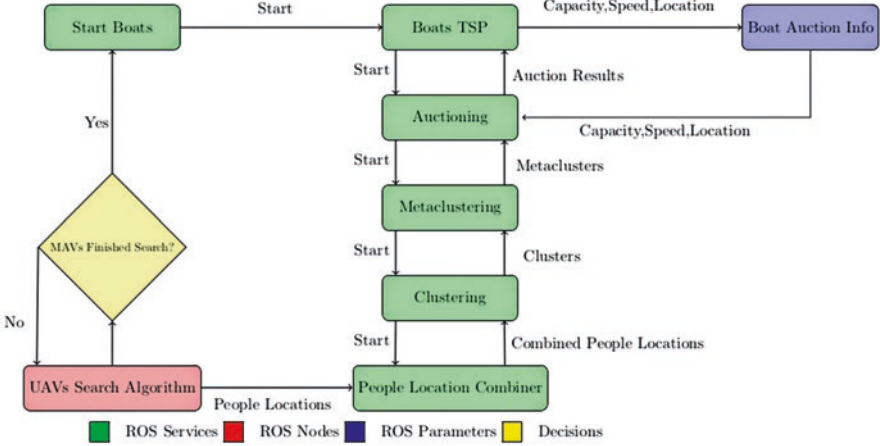


Fig. 3.1 Overview of the system in ROS

to be the same as the number of ASVs in the swarm. Once again, information about the meta-clusters such as location and size, are stored in the cloud. Using the meta-cluster information stored in the cloud, the meta-clusters are auctioned off to ASVs. After each ASV is assigned a meta-cluster, a traveling salesman solver is applied to find an optimal path between the ASVs and each cluster in the assigned meta-cluster. This system, seen in Fig. 3.1, was implemented in Robot Operating System (ROS) (Quigley et al. 2009) due to its robust message passing interface.

### 3.2.3 Localization of People

To localize the surviving people floating in the ocean, the MAVs are equipped with bottom facing stereo cameras. The people to be rescued, represented by colored circles in the simulation, are detected by using simple color thresholding techniques. Figure 3.2 shows how the simulation environment was setup.

Once a survivor is detected by the stereo camera, the location of the person in the camera frame is transformed to the camera frame using Eqs. (3.1) and (3.2), where  $x$  and  $y$  are the pixel locations of the person,  $X_{cf}$ ,  $Y_{cf}$ , and  $Z_{cf}$  are the estimated coordinates of the person with respect to the camera frame, and  $Q$  is the rectification transformation matrix.

$$\begin{bmatrix} X \\ Y \\ Z \\ W \end{bmatrix}^T = Q \begin{bmatrix} x & y & disparity(x,y) & 1 \end{bmatrix}^T \quad (3.1)$$

**Fig. 3.2** Aerial view of the simulation environment



$$\left( X_{cf}, Y_{cf}, Z_{cf} \right) = \left( \frac{X}{W}, \frac{Y}{W}, \frac{Z}{W} \right) \quad (3.2)$$

A visual representation of Eqs. (3.1) and (3.2) can be seen in Fig. 3.3. Then  $X_{cf}$ ,  $Y_{cf}$ , and  $Z_{cf}$  are transformed from the camera frame to the world frame using the traditional methods as described by Ma et al. (2003).

To validate the localization methods, the Robotic Operating System (ROS) (Quigley et al. 2009), Gazebo (Koenig and Howard 2004), and RViz (Kam et al. 2015) were used. ROS is a robust, open source message passing infrastructure used to manage the communication between the agents of the robotic swarm. Gazebo is a three-dimensional open-source multi-robot simulator. RViz is a toolkit for real domain data visualization. The localization method was tested by having three MAVs fly over an area to find and localize the victims using the method described above. The results of this simulation can be seen in Fig. 3.4, where the green circles represent the actual locations of the people to be rescued and the blue triangles represent the estimated locations of the people. The results of this simulation show that the localization methods described above work correctly, with some error. These errors include positional estimation errors along with the mistaken detection of some people. These errors could be due to environmental issues such as people being too close to each other. These errors could also be sensor related issues. Although the detection algorithm is not perfect, it can be used as a proof of concept, and needs to be improved in the future. Various tracking algorithms, such as extended Kalman filtering, can be used to improve the estimation of the location of the people. Once the people have been localized, the information can be sent to the cloud cluster to build the global map.



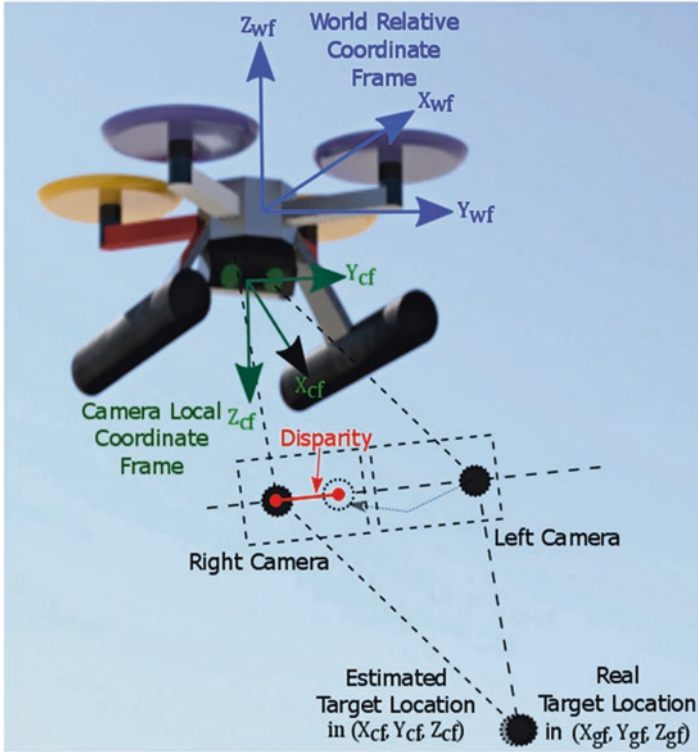


Fig. 3.3 Visual representation of the stereo camera localization model

### 3.2.4 Building the Map

Once the cloud receives the locations of the people to be rescued, a global positioning map can be made. However, calculating the optimal path to every person for each ASV would be too expensive. Therefore, we perform several clustering algorithms on the map using the cloud to create a simpler map for the ASVs.

#### 3.2.4.1 Clustering the People

To simplify the map, the cloud clusters the localized people into large groups using a modified k-means clustering algorithm. The modified k-means algorithm, Constrained Cluster Radius (CCR) K-Means Clustering, is described below in Algorithm 3.1, where  $k$  is the number of clusters that is inputted into the k-means algorithm. The CCR K-Means Clustering ensures that the cluster will be smaller than a given desired radius. This constraint is important because the clusters need to be of a reasonable size so once the ASVs arrive, the cluster size will not be too large for the rescue team to save all of the nearby people in a reasonable time frame.

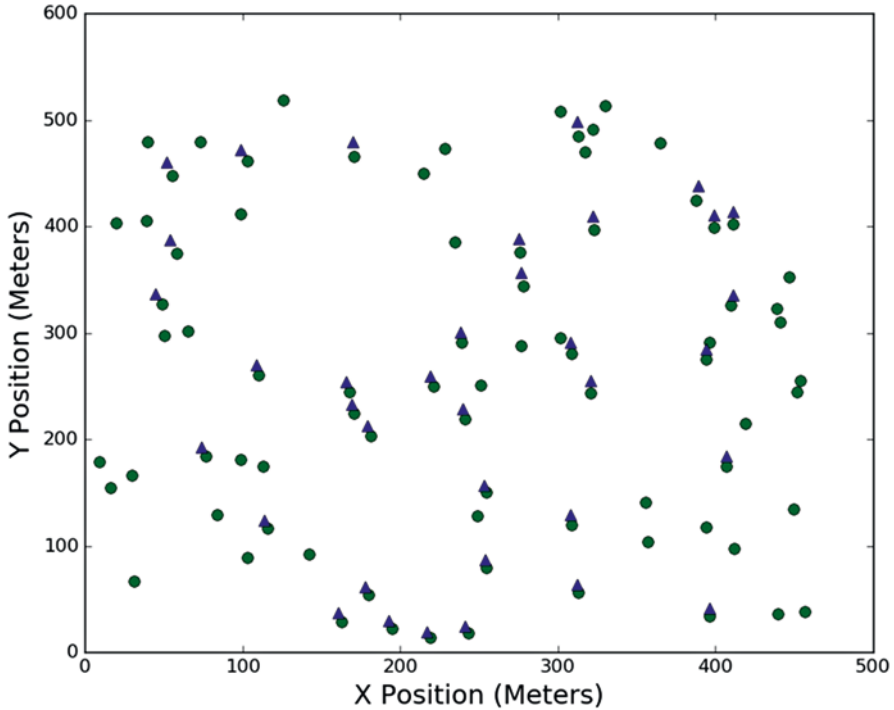


Fig. 3.4 Results of the stereo camera localization

### Algorithm 3.1 Constrained Cluster Radius (CCR) K-Means Clustering

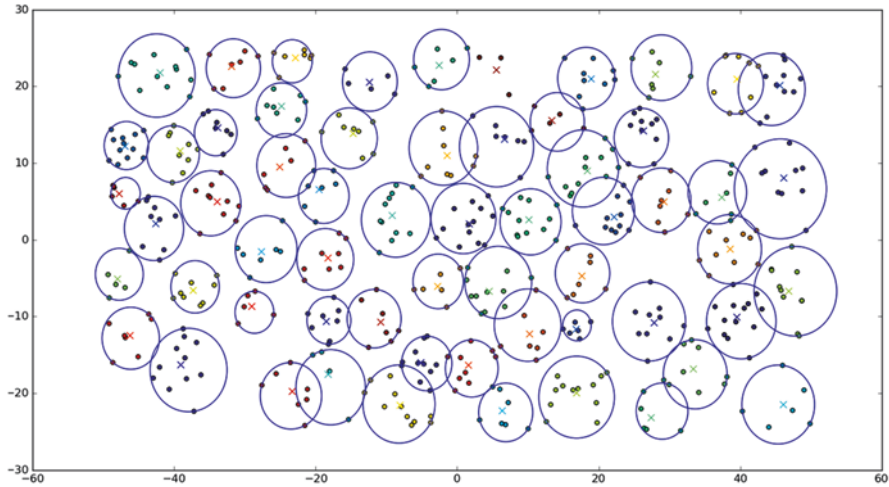
---

#### Algorithm 1 Constrained Cluster Radius K-Means Clustering

---

- 1:  $k \leftarrow 1$
  - 2: *Max Cluster Radius*  $\leftarrow \text{inf}$
  - 3: *Desired Max Cluster Radius*  $\leftarrow 10 \text{ meters}$
  - 4: **while** *Max Cluster Radius* > *Desired Max Cluster Radius*
  - 5: Select  $k$  points as initial centroids
  - 6: **repeat:**
  - 7: Form  $k$  clusters by assigning each person to its closest centroid.
  - 8: Recompute the centroid of each cluster
  - 9: **until** Centroids do not change.
  - 10:  $k+ = 1$
- 

To test the CCR K-Means Clustering algorithm, 500 locations of victims, represented by the smaller randomly colored dot, were randomly assigned. Using these locations, and a desired maximum cluster radius of 10 m, the clustering algorithm was performed. The results seen in Fig. 3.5 show that 60 m, represented by the larger



**Fig. 3.5** Results of the constrained cluster radius K-Means Clustering

blue circles, were created all with a radius of less than 10 m. The center of mass of the clusters are also represented by small Xs. Although this algorithm worked well, the runtime of the algorithm can be very slow with large numbers of people.

### 3.2.4.2 Parallelization of Clustering Algorithm

Since the CCR K-Means Clustering algorithm ran very slowly with large numbers of people, the algorithm was parallelized to shorten the run time. To do this, the Message Passing Interface (MPI) library (Forum 1994) was used. To parallelize the clustering, two algorithms were implemented. In the first algorithm, seen in Algorithm 3.2, the master process first sends a flag to each of the slave processes.

**Algorithm 3.2 Method 1, parallelized CCR K-Means Clustering****Algorithm 2 Method 1, Parallelized CCR K-Means Clustering**


---

```

1: SMCR (Slave Max Cluster Radius)  $\leftarrow 0$ 
2: finished flag  $\leftarrow False$ 
3:  $k \leftarrow 1$ 
4: MCR (Max Cluster Radius)  $\leftarrow \text{inf}$ 
5: DMCR (Desired Max Cluster Radius)  $\leftarrow 10 \text{ meters}$ 
6: if Master then
7:   while  $MCR > DMCR$  do
8:     for Each Slave Process do
9:       Send finished flag to slave process
10:      Send  $k$  to slave process
11:       $k+ = 1$ 
12:     for Each Slave Process do
13:       Receive SMCR from slave
14:       if  $SMCR > MCR$  then
15:          $MCR = SMCR$ 
16:   finished flag = True
17:   for Each Slave Process do
18:     Send finished flag to slave process
19:     Send  $-1$  to slave process
20: if Slave then
21:   while finished flag == False do
22:     Receive finished flag from master
23:     Receive  $k$  from master
24:     Perform clustering algorithm (Algorithm 3.1)
25:     Calculate SMCR
26:     Send SMCR to master

```

---

This flag is used to tell the slaves if the clustering algorithm is complete. If the flag is False, the clustering algorithm has not finished, but if it is True, the algorithm is complete. Next, the master sends the number of clusters to the slaves for them to use as the input to the k-means clustering algorithm. The master will send incrementally larger values to the slaves to use as the input until the desired max cluster radius is achieved. For example, if a system has one master and three slaves, the master will send (*False*, 1) (*False*, 2) (*False*, 3) to slaves one, two, and three, respectively. Once each slave process receives both the flag and the number of clusters to use as the input, the slaves will perform the CCR K-Means Clustering algorithm, and send back to the master the radius of the largest cluster. If the master receives a maximum cluster radius less than the desired max cluster radius, the master will set the flag to True and send it to all of the slaves. If the master does not receive a maximum cluster radius less than the desired maximum cluster radius, the master will increment the number of clusters needed, and continue the process. The second parallelized clustering algorithm, seen in Algorithm 3.3, was developed after noticing inefficiencies present in the first algorithm.

**Algorithm 3.3 Method 2, parallelized CCR K-Means Clustering****Algorithm 3 Method 2, Parallelized CCR K-Means Clustering**


---

```

1:  $k \leftarrow 1$ 
2:  $SMCR$  (Slave Max Cluster Radius)  $\leftarrow 0$ 
3:  $DMCR$  (Desired Max Cluster Radius)  $\leftarrow 10$  meters
4: if Master then
5:   Broadcast peoples locations to slaves
6:   Wait until  $k$  is received from a slave
7:   Perform clustering algorithm (Algorithm 3.1)
8: if Slave then
9:    $k \leftarrow MPI$  Rank
10:  while 1 do
11:    Perform clustering algorithm (Algorithm 3.1)
12:    Calculate  $SMCR$ 
13:    if  $SMCR < DMCR$  then
14:      Send  $k$  to master
15:      break
16:    Increment  $k$  by number of slaves

```

---

In the second algorithm, seen in Algorithm 3.3, the master initially sends the locations of the localized people to all the slaves. The master then waits until it receives a message from one of the slaves. Once the slaves receive the localized people locations, they begin calculating the clusters using an initial input value equal to their MPI rank. After each test, the slave will increment their input value by the number of slaves. Once one of the slaves calculates an output in which all of the clusters are less than a given desired radius, the slave sends this input value back to the master. This value speeds up the algorithm because only two messages are sent, which are the location of the people and the final number of clusters.

To test the effectiveness of the two Parallelized CCR K-Means Clustering algorithms, the algorithms were performed on two different machines, with varying numbers of people. The specifications of the two systems can be seen in Table 3.2.

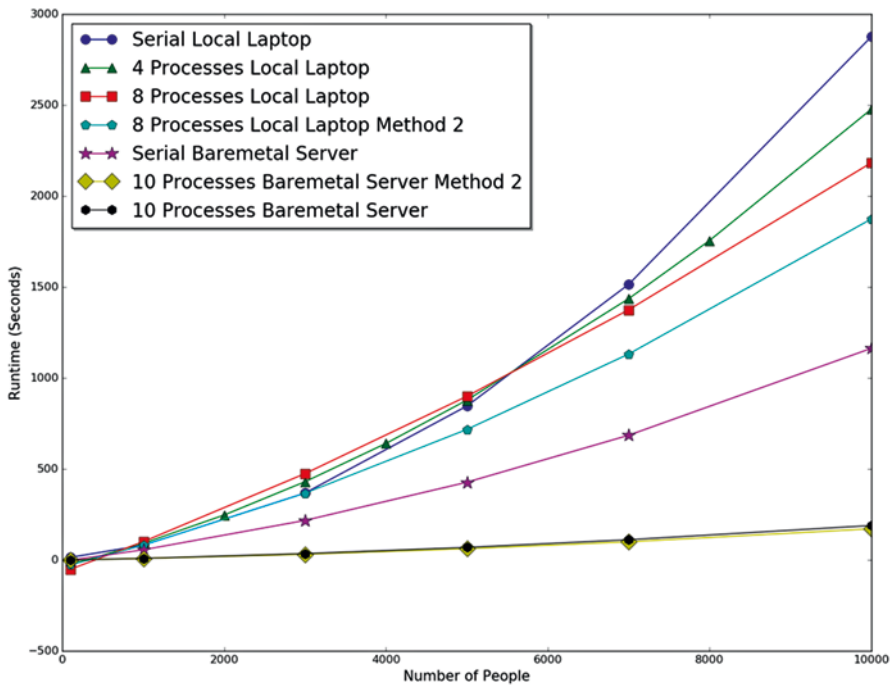
The run time of the different simulations can be seen in Fig. 3.6. The results show that parallelization improves the calculation time proportional to the number of people to be rescued. As the number of people increase, the speed-up due to parallelizing the algorithm increases.

**3.2.4.3 Meta-Clustering the Clusters**

To decide which ASV is responsible for which clusters, the cloud once again performs k-means clustering. This time, the traditional k-means clustering algorithm (Tan et al. 2005) is used, where the number of meta-clusters is chosen to be the same as the number of ASVs. To test the meta-clustering algorithm, the locations of 500

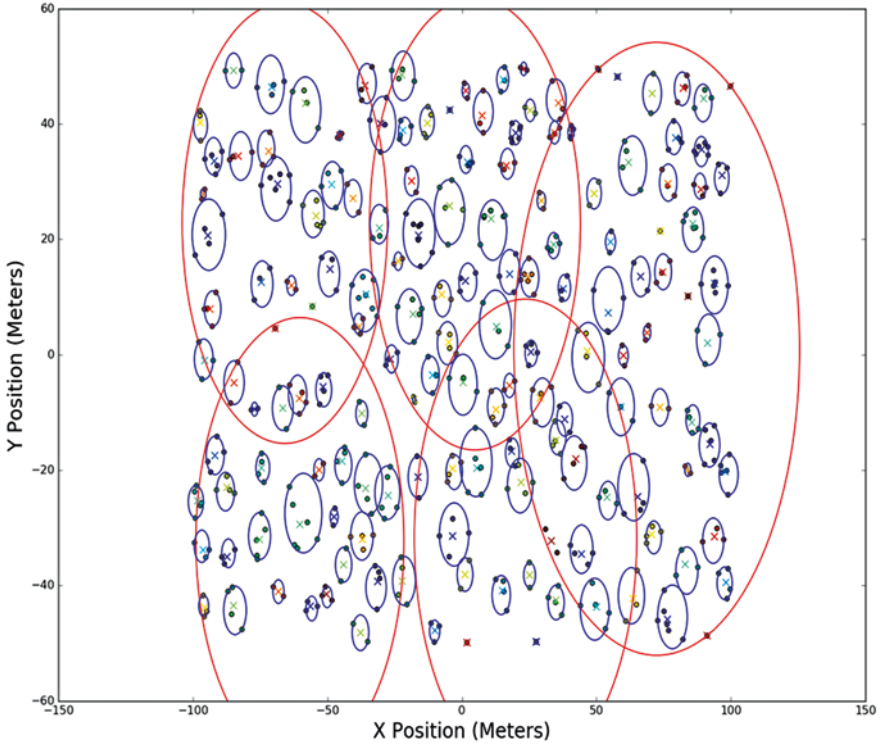
**Table 3.2** Specifications of machines

	Baremetal server	Local laptop
CPU model	Intel Xeon E5-2670	Intel Core i5-4202Y
CPU speed	3.0 GHz	1.6 GHz
Number of cores	48	4
Number of threads	96	8
Number of CPUs	2	1
RAM	128 GB	4 GB
RAM type	DDR4	DDR3L



**Fig. 3.6** Runtime of clustering algorithms of two different machines

people, represented by the smaller randomly colored dots, were randomly assigned. The parallelized CCR K-Means Clustering algorithm was performed using these locations, with a desired maximum cluster radius of 10 m. After the CCR K-Means Clustering algorithm was run, the meta-clustering algorithm was performed using five ASVs. The results of the meta-clustering can be seen in Fig. 3.7. The algorithm produced 60 clusters, represented by smaller circles, all with a radius less than 10



**Fig. 3.7** Results of the K-Means Clustering for the meta-clusters

m, and 5 meta-clusters, represented by the larger circles. Now that the original complex map of people locations has been organized into meta-clusters, these meta-clusters can be auctioned off to the ASVs.

### 3.2.5 Auctioning the Meta-Clusters

To determine which ASV is responsible for which meta-cluster, the cloud performs an auctioning algorithm. The auctioning algorithm uses the ASVs' locations, capacities, and speed along with the meta-clusters' locations, the number of people in the meta-clusters, and the location of the large mothership as inputs to the algorithm, seen in Algorithm 3.4.

**Algorithm 3.4 Auctioning algorithm for meta-clusters****Algorithm 4** Auctioning Algorithm for Meta-Clusters

---

```

1: Sort Meta-Clusters by Number of People (Max to Min)
2: Trips Needed  $\leftarrow$  0
3: Initial Distance  $\leftarrow$  0 meters
4: Mothership Distance  $\leftarrow$  0 meters
5: for Each Meta-Cluster do
6:   for Each ASV do
7:     Trips Needed = ceiling  $\left( \frac{\text{People in Meta-Cluster}}{\text{Capacity of Boat}} \right)$ 
8:     ASV.x  $\leftarrow$  ASV X Position
9:     ASV.y  $\leftarrow$  ASV Y Position
10:    MS.x  $\leftarrow$  Mothership X Position
11:    MS.y  $\leftarrow$  Mothership Y Position
12:    MC.x  $\leftarrow$  Meta-Cluster X Position
13:    MC.y  $\leftarrow$  Meta-Cluster Y Position
14:    ID  $\leftarrow$  Initial Distance
15:    MSD  $\leftarrow$  Mothership Distance
16:    ID =  $\sqrt{(\text{ASV.x} - \text{MC.x})^2 + (\text{ASV.y} - \text{MC.y})^2}$ 
17:    MSD =  $\sqrt{(\text{MS.x} - \text{MC.x})^2 + (\text{MS.y} - \text{MC.y})^2}$ 
18:    Score[ASV] =  $(\text{Trips Needed} * \frac{\text{MD}}{\text{ASV Speed}}) + \frac{\text{ID}}{\text{ASV Speed}}$ 
19:    ASV assigned to Meta-Cluster  $\leftarrow$  ASV W/ min(Score)
20:    Remove Assigned ASV from Auctioning Algorithm

```

---

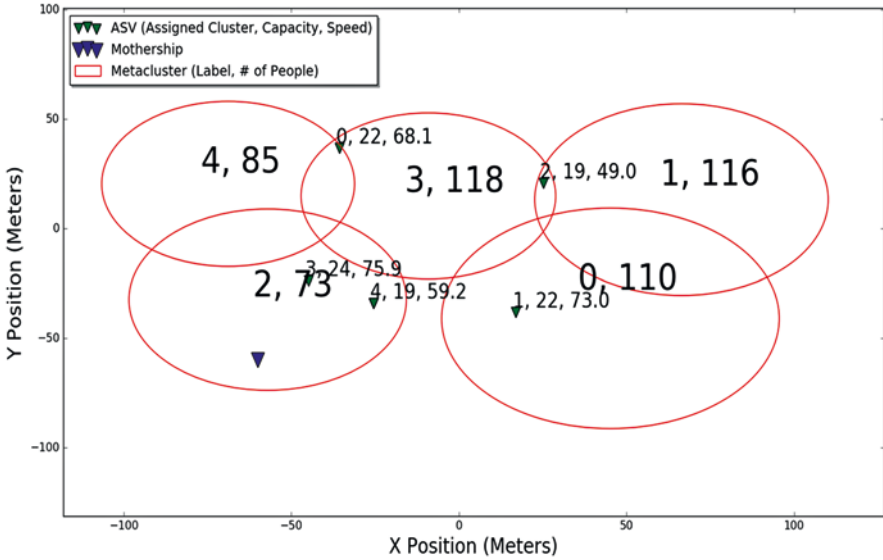
To test the auctioning algorithm, 500 peoples' locations were randomly assigned. Using these locations, and a desired maximum cluster radius of 10 m, the clustering and meta-clustering algorithms were performed. After the clustering algorithms were run, the auction algorithm was performed using five ASVs with randomly assigned capacities and normalized speeds. The results of the auction algorithm can be seen in Fig. 3.8. That figure shows which ASV is assigned to which meta-cluster. The ASVs are represented by smaller triangles, with their designated cluster, capacity and normalized speed in the upper right corner. The mother-ship is represented by the larger triangle, and the meta-clusters are represented by the larger circles. Inside each meta-cluster is the label for the meta-cluster along with the number of people inside. Now that each ASV has been assigned a meta-cluster, the ASVs can begin to travel to the clusters inside their assigned meta-clusters, and begin to rescue people.

### 3.2.6 Traveling to the Assigned Clusters

#### 3.2.6.1 Traveling Salesman Solver

To optimize the path the ASVs take to travel to the clusters inside their assigned meta-clusters, each ASV will use the nearest neighbor traveling salesman algorithm, seen in Algorithm 3.5, where the starting vertex is its current location. The





**Fig. 3.8** Results of the auction algorithm. Each of the ASVs have three numbers representing the assigned cluster, capacity and speed, respectively. The meta-clusters have two numbers representing its label, and the number of people in the meta-cluster, respectively

algorithm uses the nearest neighbor algorithm. The cost function is modified to include the number of people in each cluster and the cluster location. Finally, after the traveling salesman algorithm is completed, the ASVs can travel to each cluster and begin to save people floating in the ocean.

**Algorithm 3.5 Nearest neighbor traveling salesman algorithm**

**Algorithm 5 Nearest Neighbor Traveling Salesman Algorithm**

- 1: *Starting Vertex*  $\leftarrow$  *ASV Current Location*
- 2: **while** All vertices have not been visited **do**
- 3:     **for** Each Unvisited Cluster **do**
- 4:          $Cost = \frac{Distance\ from\ Vertex\ to\ Cluster}{Number\ of\ People\ in\ Cluster}$
- 5:     *Current Vertex*  $\leftarrow$  *Unvisited Cluster with Lowest Cost*
- 6:     *Mark Current Vertex as Visited*

To test the traveling salesman solver, 25 cluster centers, represented by the dots, were randomly generated. Using these locations, the nearest neighbor traveling salesman algorithm was used to find a short path between all the clusters. The results of the traveling salesman algorithm can be seen in Fig. 3.9, and it shows that the generated path between all the clusters is the optimal path to take.

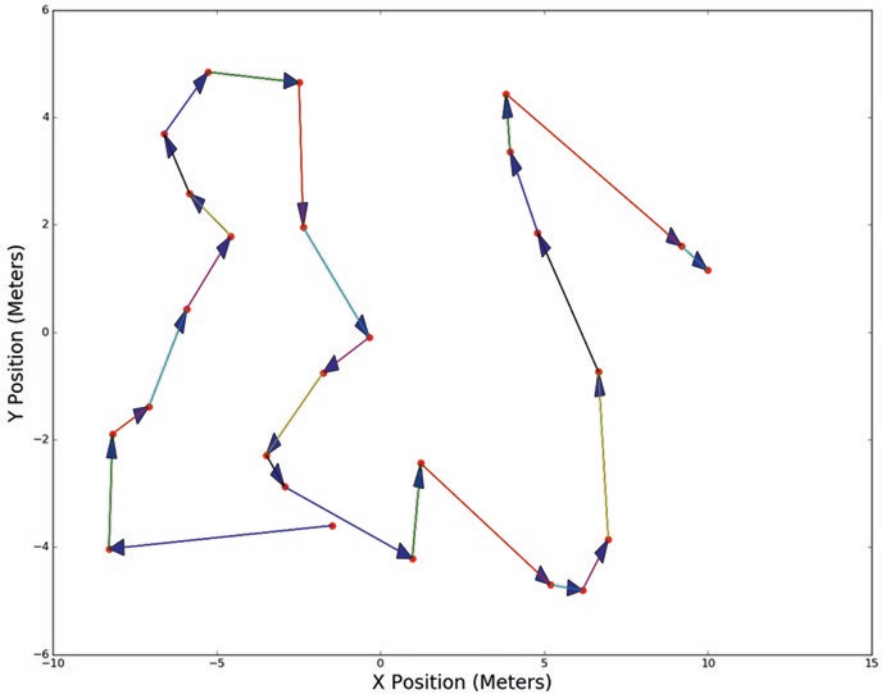
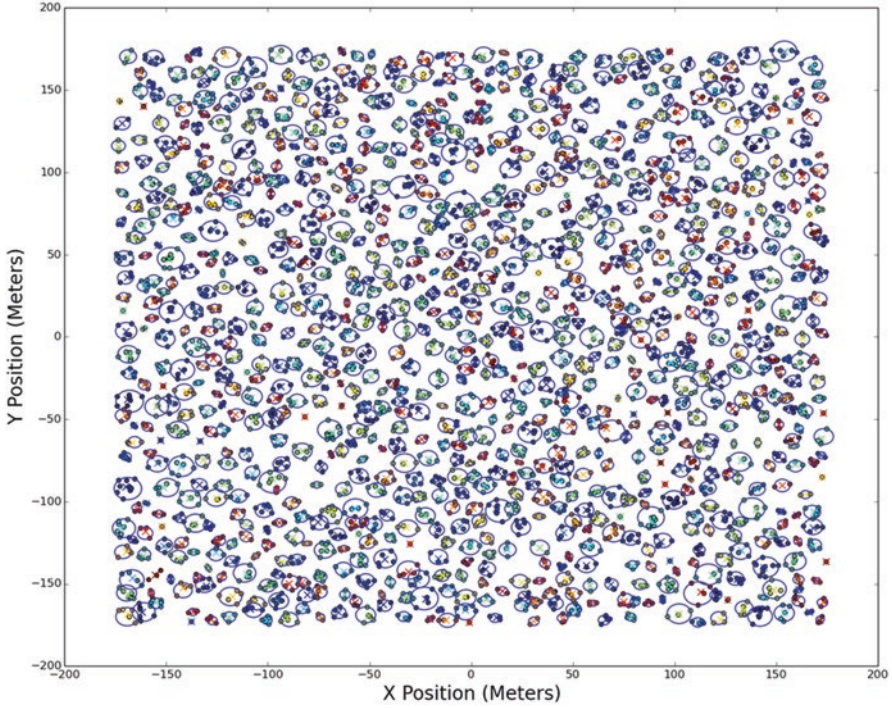


Fig. 3.9 Results of the nearest neighbor traveling salesman algorithm

### 3.2.6.2 Human Interaction with Swarm

Since the saving of people after a cruise ship crash is a very complex situation that could result in unexpected scenarios, such as someone not being detected by the MAVs, or someone needing immediate rescue, each ASV has an on-board human operator. This human operator can at any time take control of the boat. To do this, the human operator places the ASV into manual mode. To simulate this human-robot interaction, each ASV has a ROS topic responsible for determining if the ASV is in manual or autonomous mode. If at any time during the operation the human operator switches the ASV to manual mode, a message will be published to the ASV mode topic. This will cause the ASV to stop, and allow the human operator to gain control of the ASV. When the human operator switches the ASV back to autonomous mode, the ASV will rerun the traveling salesman algorithm and continue to travel to its assigned clusters.



**Fig. 3.10** Results of the CCR K-Means Clustering algorithm with 4000 people and a maximum cluster radius of 10 m, where the *small dots* represent people and the *small circles* represent clusters

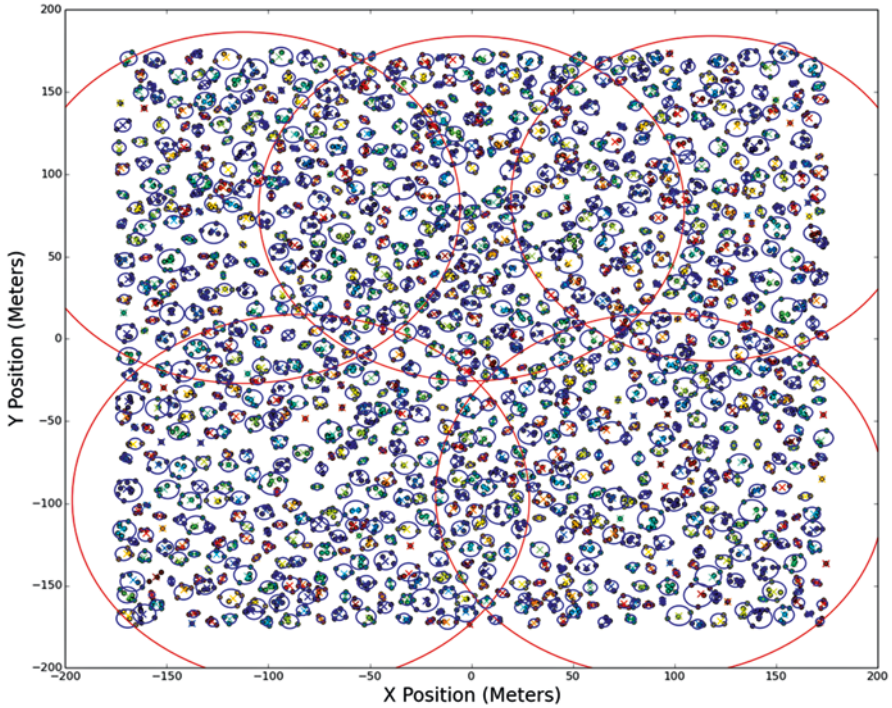
### 3.3 Experimental Results

#### 3.3.1 Simulation Results

After each part of the system was tested individually, the system components were combined and tested. The simulation randomly generated four thousand people to be rescued, as well as five ASVs with random capacities, maximum speeds, and locations. The entire system was tested together. The results of the CCR K-Means Clustering algorithm can be seen in Fig. 3.10. The CCR K-Means Clustering algorithm produced 1029 clusters all with a radius of less than 10 m.

After the CCR K-Means Clustering algorithm was run, the meta-clustering algorithm was performed. As seen in Fig. 3.11, the algorithm produced five meta-clusters, since there are five ASVs.

Lastly, the auction algorithm was performed. The results of the auction algorithm, shown in Fig. 3.12, shows that each ASV was assigned a meta-cluster. The ASVs could then use the traveling salesman solver to travel to the individual clusters inside of their assigned meta-cluster.



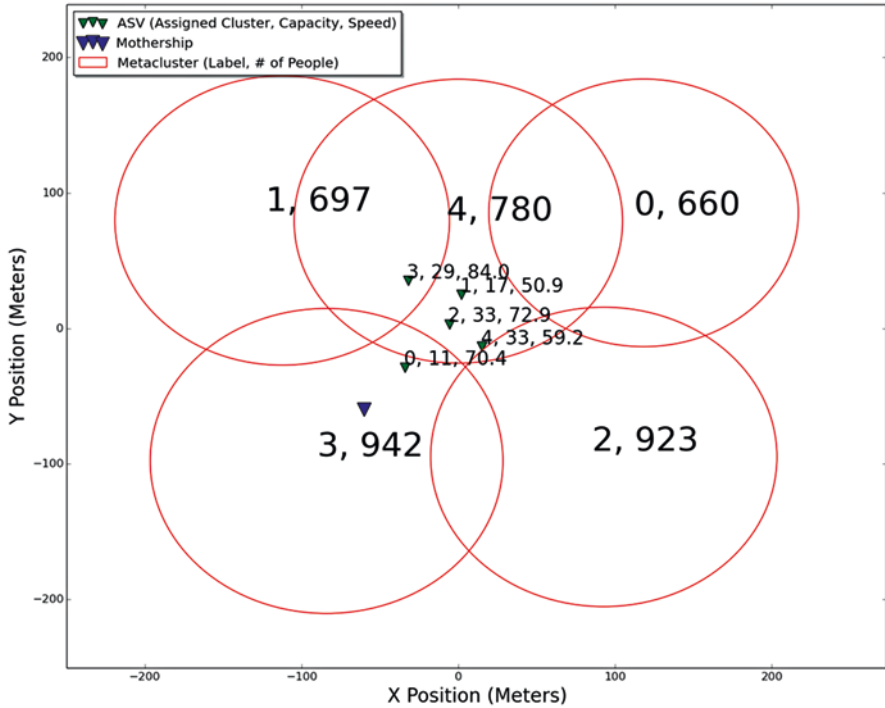
**Fig. 3.11** Results of meta-clustering algorithm with 4000 victims, a maximum cluster radius of 10 m and 5 ASVs, where the *small dots* represent people, the *small circles* represent clusters, and the *large circles* represent the meta-clusters

### 3.3.2 Hardware Emulation Results

Now that the system has been tested in simulation, the system needs to be tested using hardware. Due to hardware limitations, unmanned ground vehicles (UGVs) were used to emulate the ASVs, and an overhead camera was used to emulate GPS. Once again, each part of the system was tested separately and then the whole system was tested.

#### 3.3.2.1 Unmanned Ground Vehicle (UGV)

To emulate the ASVs, specialized UGVs were designed and created. As a base, the Adafruit Raspberry Pi Robot (Adafruit 2016) was used because of its low cost, small size, and processing power. The robot has two DC motors, a swivel caster, a Raspberry Pi 3 (which has built in Wi-Fi), and an Adafruit DC & Stepper Motor Hat. To power the robot, a 14,000 mAh battery bank was added. This battery bank has two smart USB outputs for a total rating of 5V/3.5A. These two outputs were

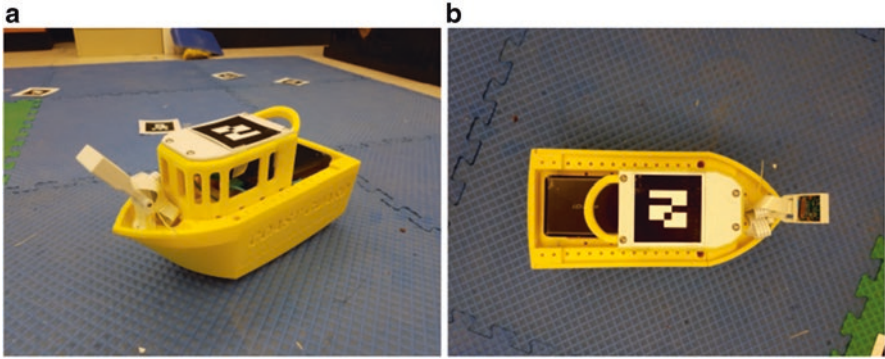


**Fig. 3.12** Results of auction algorithm with 4000 people to be rescued, a maximum cluster radius of 10 m and 5 ASVs. The *small triangles* represent the ASVs, the *larger triangle* represents the mothership, and the *large circles* represents the meta-clusters. Each of the ASVs has three numbers representing the assigned meta-clusters, boat capacity, and boat speed. Each meta-cluster has two numbers representing the meta-cluster’s label and the number of people in each of the meta-clusters

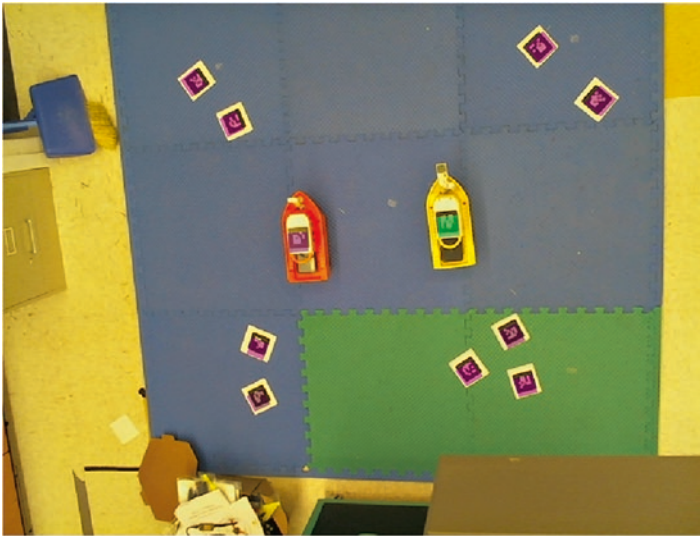
used to power the Raspberry Pi 3 and the DC motors, separately. A Bosch BNO055 (MEMS accelerometer, magnetometer and gyroscope) was also added to each UGV, but was not used for any of the experiments. Lastly, a 3D printed boat hull was added to the UGV to make it look like an actual ASV, as seen in Fig. 3.13a.

### 3.3.2.2 GPS Emulation

Since these experiments were performed indoors, the GPS needed to be emulated. To do this a ROS packaged called `ar_track_alvar` was used (ROS.org 2016). `Ar_track_alvar` is a ROS wrapper for the open source AR tag tracking library, Alvar. To use the package each UGV had an AR tag on top of it, as seen in Fig. 3.13b. An overhead web-cam could then be used to track the UGV. `Ar_track_alvar` provides an  $x, y, z$  position of each AR tag along with its orientation in quaternions. As seen in



**Fig. 3.13** 3D Printed UGV used to emulate ASV. (a) Ground view. (b) Aerial view

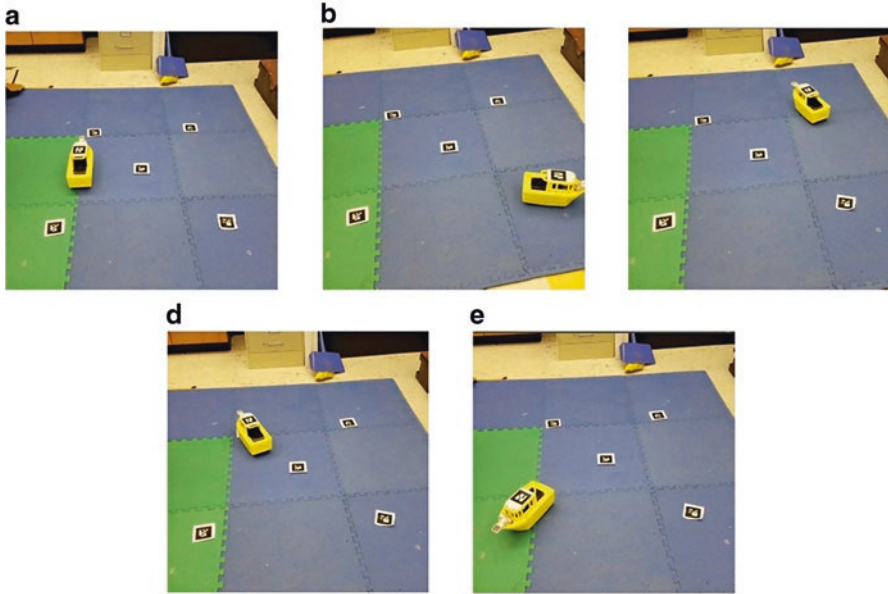


**Fig. 3.14** Tags being tracked by ar\_track\_alvar

Fig. 3.14, each tag that is being detected and tracked by ar\_track\_alvar has a purple or green square on top of it.

### 3.3.2.3 Traveling Salesman

After the UGVs and GPS emulation were determined to be operating correctly, the traveling salesman algorithm could be tested. To do this, five AR tags, to represent cluster centers, and one UGV were placed in the testing area. Ar\_track\_alvar was then used to detect the locations of the AR tags and the UGV. Once all the tags and UGV were detected, the traveling salesman algorithm was performed, and the UGV could



**Fig. 3.15** Hardware results of the traveling salesman algorithm. (a) Starting location. (b) ASV reached first tag. (c) ASV reached second tag. (d) ASV reached third tag. (e) ASV reached final tag. Video can be seen at <https://youtu.be/r9QlpZKSikY>

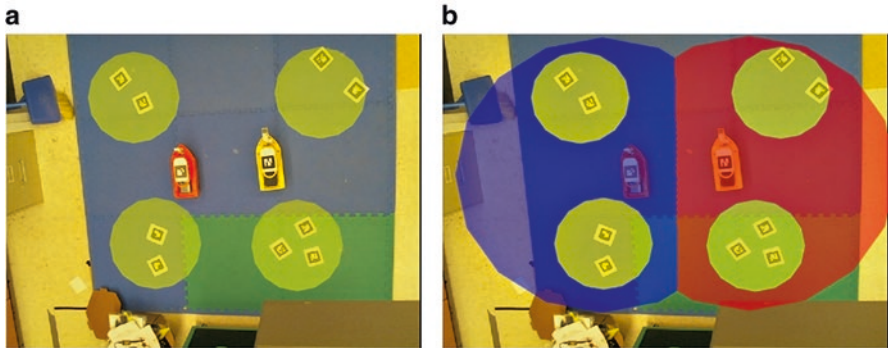
travel to each tag in an efficient manner. The results of the traveling salesman algorithm can be seen in Fig. 3.15, which shows that the UGV was successfully able to use the traveling salesman algorithm to traverse to each tag in an efficient manner.

### 3.3.2.4 CCR K-Means Clustering

After the traveling salesman algorithm was tested, the CCR K-Means Clustering algorithm needed to be tested. To test the CCR k-means clustering algorithm, ten AR tags were placed into five different clusters of varying sizes. Once again, ar\_track\_alvar was used to detect the tags. Once all the tags were detected, the CCR K-Means Clustering algorithm was performed. The CCR K-Means Clustering algorithm generated five clusters as expected, which can be seen as the circles in Fig. 3.16a.

### 3.3.2.5 Meta-Clustering

Once the CCR K-Means Clustering algorithm was successfully tested, the meta-clustering could be tested. To test the meta-clustering algorithm, the same setup was used from before, along with two ASVs. As seen in Fig. 3.16b, the meta-clustering



**Fig. 3.16** Clustering algorithms performed on hardware. (a) CCR K-Means Clustering Algorithm. (b) Meta-clustering Algorithm

algorithm generated two meta-clusters, represented by the large semi-circles, each with two clusters inside of them.

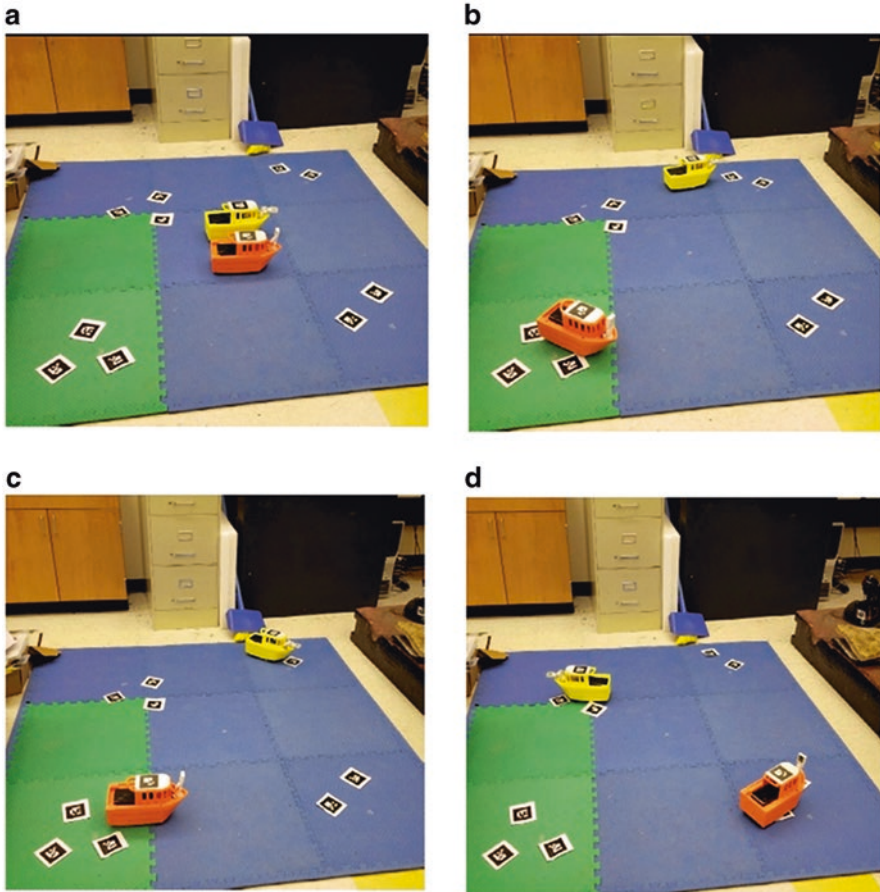
### 3.3.2.6 Auction Algorithm

Now that the meta-clustering algorithm has been tested, the auction algorithm was implemented, completing the system. Using the same setup as before, the auctioning algorithm was added to the system. To test the auctioning algorithm, the entire system was tested. The results of the test, seen in Fig. 3.17, shows that the entire system works and that the ASVs can travel to their assigned clusters after the auctioning algorithm is performed.

## 3.4 Conclusions

This paper presents a heterogeneous centralized robotic swarm to rescue human survivors after a shipwreck consisting of ASVs and MAVs. The presented system shows how a heterogeneous robotic swarm can work cooperatively with a cloud-based network. The system also shows how the symbiotic relationship between ASVs and MAVs can be used to leverage the advantages of each system to save people floating on the surface after a cruise-ship disaster. In the future, various parts of the system will be improved to increase the robustness of our algorithm. For example, the algorithm used by the MAVs to search for the victims is over simplified. This algorithm can be improved to increase the search area coverage and decrease the chances of missing a person. In the future, we also plan to develop a training-dataset and convolutional neural network to detect people floating in the ocean. These developments would allow the system to detect floating people in the ocean rather than using color thresholding, as was done in this work. We also plan





**Fig. 3.17** Hardware results of the entire completed system. (a) Starting locations. (b) First ASV reached first cluster. (c) Second ASV reached first cluster. (d) Both ASVs reached second cluster. Video can be seen at <https://youtu.be/9oAiLI6xr8>

to use an actual UAV to perform the detection and localization of the victims, along with adding autonomous underwater vehicles and robotic fish to the swarm to detect people under the surface of the water. We have demonstrated that a swarm of robots can be used to save humans involved in a shipwreck caused by human error (as with the Costa Concordia).

**Acknowledgements** This work was supported by Grant number FA8750-15-2-0116 from Air Force Research Laboratory and OSD through a contract with North Carolina Agricultural and Technical State University.

## References

- Lwowski J, Sun L, and Pack D (2016) Heterogeneous bi-directional cooperative unmanned vehicles for obstacle avoidance. In: Proceedings of the ASME 2016 Dynamic Systems and Control Conference
- Ray A, Benavidez P, and Jamshidi M (2009) Decentralized motion coordination for a formation of rovers. *IEEE Systems Journal* 3:369-381
- Gallardo N, Pai K, Erol B, Benavidez P, and Jamshidi M (2016) Formation control implementation using kobuki turtlebots and parrot bebop drone. In: 2016 World Automation Congress, pp. 1-6
- Gerkey B P, and Mataric M J (2004) A formal analysis and taxonomy of task allocation in multi-robot systems. *The Intl. J. of Robotics Research* 9:939-954
- Shia A (2011) Survey of swarm robotics techniques a tutorial <http://ieeetmc.net/r6/scv/ras/swarm-survey.pdf>
- Liu C, Kroll A (2012) A Centralized Multi-Robot Task Allocation for Industrial Plant Inspection by Using A\* and Genetic Algorithms. Berlin, Heidelberg: Springer Berlin Heidelberg 1:466-474
- Giordani S, Lujak M, and Martinelli F (2014) A distributed algorithm for the multi-robot task allocation problem. In: International Conference on Ambient Systems, Networks and Technologies
- Kuhn H W (1955) The hungarian method for the assignment problem. *Naval Research Logistics Quarterly* 2:83-97
- Sujit P B, and Sousa J B (2012) Multi-uav task allocation with communication faults. In: 2012 American Control Conference (ACC), pp. 3724-3729
- Dutta R, Sun L, and Pack D (2016) Multi-agent formation control with maintaining and controlling network connectivity. In: 2016 American Control Conference (ACC), pp. 1036-1041
- BBC News (2015) Coasta Concordia: What happened. <http://www.bbc.com/news/world-europe-16563562>
- France S (2012) Harmony of the seas oasis 4. <http://www.stxfrance.com/UK/stxfrance-reference-41-HARMONY%20OF%20THE%20SEAS%20%20OASIS%204.awp>
- Cunard (2002) Queen mary 2 fact sheet. <http://www.cunard.com/Documents/Press%20Kits/USA/Queen%20Mary%202/QM2%20Fact%20Sheet.pdf>
- US Coast Guard (2007) 47ft motor lifeboat operators handbook. [https://www.uscg.mil/directives/cim/16000-16999/CIM\\_16114\\_25B.pdf](https://www.uscg.mil/directives/cim/16000-16999/CIM_16114_25B.pdf)
- US Coast Guard (2006) Defender class operators handbook. [https://www.uscg.mil/directives/cim/16000-16999/CIM\\_16114\\_28.pdf](https://www.uscg.mil/directives/cim/16000-16999/CIM_16114_28.pdf)
- Quigley M, Conley K, Gerkey B, Faust J, Foote T, Leibs J, Wheeler R, and Ng A Y (2009) Ros: an open-source robot operating system. In: ICRA workshop on open source software
- Ma Y, Soatto S, Kosecka J, and Sastry S S (2003) An Invitation to 3-D Vision: From Images to Geometric Models. SpringerVerlag
- Koenig N, and Howard A (2004) Design and use paradigms for gazebo, an open-source multi-robot simulator. In: 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE Cat. No.04CH37566), pp. 2149-2154
- Kam H R, Lee S H, Park T, and Kim C H (2015) Rviz: a toolkit for real domain data visualization. *Telecommunication Systems* 60:337-345
- Forum M P (1994) Mpi: A message-passing interface standard. Tech. Rep.
- Tan P N, Steinbach M, Kumar V (2005) Introduction to Data Mining, First Edition. Addison-Wesley Longman Publishing Co., Inc.
- Adafruit (2016) Simple raspberry pi robot. <https://learn.adafruit.com/simple-raspberry-pi-robot>
- ROS.org (2016) ar\_track\_alvar. [http://wiki.ros.org/ar\\_track\\_alvar](http://wiki.ros.org/ar_track_alvar)

# Chapter 4

## Human Information Interaction, Artificial Intelligence, and Errors

Stephen Russell, Ira S. Moskowitz, and Adrienne Raglin

### 4.1 Introduction

Humans' interaction with information will only increase in the future and this interaction will be facilitated by artificial intelligent proxies. Because opportunities for errors most often occur at the intersections of system components, human or otherwise, the adoption of artificial intelligence (AI) mechanisms will assuredly increase the amount of error that occurs in information systems. The nature of these errors will likely manifest as latent errors and therefore be difficult to identify and resolve. Additional research in human information interaction (HII) is necessary, and can positively improve the development of AI innovations. By furthering our understanding of humans' interaction with information objects, HII research can provide advances in both the human, and machine, domains. Insights from this research are necessary to understand errors that result from the actions of humans and artificial intelligence.

Often confused with human computer interaction (HCI) and human system interaction (HSI), human information interaction has a similar but distinctly different nuance from those other fields of study. HCI is a discipline concerned with the design, evaluation, and implementation of interactive computing systems for human use and with the study of major phenomena surrounding them (ACM SIGCHI 1992). Similarly, HSI is defined as end-user or customer interaction with technology-based systems through interaction styles or modalities such as reading, writing, touch, and sound (Chang and Bennamoun 2012). Given these definitions, it is clear to see that the emphasis of both HCI and HSI is not on information, even though

---

S. Russell (✉) • A. Raglin  
U.S. Army Research Laboratory, 2800 Powder Mill Road, Adelphi, MD, USA  
e-mail: [stephen.m.russell8.civ@mail.mil](mailto:stephen.m.russell8.civ@mail.mil); [adrienne.j.raglin.civ@mail.mil](mailto:adrienne.j.raglin.civ@mail.mil)

I.S. Moskowitz  
Naval Research Laboratory, 4555 Overlook Avenue SW, Washington, DC, USA  
e-mail: [ira.moskowitz@nrl.navy.mil](mailto:ira.moskowitz@nrl.navy.mil)

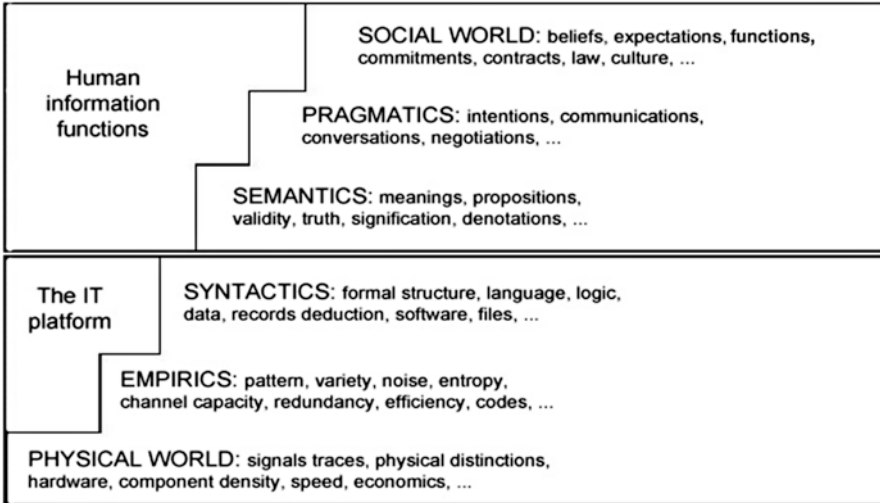


Fig. 4.1 Semiotic framework (Liu 2000)

those phenomena form a foundational basis for both computers and systems. Information is the nuanced difference between HII and HCI/HSI. HII is the field of study that is concerned with how and why humans use, find, consume, and work with information in order to solve problems, make decisions, learn, plan, make sense, discover, and carry out other tasks and activities (Sedig and Parsons 2015). It might be argued that HCI and HSI are supersets of HII, but the focus on humans' interaction specifically with information, as opposed to the computer platform, interfaces, and surrounding processes, marks a significant differentiation that is particularly critical for information systems. Information systems deal with symbolic or information-centric representations of reality. Figure 4.1 shows Liu's semiotic framework (Liu 2000) that is synonymous with an information system.

Explicitly missing from Liu's framework, but implied in "the IT platform," is the processing layer that must exist to map from the information technology to the human. Within an information system, it is this processing layer where AI finds its contributed value added. The implementations of AI are characterized by symbolic processing, non-deterministic computations, and knowledge management. Subsequently, innovations in AI are moderated by the advances in HII that directly impact the interdependence existing between humans and the AI-enabled information systems supporting them. When inconsistencies in that fundamental balance occur, errors may be generated. Nowhere is this balance more critical than in information processing environments.

The amount of complexity in the use of information has surpassed the intersection of simple computation and human's need for analytics. This has resulted in the emergent HII field of study; examining autonomous and computationally-aided problem solving within activity-contexts. The complexity of HII demands interoperability and compatibility between mixed initiative processes for information

acquisition and processing in context to aid comprehension by humans' use of information. AI innovations are one of the primary means to automate and aid interaction with information.

This chapter presents a contemporary overview of HII and discusses the need for research in this field of study that necessarily investigates the implications of AI and human error. It provides a background on HII, considers artificial intelligence and information processing, analyzes how the convergence of HII research and AI will require new notions of errors, and finally identifies potential research areas that are important to advancing human information interaction and artificial intelligence for error mitigation.

## 4.2 Human Information Interaction

The general trend towards pervasive computing will naturally result in less focus on computing devices and the boundary between them and more on humans' access to the benefits they provide. Consider how people think of their desktop or laptop computers and contrast this view with tablets and cellphones. The mobile devices are still computers providing much of the same functionality as the desktop, just more portable. When this contrast is thought of in the context of cloud computing, the diminishing emphasis on computing devices and increasing spotlight on information or information objects becomes readily apparent. Further blurring of the device-information distinction will only continue, as the pervasiveness of computing and technology continue to dissolve the barriers between information and the physical world (Limkar and Jha 2016; Barnaghi et al. 2015; Bolotin 2015; Wiberg 2015).

The commercialization of the Internet of Things marks the marketization of the focal transition away from human computer/machine/system interaction to humans' information-centric interactions (Soldatos et al. 2015). This perspective makes sense, because the world is an integral whole in which the things that exist in it are interdependent; not a collection of distinct elements isolated from each other (Fidel 2012). Moreover, information is arguably ubiquitous now and will only become more so in the future as the commercialization of Internet "things" continues to cross boundaries of business, physics, biology and other fields of science. Thus, the convergence of fields of study, already interdisciplinary in nature, mandates a similar concentration on topics of human information interaction.

Gershon (1995) was the first to establish the phrase "human information interaction" when examining HCI research, differentiating the label as placing more focus on "how human beings interact with, relate to, and process information regardless of the medium connecting the two." Marchionini (2008) extends this notion to suggest that human information interaction (HII) shifts the foci of all aspects of information work; blurs boundaries between information objects, technology, and people; and creates new forms of information. This is a significant departure from

human computer (HCI) or human system (HSI) interaction, which considers technology more broadly and places equal emphasis on physical aspects of interaction.

When considering HII, it is important to have functional definitions for the terms: human, information, and interaction. From a definitive perspective human is the best understood... it is us: individuals and people. Relative to HII, humans are the individuals or people who interplay with information and its related environments. Often considered “users” of information systems (Dervin and Reinhard 2006), humans fulfill the role of “actors” when their scope of examination includes tasks (Lamb and Kling 2003; Mutch 2002). When the interdependence of the world is factored into the definition of human, it is important to think of the second order effects, where community of actors, teamed or seemingly operating independently, impact one another through their information-centric functions. Therefore, in this sense, humans include cooperative and non-cooperative individuals and teams; bound by the scope of information.

Within HII, both human and information must be considered as nouns and thus, information must be thought of as a “thing” of a physical nature. This nature is consistent with the bit-based form of information, as defined by Shannon (1948). Although a distinct and formal definition of information has been and remains the subject of extensive philosophical debate, when the physical definition of information is adopted anything that is experienced by humans (sight, sound, taste, smell, and feel) can be considered information. Within HII, this physical definition is extended to give information context within the human experience. As such, information can be thought of as a symbolic representation that has *meaning*, is *communicated*, has an *effect*, and is used for *decision making* (Buckland 1991). *Meaning* implies some degree of understanding. *Communicated* requires transmission (not necessarily receipt). *Effect* mandates acknowledgement in the minima and action in the maxima. *Decision making* signifies purpose, relative or not (Fidel 2012). These requirements for the definition of information give information-objects state within HII processes.

Interaction is the actionable (verb) part of HII, being defined as the activity that involves both parts, humans and information. The nature of interaction extends beyond the concept of interface, which is merely the doorway to true interaction. Given this distinction, interaction is the interplay between different components (humans and information in HII), rather than a fixed and pre-specified path (Moore 1989). This view of interaction is reasonable because there are degrees of interaction and humans can inject stochasticity into a process. Yet by omitting pre-specified paths, Moore’s definition is too restrictive to hold in HII because information systems, where humans and information interact, often follow pre-specified (via a-priori programming) paths. Dourish (2004) offers a more applicable definition of interaction as: the what and the how of something being done; the means of by which activity is accomplished, dynamically, and in context.

It is this notion of dynamic activity or work that has found grounding in the HII literature. It is important to note that within HII there is an implicit understanding that information already exists, does not need to be “created,” and that it is being transformed from one state to another for the purposes of human interaction and

comprehension. This orientation on work allows HII to apply notions of Shannon's (1948) information theory to second-order concepts such as uncertainty, context, causality, and reasoning. We note that a purely quantitative approach to information is far from satisfactory. The Small Message Criterion (see footnote 1) (Moskowitz and Kang 1994) shows the danger of relying solely on bit-counting measures of information leakage. As an example, consider the ride of Paul Revere. One bit of information was enough to tell the Colonialists that the British were coming (one if by road, two if by sea). Furthermore, in Moskowitz et al. (2002) the use of bit counting metrics of hidden images is also shown to be lacking due to the way the human mind interprets images, already noisy images. Allwein (2004) attempted to provide a qualitative framework for Shannon-type theories. This paper was the first to marry Barwise and Seligman (1997) approaches to Shannon's theories using the tools of channel theory from logic.

Despite some work applying Shannon's theories in logic and computational methods, applications of Shannon's information theory have found little traction, beyond an initial foray in the 1950s, within the psychology domain (Luce 2003). Two notable early works illustrate the application of Shannon's theories to human information interaction. McGill's (1954) "Multivariate Information Transmission" and Miller's (1956) "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information" sought to address phenomena that bounded information capacity limitations, including absolute judgments of stimuli and short-term memory (Luce 2003). The idea of the human mind being an information processing network with capacity limitations has remained as a concept in the literature (Marois and Ivanoff 2005; Rabinovich et al. 2015; Serruya 2015), but these works view the human mind and its processes in more complex ways than pure information theory as quantitatively defined by Shannon. British mathematician Devlin (2001, p. 21) points out the seeming inapplicability of Shannon's information theory to complex psychological concepts and research by minimizing the notion of information to simply data:

Shannon's theory does not deal with "information" as that word is generally understood. Instead, it deals with data—the raw material<sup>1</sup> out of which information is obtained.

The lack of confluence between information theory and psychology is readily apparent in Skilling's (1989) book on Entropy and Bayesian methods. Table 4.1 summarizes the book's table of contents. Noticeably missing are any topics involving human or cognitive applications. If information is something that the human mind commonly interacts with and Shannon's theory is the grounding for one side of that interaction, more occurrences of Shannon's theories should appear in the psychology literature.

Beyond the convergence of information theory and psychology (or lack thereof) lay the purpose of methods for improving human information interaction. Human information interaction would not be as significant an issue if there were not an

---

<sup>1</sup>The Small Message Criterion is a metric used in measurements of channel capacity that indicates a degree of trade-off between lower channel capacity and channel performance.

**Table 4.1** Summary of topics (adapted from Luce 2003)

Topic	No. articles
Statistical fundamentals	17
Physical measurement	6
Time series, power spectrum	6
Thermodynamics and quantum mechanics	5
Crystallography	5
Astronomical techniques	3
Neural networks	2

impedance mismatch between the amount of information available and the optimal execution of human processes and decision making. Underlying the reasoning behind research in HII is fundamentally addressing information overload. Further issues of the inverse, information underload also exist. Information underload occurs when we do not have access to the information needed to complete a process or decision. Overload occurs when access to information is available but we are simply overwhelmed by the amount of information available of which, not all is equally valuable or applicable (Alexander et al. 2016). One could conclude that this implies that HII is most applicable to complex information situations.

Simple information interactions are those where there is a single information element and a single path to the correct answer, where the result supports complete information that has a boolean (right/wrong) output-state accounting for all the factors that might influence the answer. Albers (2015) uses the example of “look it up on Google...” as an illustration of simple information. Complex information on the other hand is a significant distance away from simple information. Albers characterizes complex information as the necessary information when there is no “single” answer. The problem space of complex information is where information needs cannot be predefined and there exists one or more other complicating conditions: (1) there are multiple paths to an answer, often with varied levels of desirable output; (2) completeness is obfuscated or even unknown, all of the factors influencing the correct answer are not known; and (3) history or temporal considerations have bearing on the problem. It is in the area of complex information, where the HII challenge is the greatest, that context becomes a dominant factor.

Contextual awareness (as shown in Fig. 4.2) is defined as the understanding of how the information fits within the current situation; the understanding of the information relationships; and the understanding of the development of the situation in the future and related predictions about interdependent effects of any decision across the entire situation (Albers 2011). Abowd et al. (1999) define context as any information that can be used to characterize the situation of an entity, where an entity can be a person, place, or physical/virtual object. Given this definition, it is clear that context is information and information describes and contains context.



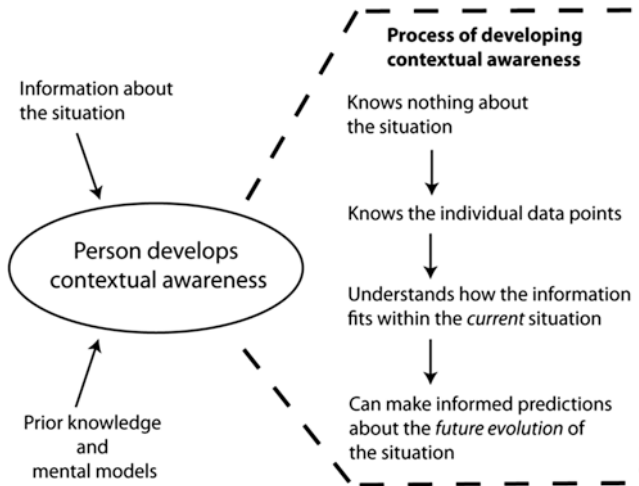


Fig. 4.2 Contextual awareness (Albers 2011)

Dey (2001) provides a definition of context as found in Abowd et al. (1999) and provides a functional example of two pieces of information—weather and the presence of other people—and use the definition to determine whether either one is context in an applications problem space. In Dey’s example, the weather does not affect the application because it is being used indoors. Therefore, according to Dey, *it is not context*. The presence of other people, however, can be used to characterize the application user’s situation and in Dey’s view *it is context*. In this example, the state of both might be viewed as simple information: weather is a factor (yes/no) and people present (yes/no). However, when these two pieces of simple information interact with an information-application’s user, they become complex information. The weather may be good or sufficiently bad such that it impacts the application, whether the user is indoors or out. Similarly, the presence of people may not have any bearing in characterizing the application user’s situation—consider people present but quietly otherwise occupied.

Interestingly, the word *context* has become a favorite in the vocabulary of cognitive psychologists (Clark and Carlson 1981). Clark and Carlson suggest that the term “context” is useful because it is sufficiently vague, general, and can accommodate many different ideas. This ambiguity is precisely why context can be associated with complex information. Other research (Henricksen et al. 2002) has even described the characteristics of context information with the same concepts as complex information: exhibiting a range of temporal characteristics; “imperfect,” i.e. may be correct or incorrect; having many alternative representations; and highly inter-related. The close relationship between complex information and context would imply that the extremes of information overload highly correlate with contextual awareness.

Because of the challenges in solving problems that involve complex information, most models of human computer interaction (HCI) do not completely express the required extensive backtracking and digressions involved in the information interaction portion of problem solving (Simon and Young 1988). Toms (1997) notes that unstructured, complex problem-solving tasks cannot be reduced in a predictable way to a set of routine Goals, Operators, Methods, and Selections (GOMS). In information interaction, users interact with a system to examine an information blueprint, analogous to traditional reader–text interaction established in a printed-paper world. This is impacted by the system’s management of the content and the system’s ability to communicate with the user (Toms 2002). This text-centric view of information interaction aligns with Toms (1997) model as shown in Fig. 4.3. While shown with a concentration on text, Toms’ model is applicable to other types of information content such as numbers, imagery, audio, and video. In information interaction, humans generally initiate a process of interaction by formulating a goal, e.g. exploration/investigation, decision making, or process/workflow. Once the information is located or provided, the individual scans the information until a cue is noted. Upon noting the cue, the individual may or may not stop to examine the specific information. This process is then potentially repeated in multiple nonlinear ways through categorical selection, cuing and extraction (Toms 2002).

According to Toms’ model of information interaction, users are likely to iterate over available information until evidence to support a viable solution or alternative is identified. This cyclic perspective on information interaction aligns with many of the theoretical models of decision-making, such as Simon’s (1960) decision-making phases shown in Fig. 4.4, as well as extensions of Simon’s model that include a monitoring phase (Mintzberg et al. 1976), and Boyd’s (1987) Observe-Orient-Decide-Act (OODA) loop. When the decision-making cycle is unmoderated, it is not difficult to see how extremes in information (underload and overload) can dramatically impact outcomes, leading to unstructured interaction dynamics.

A significant consideration in human information interaction is the interaction’s ultimate purpose: decision-making. Effective (data driven) decision making relies on a precision balance between the right amount of information, the right amount of time, and the correct ability to execute the choice. Information overload limits humans’ ability to interact with information and thus negatively impacts all three considerations (Marusich et al. 2016; Murayama et al. 2016). Moreover, this balance is predicated on a fundamental understanding of human cognitive and psychological characteristics, within the context of the decision-making situation. Information overload is just an overwhelming of human cognitive abilities (Spier 2016) and it is this overpowering that results in the negative opportunities for things such as bias, improper heuristics usage, and accuracy degradation. Most susceptible are decisions that require complex information interaction, as opposed to those arguably deterministic decisions that only require simple information. In this sense, the application of information theory to human information interaction in the

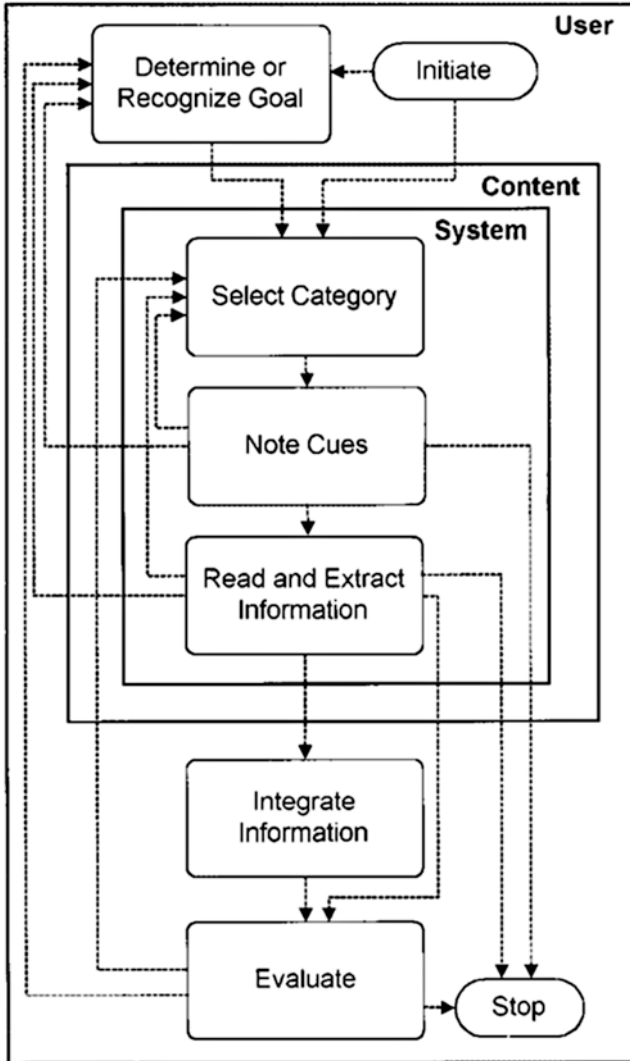
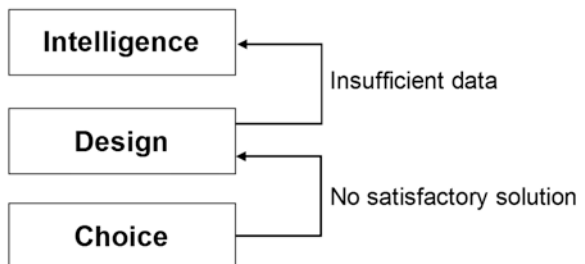


Fig. 4.3 Information interaction model (Toms 2002)

Fig. 4.4 Contextual awareness (Simon 1960)



absence of psychological theoretical grounding is likely to only provide partial solutions.

The true challenge is that our understanding of information theory and the psychological science remain separate and some argue, divergent (Luce 2003). The human information interaction domain and related-problems would underscore the convergence of information theory and theories found in experimental psychology. Although the need for understanding in the converged theoretical space is apparent, this may stem from computer and information scientists' view of humans as part of, or in itself, a complex system. Laming (2001a, b) notes:

If the human operator is viewed as a purely physical system, then statistical information theory would be applicable and analysis of information flow would provide a 'model-independent' technique for identifying the 'information-critical' operations involved.

Laming (2001a, b) goes on to refine this point by indicating how psychologists view information, suggesting their perspective is that of simple, and not complex, information:

Under the influence of Shannon's theory, psychologists are wont to suppose that information is an absolute. Not so! Data is absolute, but information is always relative to the two hypotheses between which it distinguishes.

Laming's statement nuances the difference in complex versus simple information in his statement about information distinguishing between *two* hypotheses. By definition a single hypothesis tests the state of a yes or no outcome or effect, whereas a comparison between two, or potentially more, hypotheses suggest a much richer output state. In a perhaps overly simplistic synthesis of these notions, a hypothesis test results in a Boolean (right/wrong, null-hypothesis or not) output-state not altogether different than Albers (2015) view of simple information. Comparison between hypotheses requires an understanding of the output state value and the compared hypotheses themselves—clearly a much more complicated scope of information. Thus, the human information interaction required to actionably understand one hypothesis versus the comparison of multiple hypotheses is far more complex. This conclusion suggests approaches for both information theory (e.g., measurement, quantification, analysis, etc.) and psychology (e.g., synthesis, comprehension, context, etc.) would be mandated.

The mapping between information theoretic notions and psychological effects is not without their conceptual parallels. Laming (2010) provides a table (Table 4.2) of exemplars that illustrate conceptual alignments. It is noteworthy that even in Laming's work the concept of tasks appears, underscoring the introduction of context. However, this is necessary to consider the human as a physical system (channel) where the stimulus is the message to be transmitted and the response is the message actually received. Nonetheless, Laming posits when the human is considered a physical system, performance is no longer dominated by a capacity limitation; instead, the efficiency of performance depends on the pre-existing match between stimulus and channel characteristics. In the context of HII, the notion of capacity's relationship to performance complicates the study of interaction phenomena. Moreover, typical [rigorous, scientific] study of human interaction effects

**Table 4.2** Conceptual parallels between information theory, an ideal communication system, thermodynamics, and psychological experiments (Laming 2010)

Information theory	Communication theory	Thermo-dynamics	Psychological experimentation
	Message sent		Stimulus
	Message received		Response
Data	Transmission frequencies		Performance data
Null hypothesis	Channel open-circuit		Independence
Alternative hypothesis	Channel functioning but subject to errors		Task completed w/errors
Information statistic	Information transmitted	Work done	Measure of task performance
	Uncertainty	Entropy	Maximum yield of information

seeks to be broadly generalizable, which may be as fleeting as generalizing individual human cognitive perception given dynamic stimuli.

Viewing humans as part of, or simply as, a physical system does have its benefits in bounding relevant variables and would potentially allow information theory to be applicable to complex cognitive problems. However, this does require careful construction of the research methodology where the stimuli are carefully controlled and considered as a flow through that system. In this case, according to the Laming (2010):

The summary of discrimination results between two separate stimuli poses the question: What information is lost in transmission to the discrimination response and how? If that loss of information can be accurately characterized (this is ultimately no more than an analysis of experimental data), the theoretical possibilities are correspondingly constrained.

Given qualitatively constrained problems where quantification of human performance is directly relatable to channel capacity, Shannon’s theory allows the human-system to be modeled. Yet such experimentation and evaluation would be insufficient to broadly describe complex behavioral phenomenon that would naturally exist in human-in-the-loop processes such as human information interaction. It is not unreasonable to conclude that the application of information theory’s relevance to psychology and complex information problems can be constrained to conditions where the information accurately represents a measure of the messaging sent through a physical system or physical-human systems that are strictly analogous to physical systems.

While providing some explanation of the divergence between psychology and information theory, this discussion underscores the need for grounding theories that converge the two domains and provide better understanding of human information interaction. One might argue that, if the interaction is sufficiently decomposed and properly sequenced, Shannon’s information theory should be applicable. This argument assumes that the amount of information (or messages) needed to address a complex information problem is defined and likely known a priori. The relationship of information overload to human information interaction highlights the rarity of

such conditions. In much of the research literature, the investigation of information overload tends to follow the messaging-system model that is appropriate for Shannon's theories (Jones et al. 2004; Sharma et al. 2014; Asadi 2015). When information overload is investigated in situations involving complex information, partitioning of the problem is necessary. This is shown in Jackson and Farzaneh's (2012) work, where they separate intrinsic factors and extraneous factors affecting information overload. They consider intrinsic factors as "information quantity, processing capacity, and available time" and extraneous factors such as information characteristics and quality and task parameters. In the work, they constructed a model that provided quantitative measures of the intrinsic factors, as well as the extraneous factors in terms of their level of contribution and way of interaction. While the intrinsic factors had deterministic measures, Jackson and Farzaneh's model makes significant assumptions about measuring extraneous factors. For example, they consider "Quality of Information" equal to the product of "Validity  $\times$  Relevancy." Given a complex information problem, the validity and relevancy of available information is often unknown until the information is discovered, explored, and understood (Saracevic 2016).

Information overload is an aspect of human information interaction that remains an active area of study. Moreover, this ongoing activity is an indication of the need for additional research on grounding theories in human information interaction. In particular, as innovations in augmenting human information interaction to mitigate challenges from information overload are developed and matured, technologies such as machine reasoning and artificial intelligence will suffer from the same problems as the humans they proxy. These problems will not manifest as effects on the technologies themselves, but will be propagated indirectly as poor performance, limitations, or errors to human users.

Despite recent advances in computational reasoning technologies intended to aid human information interaction, there remains a gap between information theory and the psychological and cognitive sciences. We have discussed two important situations where Shannon's information theory is lacking: one is using channel capacity as a metric for information knowledge, and the other is the use of Shannon theory in the psychological sciences. Shannon himself warned of the shortcomings of his information theory. He cautioned researchers in his famous and short Bandwagon paper (Shannon 1956), and there is ample additional evidence in the literature to support his assertions. The incompatibilities between Shannon's quantitative information theory and our understanding of human cognition underscore the difficulties facing the HII research domain. Further, current work in HII often avoids, or obfuscates, the processing layer between the information and humans. Shannon's information theory aptly describes simple information but lacks the ability to characterize complex information. Particularly when context is considered, applications of Shannon's theory tend to fall short. One might argue that this limitation is why it has been challenging to apply Shannon's theory to problems in psychology. Yet complex information will increasingly be the focus of human information interactions and thus the diversity of theoretical representation will present another barrier to the development and adoption of theories in the field.

Researchers' understanding of HII would seem to be in its early stages, particularly in complex and/or unstructured situations. Yet advances in the computational and information sciences are driving humans' interaction with information at an accelerating pace. The speed of this trend is readily apparent in the prevalence of information analytics and processing. Assuming that all of the information in the world is already in existence (a physics-based view) and only requires contextual transformation to make it "interact-able" with humans, processing is essential and particularly critical. However, this processing must be done with context and situational awareness, which mandates computational methods for learning and reasoning that can produce rationally acting behaviors and outcomes that align with humans' cognitive models and expectations.

### 4.3 HII and Artificial Intelligence

Artificial intelligence is a program, which in an arbitrary world, will cope no worse than a human (Dobrev 2005). While clearly denoting AI as a "program," this definition sets the standard as being bounded by a human reference. Given the range of all AI definitions, and there are many, they all consistently frame AI as a proxy for humans. Thus, HII research is relevant for not only AI but also the relationship that AI has with humans. Wah (1987) characterizes AI processing as requiring symbolic processing, deterministic computations, dynamic execution, parallel processing, open systems, and knowledge management. While HII as a field of study was not considered at the time of Wah's work, his description of AI is indicative of the processing necessary to facilitate humans' interaction with information. Wah (1987) points to knowledge management as an important element of AI as a means to reduce a problem's scope. This statement about problem reduction illustrates Wah's implication that AI would potentially be impacted by information overload and underload. This is not an unreasonable implication because AI is the computational proxy for human information interaction. Furthermore, humans are never completely removed from a process or workflow. In the limit, humans exist on the boundaries of AI activities, if only to initiate or receive benefits of AI-augmented capabilities.

Most implementations of artificial intelligence rely on machine learning methods to create their ability to reason and learn. Machine learning depends on three approaches to achieve pattern recognition: neural, statistical, and structural (Schalkoff 1992). Like a human, a machine's pattern recognition requires training or exemplars on which to build repeatable models. When machine learning is considered in this manner, the challenges of information underload and overload are readily apparent. Key problems that limit machine learning effectiveness involve too little exemplar data (information underload) resulting in precision issues, and too much exemplar data (information overload) resulting in recall issues. Also, similar to humans, machines require additional processing to deal with imbalances in information loads in order to produce preferable outcomes. Due to their inherent

complexity and because complex problems require exceedingly large amounts of useful knowledge (Du and Pardalos 2013), AI problems demand significant computational power. Advances in AI allow increasingly difficult reasoning problems to be addressed (Bond and Gasser 2014; Nilsson 2014) and advanced cognitive activities, such as those that exist naturally in the human brain, represent some of the most difficult reasoning capabilities to artificially recreate.

Even with the challenges and tradeoffs of precision and recall in machine learning implementations, it is managing information overload and underload is where artificial intelligence finds one of its greatest utilities. To address issues of information overload in human interactions, the use of artificial reasoning agents (software AI) has become a dominant contemporary solution (Maes 1994; Aljukhadar et al. 2012; Lohani et al. 2016). Instead of user-initiated interaction via commands and/or direct manipulation, the user is engaged in a cooperative process in which human and computer agents both initiate communication, monitor events and perform tasks. Because information space is the primary environment for AI agents, if they possess the ability to learn, reason, and adapt, the agents can find ways to solve problems with minimal human interaction. In problems involving complex information, intelligent agents are particularly useful, as they have the ability to apply what they have learned to new, unforeseen, or dynamic situations. Complex information domains are domains in which there is constant change, and domains in which many players may interact in solving a problem. Thus, it is not surprising that many of the most successful AI solutions to complex information problems are being led by this segment of the AI community (Hendler 1996). Most intelligent agents are implemented to act on behalf of their human taskers where potential issues of information overload exist as symptoms of higher order activities and goals. In accomplishing these goals and tasks, the agents ideally perform at a higher level of proficiency, efficiency and expediency than human, while still delivering outcomes that are consistent with human belief structures and conceptual models that are cognitively consistent such that trust and acceptance are not issues.

Learning is the means to moderate information overload by lessening the need for human information gathering and other information interaction activities. Minsky (1968) defines AI as “the science of making machines capable of performing tasks that would require intelligence if done by humans” and the smarter the AI becomes, as the result of learning, the greater the scope of assistance provided by AI. Learning and adaptation are critical capabilities for both AI and HII and it is in these areas where the two fields find significant overlap. With the advance of big data analytics and the overwhelming prevalence of available information, machine learning has emerged as a trendy method for giving humans greater interaction with information and as a driver for increased innovations in AI. As an example, content analysis, a fundamental activity in HII, employs a myriad of machine learning approaches to enable artificial intelligence to perform content analysis in volume and autonomously. General definitions of machine learning focus on the design and development of algorithms to create re-applicable models based on generalizations



from limited sets of empirical data in order to adapt to new circumstances and to detect and extrapolate patterns (Russell and Norvig 2003). Therefore, machine learning is the way AI implements the human learning, reasoning, and adapting functions to perform human-like tasks involving information.

Artificial intelligence encompasses other areas of research apart from machine learning, including knowledge representation, natural language processing/understanding, planning. These same areas also have overlap with HII, particularly when one considers a work or an activity context. The purpose of AI is often to automate HII for the purposes of decision-making or work. However, no AI-enabled autonomous system is completely autonomous because at some point a human enters the loop. Ideally, at the end points of the autonomous functionality, the ultimate purpose of any contemporary autonomous activity is to aid or augment a human process. Artificial intelligence without purpose is pointless, in the same manner human information interaction without an objective is merely iterating over data. The objective and purpose form the goals of work that humans, and the artificial intelligence that aid them, execute. The humans and AI that interact with information behave as actors involved in work related actions. This perspective assumes that to be able to design systems that work harmoniously with humans, the work human actors do, their information behaviors, the context in which they work, and the reasons for their actions must be understood (Fidel 2012).

The literature documents many ways that artificial intelligence is applied to information centric activities such as text processing (Vasant 2015), searching (Shrobe 2014), decision making (Hendler and Mulvehill 2016), and planning (Rich and Waters 2014; Kerr and Szelke 2016). In larger systems, artificial intelligences that provide low-level functionality are connected and integrated to deliver bigger solutions and greater functionality. In these large systems and application both tasks and information are decomposed into digestible bits and coupled with learning. Table 4.3 describes levels of automation that support decision making and action.

In Table 4.3, if the word “computer” is replaced with “artificial intelligence” the levels descriptions, still make sense and would be applicable. Automation is not an all-or-nothing phenomenon; there are degrees of automation. In the lowest level of automation, no support is provided. At the highest level, the human is completely removed from the decision activity. When artificial intelligence is operating on behalf of the human at the highest level, level 10, the artificial intelligence decides everything, acts autonomously, and has no human interaction. Level 10 is currently achieved in many low-level information interaction and related decision-making activities. The more complex the task and thus the information, the more challenging it is to reach the higher levels of autonomy. Moreover, as the human is increasingly removed from an activity, there is less confirmation and opportunities for human re-direction. It is generally when a system exhibits behaviors above Level 6, in complex situations, where issues of human trust and ethics become a concern (Alaieri and Vellino 2016).

**Table 4.3** Levels of automation (adapted from Parasuraman et al. 2000)

Levels of automation of decision and action selection	
High	10. The computer decides everything, acts autonomously, ignoring the human
	9. The computer informs the human only if it (the computer) decides to
	8. The computer informs the human only if asked (by the human) or
	7. Executes automatically, then necessarily informs the human, and
	6. Allows the human a constrained time to veto before automatic execution, or
	5. Executes the suggestion if the human approves, or
	4. Suggests one alternative, or
	3. Narrows the selection down to a few, or
	2. The computer offers a complete set of decision/action alternatives or
	1. The computer offers no assistance: the human must make all decisions and actions
Low	

According to Parasuraman et al. (2000), automation, implying artificial intelligence, is not an exact science and neither does it belong in the realm of the creative arts. Systems providing solutions that deliver accurate answers that humans trust and subsequently utilize often require a deep understanding of the relevant goal. Going beyond an understanding of the goal, goal decomposition and contextually related tasks necessary to achieve an objective must be completely and fully understood as well (Anderson 2014; Harkin et al. 2015). If a goal's tasks and its requirements are fully understood, it becomes possible to use artificial intelligence to learn a problem and apply what it has learned to new challenges.

Goal attainment when intelligent software or physical, e.g., robotic, agents are considered is even more complex than simple problem solving. However, nearly all goals can be better achieved given more resources (Omohundro 2008). This suggests that artificially intelligent agents requiring information interaction would have an incentive to acquire additional resources, even those that may be in use by humans. Thus, it's not unreasonable to envision situations where some goals would put artificial intelligence at odds with human interests, giving the AI incentives to deceive or manipulate its human operators and resist interventions designed to change or debug its behavior (Bostrom 2014). Reliable and error-tolerant artificially intelligent agent designs are only beneficial if the resulting agent actually pursues desirable outcomes (Soares and Fallenstein 2014). This competitive goal seeking is often the topic of science fiction movies where the artificial intelligence takes over society and deems humans non-essential. While movies frequently portray the competitive situation in the extreme, humans' increasing dependence on smart automation coupled with artificial intelligence's lack of human emotion, bias, morals, and psychological limitations make these storylines plausible.

Automation emphasizes efficiency, productivity, quality, and reliability, focusing on systems that operate fully-autonomously, often in structured environments over

extended periods, and on the explicit structuring of such environments (Goldberg 2012). However, information environments vary broadly in terms of their structure, forming the rules and association of information objects that artificial intelligence operations act on (Stonier 2012). In this manner, artificial intelligence is a proxy for humans in their interaction with information, but with the added dimension that humans interact with the artificial intelligence, essentially creating a recursive HII loop. AI interacts with information and humans interact with AI, which is itself information. This recursive relationship can both minimize and amplify opportunities for errors.

Ideally, an AI implementation delivers seamless interaction with the information environment and the real world (in some cases) to accomplish human intent. Because AI is driven by machines, the number of information transactions will be much higher than those generated by humans. As a proxy for humans, AI's interaction with information will similarly increase, if only in the encoding and translation of representations. Within the AI-proxy, there is a potential for errors in the interdependence between information and intent. Since system errors occur at the intersection of logic and data, increases in information interaction (human or otherwise) can increase the potential for errors.

#### 4.4 HII, AI, and Errors

The lack of understanding in human information interaction coupled with increasing dependence on automation and sophistication of artificial intelligence technologies will likely lead to unpredictable system behaviors and subsequent outcomes. As artificial intelligence becomes more effective at making decisions involving complex information and highly variable environmental conditions, the need for a theoretical understanding of HII will be necessary to ensure that artificially intelligent systems performs as expected by human operators. Information interaction is an opportunity for automation that without deep understanding of the tasks and/or context increases the likelihood of errors. There is little insight in the revelation that people make errors when using information systems. Nonetheless, errors can be serious, both in the sense that they can lead to significant mishap and even physical harm, and in the social sense that they frustrate or inconvenience humans (Norman 1983).

Figure 4.5 illustrates classes of errors that result from interactions. From an information interaction perspective, regardless of whether the interaction involves humans or their AI proxies, the implications of the classification on error understanding, handling and mitigation apply. Referring to Fig. 4.5, mistakes characterize the direct result of information interaction. Even slips mark the domain of AI outcomes, as well as human decision making resulting from information interaction. Moreover, as noted previously, humans are never completely removed from AI-enhanced automated processes. Thus, humans are always responsible for system disasters, if only because they are the visible element of system performance. While

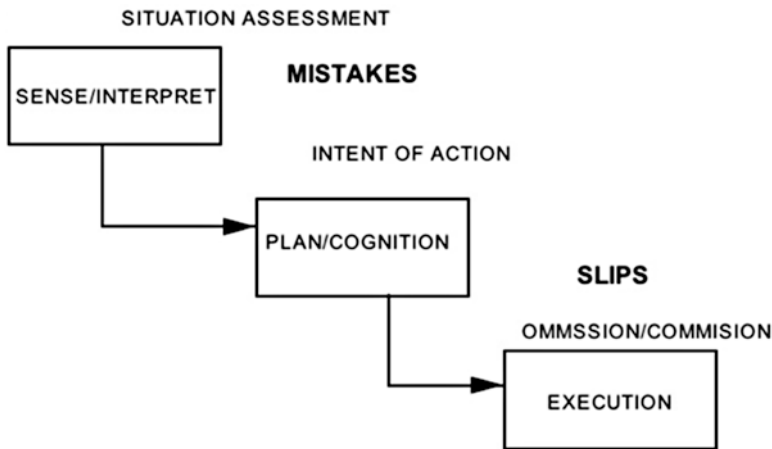


Fig. 4.5 Classes of errors (Norman 1983)

generalized notions of errors would spur debate about trust, blame, and culpability when errors are considered in the abstraction of information interaction and AI automation, the distinction of active and latent errors will become increasingly difficult to partition.

Active errors are those associated with the performance of frontline operators/actors and latent errors are those related to activities removed in time and/or space from the direct control interface (Reason 1990). Moreover, active and latent errors in the context of information interaction and AI will conflate the two approaches as the problem of human fallibility: the person and the system. The person approach focuses on the errors of individuals, e.g., forgetfulness, inattention, moral weakness, etc. The system approach concentrates on the conditions under which individuals work (Reason 2000). Errors resulting from AI conducting HII will obfuscate the line between the system and the humans. A contemporary view of artificial intelligence that supports complex activities, such as automated driving, questions the safety of the systems: Researchers studying errors no longer assume the “system” is safe, properly attributing problems to human or mechanical failure. They generally believe the line between human error and mechanical failure will be blurry (Dekker 2014). Moreover, with regard to cause, Dekker posits human error is not the cause of failure, but is a symptom of failure.

Detailed analysis from recent accidents, many of which resulted in a loss of life and extensive long-lasting damage, have made it apparent that latent errors pose the greatest threat to the safety of a complex system (Reason 1994). The amalgamation of the person approach and system approach will make error understanding and mitigation more difficult. Consider how errors in information interaction may be perceived or attributed when the device boundaries are dissolved and the information objects are given focus. Issues of misinformation, misuse, improper processing, and out-of-context alignment will increasingly be the norm. The promise of artificial intelligence will find its grounding in machine learning, which can obfuscate as

well as enhance human information interaction. The susceptibility of machine learning techniques to their underlying information (data and distribution) cascades into the AI that depends on it. This weakness is subsequently transferred to the automation and humans that rely on it, ultimately manifesting as latent errors when conditions arise that meet the unforeseen convergent of knowledge and logic.

The same power that machine learning brings to AI is the same weakness and subsequent susceptibility that it embeds in AI systems. To this end, it is possible for malicious users and designers to deliberately generate errors and problems that allow systems to be compromised and/or manifest behaviors that are harmful to benign users. Consider how much artificial intelligence functionality is implemented in software. This software relies on libraries that implement machine learning algorithms (Marsland 2015). That means the implementation of artificial intelligence execution code will be widely understood by many developers as well as those seeking to hack systems. These same algorithms are also being implemented in hardware, creating further embedded potentials for problems that are significantly more difficult to address post-implementation and operation. Contemporary research has shown that it is possible to build a meta-classifier and train it to hack other machine learning classifiers and infer information about their training sets; setting the hacked systems up for manipulated operations (Ateniese et al. 2015). In this sense, machine learning displays problems of a similar nature to human biases and other cognitive framing limitations. This explanation implies that artificial intelligence may ultimately be susceptible to errors similar to their human proxies. Thus, investigation into theoretical aspects of errors in HII may provide insights and solutions that mitigate the errors resulting from artificial intelligence processes.

Figure 4.5 illustrates classes of errors, but primarily procedural and process errors. Errors involving information interaction may require a broader taxonomy of error classification. Primiero (2014) suggests a taxonomy for information systems that extends Norman's (1983) model with three additional categories that represent system correctness: (1) conceptual validity, relating to the conceptual description and design of the system goals; (2) procedural correctness, relating to the functional aspects of the system goals; and (3) contextual admissibility, relating to both the conceptual and procedural aspects in the systems' execution environment. Primiero considers aspects of conceptual validity to include system and problem design, procedural validity inclusive of data and semantics, and contextual validity essentially the appropriateness for the match of concept and procedure. Primiero goes on to generalize this basic categorization in aligning his taxonomy with Norman's (e.g., see Table 4.4).

Primiero's taxonomy provides a structure for validating errors. In this taxonomy, Primiero considers problems of design or structuring to be mistakes. This is whether design or structuring is applied to a decision problem or a system. In a system context, mistakes typically manifest as defects that have a role in a failure. In Primiero's taxonomy, failures are errors that occur during the evaluation and resolution of the problem or a system's function. Slips are errors that occur as exceptions such as reduced efficiency or less than expected performance. As can be seen in failures that

**Table 4.4** Categories of errors (Primiero 2014)

Type of error	Conceptual	Material
Mistakes	Problem description: categorization	Problem design: category structuring
Failures	Procedure definition: form of main process	Procedure construction: accessibility of dependent processes
Slips	Algorithm design: efficiency	Algorithm execution: performance

result from a mistake, Primiero's categories are not discrete. Slips can also cause failures. Consider a slip where system efficiency degrades to the point of halting the system.

As an example of applying Primiero's taxonomy to a problem readily solvable by artificial intelligence (Nilsson 1969; Gil et al. 2004; Strong 2016), consider an organization's scheduling assistant software. AI's role in providing this automated function is to be a proxy for its human users, to understand the appointment/resource constraints, and to identify convergences without conflict. When all the planning conditions and constraints are known by the system, artificial intelligence is not appropriate for the problem. However, when some of the conditions are unknown AI can be used to adapt policies and models. The linkages to human information interaction is clear in this problem because the AI would necessarily require and access the same types and sources of information that a human would use to arrive at a solution. Examining problems that might occur in human processing of meeting-planning activities, an incorrect or incomplete understanding of the schedules for other people whose attendance was required would result in errors in the meeting's completeness, perhaps preventing the purpose of the meeting from being achieved. In Primiero's taxonomy this error would stem from a mistake. The mistake would have been incorrect or missing elements in the schedule model, a misstructuring of the multi-constraint resolution, or a limited problem design that could not handle ambiguous or incorrect schedule constraints. This mistake may result in a failure of the meeting purpose/process, or a slip if the meeting achieved less than desirable results or required additional meetings. While this example greatly simplifies the problem of scheduling and does not provide much detail of the human information interaction required, it does illustrate how Primiero's taxonomy would be applicable to AI's algorithmic human information interaction functions. Despite the simplicity of this example, even in uncomplicated situations the activities of algorithms embedded in complex automated systems can have greater obfuscation than human activities. The complexity of the system, the degree of which tends to be proportionate to the sophistication of the system (Lloyd 2001), increases the likelihood of latent errors. There is little question that systems with artificial intelligence operating on information and intending to automate sophisticated human processes fall into the category of highly complex systems. The relationship between complexity and errors is a well-studied phenomenon (Ferdinand 1974; Basili and Perricone 1984; Pincus 1991; Meseguer 2014).

Unless a complex process is entirely automated, it is the product of technology, the human user and how well each fits the other. When the technology has a capability to reason and learn, then the product may be viewed as a “knowledge coupling” between human and machine (Meseguer 2014). As discussed earlier, problems involving complex information interaction require complex systems to provide solutions. This implies the relationship between increased complexity and likelihood of error would have a high degree of correlation. According to Dekker (2016), in complex systems, decision-makers (automated or otherwise) are locally, rather than globally, rational. Even in the case of AI, this scoping of rational understanding is a factor in increased probabilities of errors because in complex systems local decisions and actions can lead to global effects. Dekker (2016, p. 29) suggests this is because of an intrinsic property of complex systems: “the multitude of relationships, interconnections and interdependencies and interacting of interacting and interdependent agents, or components.” While Dekker focuses on system failures, his summarization underscores the core problem that emphasizes the need for increased research on human information interaction and the challenge of preventing AI errors as AI becomes the dominant proxy for humans in automated processes. Dekker (2016 p. 29) states “adaptive responses to local knowledge and information throughout the complex system can become an adaptive, cumulative response of the entire system – a set of responses that can be seen as a slow but steady drift into failure.”

In information systems design, the notion that errors occur at the intersection of data and logic forms a basis for the problem with AI automating and supporting human information interaction. Considering this notion, the strength of AI (its ability to reason, learn, and understand) is also its weakness. This reasoning is because it is impossible to achieve completeness in descriptions of complex systems—whether before, during or after their lifetime (Cilliers 1998). The underlying issue is traceability in complex situations involving a variety of information. When errors occur in a complex system, we should gather as much information on the issue as possible. Of course, complexity makes it impossible to gather “all” of the information, or for us to even know how much information we have gathered (Dekker 2016). Moreover, these conditions impede attribution of the cause and complicate the resolution of the problem. Latent errors that occur only under certain information, environmental, and operational conditions will be difficult to anticipate, identify, or resolve, ultimately leading to additional errors.

Consider the case of commercial airline piloting, where AI and automation have had a long-standing role. While there is still much debate about AI removing human pilots altogether, flight-system automation is one of the most mature complex applications of AI-supported HII in existence. The disappearance of flight MH370 in March 2014 has never been solved, nor has the wreckage ever been found. This may seem unlikely in an age of persistent surveillance (e.g. satellite imagery, electromagnetic scanning, etc.) but it occurred. Searches have looked for the wreckage for more than 2 years. This type of catastrophic failure has all the markings of the manifestations of errors when AI is involved. Further, there are two dimensions of this problem. The first is the cause of disappearance of the plane, which has been

attributed to everything from a pilot bathroom break, to terrorism, to equipment failure (McNutt 2014). The second dimension is that of the search itself, which at least in part, is a human information interaction activity because of the existence of persistent surveillance technologies (Davey et al. 2015). After more than 2 years of physical and information searching, the plane remains undiscovered. Similarly, as artificial intelligence technologies enable automated driving, latent AI/human information interaction errors are leading to catastrophic unanticipated outcomes. This problem was evident in the first fatality involving a Tesla automated-driving passenger vehicle (Singhvi and Russell 2016). While the loss of life was significant, a second significant outcome of the accident was causal attribution. In this Tesla accident, a driver had engaged the auto-pilot mechanism and may have been distracted watching a DVD when a large freight truck turned in front of the car. The car did not stop. Instead it collided with and went under the truck, then skidded off the road, went through a fence and finally collided with a telephone pole.

Investigators ultimately, but not conclusively, determined that there was a problem with the AI-enabled laser radar (lidar) system in the Tesla vehicle. But that conclusion was not deterministic, as investigators also faulted the brakes, the crash-avoidance system, and even the driver (Ackerman 2016). The various conclusions of the investigation underscore Dekker (2016) description of failures involving complex system:

In a complex system, there is no objective way to determine whose view is right and whose view is wrong, since the agents effectively live in different environments. This means that there is never one “true” story of what happened.

It is interesting to note that in the literature regarding the accident, few people highlighted the importance or relevance of the information interactions that the AI in Tesla’s systems had to perform. This interaction, if it was discussed at all, was framed as a mechanical interaction with signaling but not as a proxy for human interaction. There were many “local” decisions being made by many subsystems in the car that could have affected the “global” outcome of the complex AI system. It is likely that information overload and underload conditions existed and were relevant to the ultimate outcome. The lidar system error conclusion focused on the fact that the system could not effectively distinguish between a truck’s white color and a brightly sunlit background sky (The Tesla Team 2016). While it is possible that the lidar AI simply “did not see the truck,” much like a person, it is equally likely that other information was available that indicated the presence of the truck. Even given the cascading lidar mistake, a better understanding of human information interaction might have added additional controls for just-in-time interactions or more sophisticated handling of the uncertainty involved in the complex situation.

Latent errors pose the greatest threat to safety in a complex system because they are often unrecognized, but they have the capacity to result in multiple types of active errors (Reason 1990; Kohn et al. 2000), such as those identified by Primiero. Latent errors are even more difficult to diagnose, address and resolve because errors in complex AI systems tend to represent latent failures coming together in unexpected ways to produce unique or infrequent results. Since the same confluence of



latent error factors are likely only with a low or obfuscated frequency, identification and prevention can be virtually impossible. Strategies to predict latent errors will be increasingly critical to AI systems. Extensive simulation may be one approach to minimize the proliferation of latent errors (Khoshgoftaar and Munson 1990), but given the volume and heterogeneity of data in human information interaction activities, simulation is likely to face scalability problems.

Artificial intelligence that learns errors may be another effective strategy (Arora and Ge 2011); proponents of AI would argue that AI itself may solve latent errors, essentially reducing or eliminating “human” errors. However, that is an overly optimistic view. It is likely that AI may reduce human errors, but one might also argue it simply shifts the source of the error from the human to the AI system. Worse, errors involving AI are more likely to be latent errors that are very difficult to address. The need for additional research in HII and the rapid advancement of AI implementation may paint a picture that is rife with error riddled challenges. Yet a better understanding and the adoption of computational explanations may provide a way forward that builds defenses to the problems of latent error intrinsically into complex AI systems.

It is clear that, in fact, the power to explain involves the power of insight and anticipation, and that this is very valuable as a kind of distance-receptor in time, which enables organisms to adapt themselves to situations which are about to arise. (Craik 1967, p. 51)

If interpretable explanations were tightly integrated with AI’s reasoning and outcome mechanisms, when problems occur there would be traceability that would naturally have produced insights to the myriad of relevant variables that may have led to the unexpected outcome. Such explanations could illuminate AI strengths and weaknesses, as well as convey information about how the AI will behave in future situations with alternate conditions. Explanations of this nature would necessarily have to go beyond simple rules or counterfactual causation; they would likely have to be probabilistic. The need for probabilistic explanation approaches is due to the nature of complex HII problems. Ample evidence has already been presented showing the challenges of conclusively identifying the factors and circumstances of a problem situation and its outcome. Humphreys (2014) refers to this as the multiplicity of causes. He uses the example of a medical problem, another area where AI as an HII proxy has found some measured success, to illustrate the multiplicity of causal influences:

Successively adding 1) a smoking level of twenty cigarettes a day, 2) medium-high blood pressure (140/88) and 3) medium-high serum cholesterol levels (250 mg/dl) increases the probability of having a heart attack within the next twelve years for a forty-six-year-old man from .03 to [given the above factors, additively] 1) 5%, 2) 7.5%, [to] 3) 15%.

This example shows how combined factors can affect the likelihood of an outcome and subsequently its explanation. Modern medical expert systems often employ this approach in providing the guidance for diagnoses guidance. In the medical example the explanation is allowed to be incomplete but with probabilistic bounding. It is interesting to note that this is the same languages that medical doctors often use, even in non-AI augmented diagnoses: “given your symptoms it is

*likely* that you just have a cold;” or perhaps where language such as “it *seems* you have a cold.” Modern doctors seldom have a deterministically conclusive presentation of a diagnosis (Scheff 1963), and the notion of second opinion underscores the probabilistic nature that humans are accustomed to in medical problem solving. As medical practice evolves, it is increasingly becoming an application of HII—just in the process of symptom pattern matching alone, doctors have become reliant on computational automation.

Humphreys also acknowledges the role of information interaction in explanation. He touches on the notion of completeness by highlighting the delineation between causal explanations is a fuzzy boundary. Hempel formulates the requirement of maximal specificity to find the right balance between background information and relevant information. As an example, Hempel (1966, p. 299) cites a historical exchange:

The astronomer Francesco Sizi claimed, against his contemporary Galileo, that there could not be satellites circling around Jupiter and offered the following argument: ‘There are seven windows in the head, two nostrils, two ears, two eyes and a mouth; so in the heavens there are two favorable stars, two unpropitious, two luminaries, and Mercury alone undecided and indifferent. From which ... we gather that the number of planets is necessarily seven’.

These features, Hempel complained, did not have a proper relation to the planets; and thus, they are not the kind of features that could explain the arrangement observed by Galileo. Probabilistic explanation could suffer from similar relational bias. However, with regard to explainable AI, issues of this nature are not uncommon in machine learning. These issues manifest as problems of overfitting, incorrect feature selection, or training data. If the machine learning on which AI relies for reasoning had explanations for their models, then errors cascading from these models could be used to infer their relationship to generated outcomes. In most AI implementations, formal models of machine learning exist, but they are difficult to interpret resulting in opportunities for latent errors. Like HII, explainable machine learning, and thus explainable AI, is a research area that is still not fully understood (Mooney and Ourston 1989). Nonetheless, the U.S. Defense Advanced Research Programs Agency (DARPA) (Gunning 2016) has a program that seeks to investigate methods of embedding explanation in machine learning algorithms. Figure 4.6 describes the purpose of the DARPA research program. Contemporary machine learning techniques use deep-learning feature detection and also interpretable models with model induction to increase their explainability, while not reducing algorithmic performance. Although the purpose of DARPA’s program is not focused on error reduction or elimination, elements of the program employ the outputs of the explainable models through an explanation interface that is grounded in HCI and psychological theories of explanation (Kulesza et al. 2015). Key metrics of the explanation effectiveness include human centered measures such as user satisfaction, mental model factors, and trust, in addition to quantitative measures such as performance and correct-ability. It is noteworthy that elements of the DARPA program include identifying and correcting errors as an important effectiveness measure. Despite explicitly concentrating on HII issues, the explainable artificial

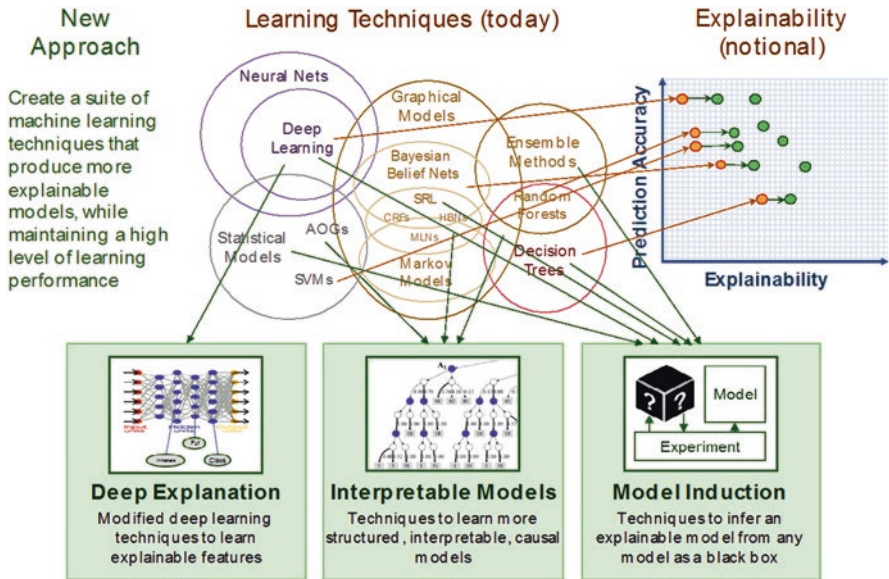


Fig. 4.6 Explainable artificial intelligence (Gunning 2016)

intelligence research program includes many aspects of those concerns. If successful the DARPA program will likely demonstrate that artificial intelligence embedded with explanation will improve human information interaction, reduce errors, and increase traceability when they do occur.

## 4.5 Conclusion

The lack of agreed upon definitions in the emergent HII domain presents a significant impediment to understanding the interdisciplinary complexity of this research area. However, the trending computational nature of all sciences (e.g., physics, biology, chemistry, etc.) will force the need for a better theoretical and practical understanding of HII. The idea of designing information as an activity, separate from the design of the machines containing the information, will move beyond an emergent research area to one that defines not only information science, but other sciences as well. A world characterized by computation will drive the notion of “things” made from information; shifting human models of artificial intelligence and its application for autonomous and seamless work.

The cost of voluminous “information things” will be realized in a dramatic rise in information imbalances (overload and underload) that will impact decision-making processes, whether they be made by human or artificially intelligent decision makers. In information interaction activities and automation, the effects of information imbalances will affect AI in the same way it affects humans. Automated

goal attainment is a complex problem that requires a deep understanding of not only the objective but the underlying tasks and processes, as well. Where gaps exist in the understanding of the process, tasks, or precise information requirements, failures, mistakes, and slips will occur. Because of the difficulty in achieving the understanding necessary for AI solutions and the complexity of information interaction problems, hard to identify and resolve latent errors will become the dominant type of error.

AI will be a key enabler as human information interaction expands. Advances in AI will accelerate the need for fundamental research in HII. Increasingly fulfilling the role of humans, AI will not ever likely completely remove humans from information interactions. Thus, the interdependence between humans, AI and information processing will result in increased latent errors that conflate system and person errors. It is not unreasonable to expect systems to have excessive occurrences of latent errors in information interaction, resulting from increased AI usage. Greater understanding of information interaction can reduce latent errors and potentially minimize interdependence between person and system approaches to fallibility.

Computational implementation of probabilistic explanation is a promising approach that will incorporate the fundamentals of human information interaction with methods that provide insight to machine learning and AI reasoning. Probabilistic explanation embeds explanation in a learning subsystem. This provides traceability at the critical juncture between information and the AI logic that acts on it. Effective implementation of computational explanations that deal with issues of completeness and relativity will be necessary for AI system errors to be permanently resolved.

This chapter provided a review of human information interaction and showed how AI is, and will continue to be, a proxy for humans in that context. HII is where the intersection of AI and human error occur. The opportunity for AI to address (and potentially cause) errors will force the demand for new models of human error and methods for causal explanation. Given these trends, increased research focus on applying Shannon's seminal theories to psychological advances, providing a theoretical grounding for HII, and developing new methods of computational explanation, will both become progressively important.

## References

- Abowd, G. D., Dey, A. K., Brown, P. J., Davies, N., Smith, M., & Steggle, P. (1999, September). Towards a better understanding of context and context-awareness. In *International Symposium on Handheld and Ubiquitous Computing* (pp. 304-307). Springer Berlin Heidelberg.
- Ackerman, E. (2016). Fatal Tesla self-driving car crash reminds us that robots aren't perfect. *IEEE-Spectrum*, 1.
- ACM SIGCHI. (1992) *ACM SIGCHI curricula for human-computer interaction*. New York, NY, USA.
- Alaieri, F., & Vellino, A. (2016). Ethical Decision Making in Robots: Autonomy, Trust and Responsibility. In *International Conference on Social Robotics* (pp. 159-168). Springer International Publishing.

- Albers, M. J. (2011). Usability and information relationships: Considering content relationships and contextual awareness when testing complex information. *Usability of complex information systems: Evaluation of user interaction*, 3-16.
- Albers, M. J. (2015, June). Human-Information Interaction with Complex Information for Decision-Making. In *Informatics* (Vol. 2, No. 2, pp. 4-19). Multidisciplinary Digital Publishing Institute <http://www.mdpi.com/2227-9709/2/2/4/htm>.
- Alexander, B., Barrett, K., Cumming, S., Herron, P., Holland, C., Keane, K., Ogburn J., Orlowitz, J., Thomas MA., Tsao, J. (2016). Information Overload and Underload. *Open Scholarship Initiative Proceedings*, 1.
- Aljukhadar, M., Senecal, S., & Daoust, C. E. (2012). Using recommendation agents to cope with information overload. *International Journal of Electronic Commerce*, 17(2), 41-70.
- Allwein, G. (2004) A Qualitative Framework for Shannon Information Theories, In *Proceedings Of The 2004 Workshop On New Security Paradigms*, 23-31.
- Anderson, J. R. (2014). *Rules of the mind*. Psychology Press.
- Arora, S., & Ge, R. (2011, July). New algorithms for learning in presence of errors. In *International Colloquium on Automata, Languages, and Programming* (pp. 403-415). Springer Berlin Heidelberg.
- Asadi, M. (2015). "Information Theory" Research Trend: A Bibliometric Approach. *SLIS Connecting*, 4(1), 45.
- Ateniese, G., Mancini, L. V., Spognardi, A., Villani, A., Vitali, D., & Felici, G. (2015). Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers. *International Journal of Security and Networks*, 10(3), 137-150.
- Barnaghi, P., Sheth, A., Singh, V., & Hauswirth, M. (2015). Physical-Cyber-Social Computing: Looking Back, Looking Forward. *Internet Computing, IEEE*, 19(3), 7-11.
- Barwise, J. & Seligman, J. (1997) Information Flow: The Logic of Distributed Systems, *Cambridge Tracts in Theoretical Computer Science* 44.
- Basili, V. R., & Perricone, B. T. (1984). Software errors and complexity: an empirical investigation. *Communications of the ACM*, 27(1), 42-52.
- Bolotin, A. (2015, January). Any realistic model of a physical system must be computationally realistic. In *Journal of Physics: Conference Series* (Vol. 574, No. 1, p. 012088). IOP Publishing.
- Bond, A. H., & Gasser, L. (Eds.). (2014). *Readings In Distributed Artificial Intelligence*. Morgan Kaufmann.
- Bostrom, N. 2014. *Superintelligence: Paths, Dangers, Strategies*. New York: Oxford University Press.
- Boyd, J. R. (1987). Organic design for command and control. *A discourse on winning and losing*.
- Buckland, M. K. (1991). Information as thing. *Journal of the American Society of Information Science*, 42(5), 351-360.
- Chang, E., & Bennamoun, M. (2012 June), Message from the Conference General Chairs. *5th International Conference on Human System Interactions (HSI) 2012*
- Cilliers, P. (1998). *Complexity and postmodernism: Understanding complex systems*. London: Routledge.
- Clark, H. H., & Carlson, T. B. (1981). Context for comprehension. *Attention and performance IX*, 313-330.
- Craik, K. J. W. (1967). *The nature of explanation* (Vol. 445). CUP Archive.
- Davey, S., Gordon, N., Holland, I., Rutten, M., & Williams, J. (2015). Bayesian Methods in the Search for MH370. *Defence Science and Technology Group*.
- Dekker, S. (2014). *The field guide to understanding 'human error'*. Ashgate Publishing, Ltd.
- Dekker, S. (2016). *Drift into failure: From hunting broken components to understanding complex systems*. CRC Press.
- Dervin, B., & Reinhard, C. D. (2006). Researchers and practitioners talk about users and each other. Making user and audience studies matter paper. *Information research*, 12(1), 1.
- Devlin, K. (2001). Claude Shannon 1916-2001. *Focus: The Newsletter of the Mathematical Association of America*, 21, 20-21.

- Dey, A. K. (2001). Understanding and using context. *Personal and ubiquitous computing*, 5(1), 4-7.
- Dobrev, D. (2005). *Formal definition of artificial intelligence*.
- Dourish, P. (2004). *Where The Action Is: The Foundations Of Embodied Interaction*. MIT press.
- Du, D. Z., & Pardalos, P. M. (Eds.). (2013). *Handbook of combinatorial optimization: supplement* (Vol. 1). Springer Science & Business Media.
- Ferdinand, A. E. (1974). A theory of system complexity. *International Journal of General System* 1(1), 19-33.
- Fidel, R. (2012). *Human Information Interaction: An Ecological Approach To Information Behavior*. MIT Press.
- Gershon, N (1995). Human Information Interaction, *WWW4 Conference*, December 1995.
- Gil, Y., Deelman, E., Blythe, J., Kesselman, C., & Tangmunarunkit, H. (2004). Artificial intelligence and grids: Workflow planning and beyond. *IEEE Intelligent Systems*, 19(1), 26-33.
- Goldberg, K. (2012). What is automation? *IEEE Transactions on Automation Science and Engineering*, 9(1), 1-2.
- Gunning, D. (2016). Program Information: Explainable Artificial Intelligence (XAI), *Defense Advanced Research Projects Agency*. Available at [http://www.darpa.mil/attachments/XAIIndustryDay\\_Final.pptx](http://www.darpa.mil/attachments/XAIIndustryDay_Final.pptx) Accessed January 10, 2017.
- Harkin, B., Webb, T. L., Chang, B. P., Prestwich, A., Conner, M., Kellar, I., Benn, Y., & Sheeran, P. (2015). Does monitoring goal progress promote goal attainment? A meta-analysis of the experimental evidence. *Psychological Bulletin*, 142(2), 198-229.
- Hempel, C. G. (1966). Philosophy of natural science.
- Hendler, J. A. (1996). Intelligent agents: Where AI meets information technology. *IEEE Expert*, 11(6), 20-23.
- Hendler, J., & Mulvehill, A. (2016). *Social Machines: The Coming Collision of Artificial Intelligence, Social Networking, and Humanity*. Apress.
- Henricksen, K., Indulska, J., & Rakotonirainy, A. (2002, August). Modeling context information in pervasive computing systems. In *International Conference on Pervasive Computing* (pp. 167-180). Springer Berlin Heidelberg.
- Humphreys, P. (2014). *The chances of explanation: Causal explanation in the social, medical, and physical sciences*. Princeton University Press.
- Jackson, T. W., & Farzaneh, P. (2012). Theory-based model of factors affecting information overload. *International Journal of Information Management*, 32(6), 523-532.
- Jones, Q., Ravid, G., & Rafaeli, S. (2004). Information overload and the message dynamics of online interaction spaces: A theoretical model and empirical exploration. *Information systems research*, 15(2), 194-210.
- Kerr, R., & Szelke, E. (Eds.). (2016). *Artificial intelligence in reactive scheduling*. Springer.
- Khoshgoftaar, T. M., & Munson, J. C. (1990). Predicting software development errors using software complexity metrics. *IEEE Journal on Selected Areas in Communications*, 8(2), 253-261.
- Kohn, L. T., Corrigan, J. M., & Donaldson, M. S. (2000). Why Do Errors Happen?
- Kulesza, T., Burnett, M., Wong, W.-K., & Stumpf, S. (2015). Principles of Explanatory Debugging to Personalize Interactive Machine Learning. *IUI 2015, Proceedings of the 20th International Conference on Intelligent User Interfaces* (pp. 126-137).
- Lamb, R., & Kling, R. (2003). Reconceptualizing users as social actors in information systems research. *MIS Quarterly*, 197-236.
- Laming, D. (2001a). Statistical information, uncertainty, and Bayes' theorem: Some applications in experimental psychology. In S. Benferhat & P. Besnard (Eds.), *Symbolic and Quantitative Approaches to Reasoning With Uncertainty* (pp. 635- 646). Berlin: Springer-Verlag.
- Laming, D. (2001b). Statistical information, uncertainty, and Bayes' theorem: Some applications in experimental psychology. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty* (pp. 635-646). Springer Berlin Heidelberg.

- Laming, D. (2010). Statistical information and uncertainty: A critique of applications in experimental psychology. *Entropy*, 12(4), 720-771.
- Limkar, S., & Jha, R. K. (2016). Technology Involved in Bridging Physical, Cyber, and Hyper World. In *Proceedings of the Second International Conference on Computer and Communication Technologies* (pp. 735-743). Springer India.
- Liu, K. (2000) *Semiotics in Information Systems Engineering*, Cambridge University Press, Cambridge.
- Lloyd, S. (2001). Measures of complexity: a nonexhaustive list. *IEEE Control Systems Magazine*, 21(4), 7-8.
- Lohani, M., Stokes, C., Dashan, N., McCoy, M., Bailey, C. A., & Rivers, S. E. (2016). A Framework for Human-Agent Social Systems: The Role of Non-Technical Factors in Operation Success. In *Advances in Human Factors in Robots and Unmanned Systems* (pp. 137-148). Springer International Publishing.
- Luce, R. D. (2003). Whatever happened to information theory in psychology? *Review of general psychology*, 7(2), 183.
- M. L. Minsky, editor. *Semantic Information Processing*. MIT Press 1968.
- Maes, P. (1994). Agents that reduce work and information overload. *Communications of the ACM*, 37(7), 30-40.
- Marchionini, G. (2008) Human-information interaction research and development, *Library & Information Science Research*, Volume 30, Issue 3, September 2008, Pages 165-174, ISSN 0740-8188
- Marois, R., & Ivanoff, J. (2005). Capacity limits of information processing in the brain. *Trends in Cognitive Sciences*, 9(6), 296-305.
- Marsland, S. (2015). *Machine learning: an algorithmic perspective*. CRC press.
- Marusich, L. R., Bakdash, J. Z., Onal, E., Michael, S. Y., Schaffer, J., O'Donovan, J., Gonzalez, C. (2016). Effects of Information Availability on Command-and-Control Decision Making Performance, Trust, and Situation Awareness. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 58(2), 301-321.
- McGill, W. J. (1954). Multivariate information transmission. *Psychometrika*, 19, 97-116.
- McNutt, M. (2014). The hunt for MH370. *Science*, 344(6187), 947-947.
- Meseguer, J. (2014). Taming distributed system complexity through formal patterns. *Science of Computer Programming*, 83, 3-34.
- Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63, 81-97.
- Mintzberg, H., Raisinghani, D., & Theoret, A. (1976). The structure of "unstructured" decision processes. *Administrative science quarterly*, 246-275.
- Mooney, R., & Ourston, D. (1989). Induction over the unexplained: Integrated learning of concepts with both explainable and conventional aspects. In *Proceedings of the sixth international workshop on machine learning* (pp. 5-7). Morgan Kaufmann Publishers Inc.
- Moore, M. G. (1989). *Editorial: Three types of interaction*.
- Moskowitz, I.S. & Kang, M.H. (1994) Covert Channels - Here to Stay? *COMPASS'94*, 235-243.
- Moskowitz, I.S., Chang, L.W. & Newman, R.E. (2002) Capacity is the Wrong Paradigm, *NSPW'02*, 114-126.
- Murayama, K., Blake, A. B., Kerr, T., & Castel, A. D. (2016). When enough is not enough: Information overload and metacognitive decisions to stop studying information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42(6), 914.
- Mutch, A. (2002). Actors and networks or agents and structures: towards a realist view of information systems. *Organization*, 9(3), 477-496.
- Nilsson, N. J. (1969). *A mobile automaton: An application of artificial intelligence techniques*. SRI INTERNATIONAL MENLO PARK CA ARTIFICIAL INTELLIGENCE CENTER.
- Nilsson, N. J. (2014). *Principles Of Artificial Intelligence*. Morgan Kaufmann.
- Norman, D. A. (1983). Design rules based on analyses of human error. *Communications of the ACM*, 26(4), 254-258.

- Omohundro, S. M. (2008). "The Basic AI Drives." In *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, edited by Pei Wang, Ben Goertzel, and Stan Franklin, 483–492. Frontiers in Artificial Intelligence and Applications 171. Amsterdam: IOS.
- Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans*, 30(3), 286-297.
- Pincus, S. M. (1991). Approximate entropy as a measure of system complexity. *Proceedings of the National Academy of Sciences*, 88(6) 2297-2301.
- Primiero, G. (2014). A taxonomy of errors for information systems. *Minds and Machines*, 24(3), 249-273.
- Rabinovich, M. I., Simmons, A. N., & Varona, P. (2015). Dynamical bridge between brain and mind. *Trends in cognitive sciences*, 19(8), 453-461.
- Reason, J. (1990). *Human error*. Cambridge university press.
- Reason, J. (1994). Latent errors and systems disasters. In *Social Issues In Computing* (pp. 128-176). McGraw-Hill, Inc.
- Reason, J. (2000). Human error: models and management. *BMJ*, 320(7237), 768-770.
- Rich, C., & Waters, RC. (Eds.). (2014). *Readings in artificial intelligence and software engineering*. Morgan Kaufmann.
- Russell, S. & Norvig, P., (2003). *Artificial Intelligence – A Modern Approach*. New Jersey: Prentice-Hall, 9 (2) (2003)
- Saracevic, T. (2016). The Notion of Relevance in Information Science: Everybody knows what relevance is. But, what is it really? *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 8(3), i-109.
- Schalkoff, R. J. (1992). *Pattern recognition*. John Wiley & Sons, Inc.
- Scheff, T. J. (1963). Decision rules, types of error, and their consequences in medical diagnosis. *Systems Research and Behavioral Science*, 8(2), 97-107.
- Sedig, K., & Parsons, P. (2015, March). Human-Information Interaction—A Special Issue of the Journal of Informatics. In *Informatics* (Vol. 2, No. 1, pp. 1-3). Multidisciplinary Digital Publishing Institute.
- Serruya, M. D. (2015). As we may think and be: brain-computer interfaces to expand the substrate of mind. *Frontiers in systems neuroscience*, 9.
- Shannon C. (1956). "The Bandwagon," Institute of Radio Engineers, *Transactions on Information Theory*, March 1956; IT-2:3
- Shannon, C.E. (1948). "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, July–Oct. 1948.
- Sharma, A., Jagannathan, K., & Varshney, L. R. (2014, June). Information overload and human priority queuing. In *2014 IEEE International Symposium on Information Theory* (pp. 831-835). IEEE.
- Shrobe, HE. (Ed.). (2014). *Exploring artificial intelligence: survey talks from the National Conferences on Artificial Intelligence*. Morgan Kaufmann.
- Simon, H. A. (1960). The new science of management decision.
- Simon, T., & Young, R.M. (1988). GOMS meets STRIPS: The integration of planning with skilled procedure execution in human-computer interaction. In *People and Computers IV: Proceedings of the fourth conference of the British Computer Society Human-Computer Interaction Specialist Group*, University of Manchester, 5–9 September 1988 (pp. 581–594). New York: Cambridge University Press.
- Singhvi A, Russell K. (2016). Inside the self-driving Tesla fatal accident, *The New York Times*. Available at [www.nytimes.com/interactive/2016/07/01/business/inside-tesla-accident.html](http://www.nytimes.com/interactive/2016/07/01/business/inside-tesla-accident.html). Accessed December 2, 2016.
- Skilling, J. (1989). *Maximum entropy and Bayesian methods*. Cambridge England 1988. Dordrecht, the Netherlands: Kluwer.
- Soares, N., & Fallenstein, B. (2014). Aligning superintelligence with human interests: A technical research agenda. *Machine Intelligence Research Institute (MIRI) technical report*, 8.



- Soldatos, J., Kefalakis, N., Hauswirth, M., Serrano, M., Calbimonte, J. P., Riahi, M., Aberer, K., Jayaramen, P.P., Zaslavsky, A., Zarko, I.P., Skorin-Kapov, L. & Herzog, R. (2015). OpenIoT: Open Source Internet-of-Things in the Cloud. In *Interoperability and Open-Source Solutions for the Internet of Things* (pp. 13-25). Springer International Publishing.
- Spier, S. (2016). From Culture Industry to Information Society: How Horkheimer and Adorno's Conception of the Culture Industry Can Help Us Examine Information Overload in the Capitalist Information Society. In *Information Cultures in the Digital Age* (pp. 385-396). Springer Fachmedien Wiesbaden.
- Stonier, T. (2012). *Beyond information: The natural history of intelligence*. Springer Science & Business Media.
- Strong, A. I. (2016). Applications of Artificial Intelligence & Associated Technologies. *Science [ETEBMS-2016]*, 5, 6.
- Toms, E. G. (2002). Information interaction: Providing a framework for information architecture. *Journal of the American Society for Information Science and Technology*, 53(10), 855-862.
- Toms, E.G. (1997). Browsing digital information: examining the "affordances" in the interaction of user and text. Unpublished Ph.D. Dissertation, University of Western Ontario.
- The Tesla Team. (2016). A Tragic Loss, *Tesla Blog*. Available at [www.tesla.com/blog/tragic-loss](http://www.tesla.com/blog/tragic-loss). Accessed January 8, 2017.
- Vasant, P. (2015). *Handbook of Research on Artificial Intelligence Techniques and Algorithms, 2 Volumes*. Information Science Reference-Imprint of: IGI Publishing.
- Wah, BW. (1987). "Guest Editor's Introduction: New Computers for Artificial Intelligence Processing," in *Computer*, vol.20, no.1, pp.10-15, Jan. 1987
- Wiberg, M. (2015). Interaction, New Materials & Computing—Beyond the Disappearing Computer, Towards Material Interactions. *Materials & Design*.

# Chapter 5

## Verification Challenges for Autonomous Systems

Signe A. Redfield and Mae L. Seto

### 5.1 Introduction

Autonomy and artificial intelligence are quite different. Autonomy is the ability of a physically instantiated robot (autonomous system) to make decisions and reason about its actions based on its in-situ sensor measurements. The objective is to adapt to changes in itself or other systems it interacts with, the environment it operates in, or its tasking (mission). Artificial intelligence, in the broad sense, refers to abstract capabilities associated with problem-solving and does not necessarily require a reference to the physical world. An autonomous robot might use artificial intelligence tools to solve its problems but it is grounded in the physical environment it shares with other objects. An artificial intelligence itself might use an autonomous robot to implement a solution it devises or to gather data to solve a problem but it does not have to ground itself in the physical world for this. This chapter addresses challenges in transitioning autonomous robots, enabled with autonomy which may have artificial intelligence, from the laboratory to real-world environments.

Robotics has been a recognized interdisciplinary area of study since the mid-1900s. In the 1970s the first wave of industrial robots went from the research community to the factory floors (Engelberger 1974). These robots were automated. To overcome safety issues due to their sensory and processing limitations, they were segregated from their human co-workers in physical safety cages. Even with relatively predictable controllers governing their actions, it was not possible to verify their safety sufficiently to operate near humans. Today, robot systems are more capable (Miller et al. 2011), complex (Ferri et al. 2016), and thus less comprehensible

---

S.A. Redfield (✉)  
U.S. Naval Research Laboratory, Washington, DC, USA  
e-mail: [signe.redfield@nrl.navy.mil](mailto:signe.redfield@nrl.navy.mil)

M.L. Seto  
Dalhousie University, Halifax, NS, Canada  
e-mail: [mae.seto@dal.ca](mailto:mae.seto@dal.ca)

to non-specialists. Research and development has pushed the boundaries on what autonomy can confer on robots in all environments. However, it has not similarly pushed boundaries for how to certify and assure these robots' functions.

Research addresses user needs at the design stage as motivation for an autonomous robot to address a problem. Aspects peripheral to the problem become a lower priority. However, with increase interest in long duration autonomy (Dunbabin and Marques 2012), complex missions, and driverless cars, one of these peripheral aspects have risen in importance. This is the requirement to verify the safety and bounds on the operational capabilities of autonomous systems.

This chapter introduces autonomy for autonomous systems, verification in general then verification implications for autonomy. Next, verification challenges applicable to most robot operating environments (land, sea, air, and space) are outlined with the simple ground robot as an illustrative example.

## 5.2 Autonomy

Autonomy adds complexity to autonomous systems which adds expense and uncertainty about the system performance, its safety, and when it should be used. Despite this, there are robot situations where autonomy provides a viable solution. These include situations that involve:

- uncertainty about the environment: for example, in rooms, doors may be opened or closed, they can contain people acting within it
- uncertainty about the robot state within the environment: inaccurate or incomplete sensor data on its self-position so that even with a complete map of the environment, the robot cannot navigate to a desired location, and
- communications latency: the robot does not have a human to interpret sensor data or make decisions in new or ambiguous situations.

Autonomy refers to a category of control mechanisms and behaviors (in the context of the behavior-based robot control paradigm) that provides robustness against this uncertainty and enables the robot to operate with little or no human intervention to interpret sensor data or make decisions.

The following terms are used to discuss elements of autonomous systems.

### Definitions

*System*—immobot,<sup>1</sup> robot, group of immobots or group of robots, centralized or decentralized. The hardware, software, perception, actuation, communications, and decision-making that are abstracted as a unit and act in the world. For example: the robot that turns the doorknob is a *system*, but the doorknob is part of the environment rather than the system. A team of robots with a single stationary

---

<sup>1</sup>Immobot—a robot that is not capable of moving from one location to another within its environment but is capable of modifying its environment in some way, e.g. a smart house.

computer acting as a centralized controller includes both the robots and the computer in the system. A smart house is a system but the people inside of it are part of its environment. The user interface may or may not be part of the system but the human using it is not. The terms autonomous system, system, autonomous robot, and robot are used interchangeably here.

*Autonomous system*—a system that makes decisions based on local environmental information and **has an intractably complex interaction with its world (environment)**.

*Behavior*

1. (*robotics*)—the algorithms, software modules and/or actions of a robot in each context (designed and observed)
2. (*verification*)—the actions of a system in an environment

These definitions are unusually specific; a more typical definition of *autonomous* simply means the system makes decisions about its actions based on local environmental data (IEEE Standard Ontologies for Robotics and Automation 2015a, b). Since the focus is systems with no verification tools, the more specific definition for *autonomous* will be used.

Simple systems can fall into the ‘autonomous’ category while at the same time, complex ones may not. For example, robotic arms in a factory have their physical structure and/or environment constrained so the verification problem is tractable. Similarly, their instantiated behaviors are not subject to any constraints so their system architects build the autonomy as they see fit. However, the purpose of this chapter is to identify verification challenges for difficult cases where formal methods-based design tools are, for whatever reasons, not feasible. While there is complexity and cost to autonomy its benefits on-board autonomous systems are notable.

### 5.2.1 Benefits of Autonomy

One reason to deploy a mobile autonomous system for a task is the difficult environment (space, underwater, under-ice, etc.). In dynamic environments, autonomous systems operate with limited human interaction to control complex, real-time, and critical processes over long durations. In addition to enabling operations in adverse environments, autonomy also has the potential for increased capability at reduced operational cost. The number of human operators required, a major cost, is reduced. As well, the reliance on the communications link between the robot(s) and its operators is also reduced. An autonomous system is faster than an operator (especially given latencies due to distance or environment) and can function even when communications with its operator is poor. Communication has a cost (energy, at the very least) and is imperfect as channels can be directionally-dependent, lossy, range dependent, and introduce delays. Autonomy can mitigate some of this compromised

communications. However, another cost of on-board autonomy is in the complexity, reliability, and cost of the verification design and implementation.

Verification addresses whether the autonomy was designed and implemented correctly whereas validation is concerned with whether the autonomy (e.g. a robot behavior) meets its requirements to begin with. Verification is the focus in this chapter. Current verification and validation, or V&V, techniques struggle with existing autonomous systems. For example, in the past, the implementation of embedded systems was conservative and dynamic memory allocation was only permitted at start time. Now, the requirement is to verify and validate autonomous systems that exhibit large sets of interacting behaviors and not all of them deterministic.

Autonomy facilitates the autonomous system adapting to a larger set of situations—not all of which are known at design time. This is a key point as one of the purposes of autonomy is to provide contingencies for situations that cannot be specified precisely a priori. Unfortunately, current verification processes require a complete specification of what the system is expected to do in all situations.

Analytic V&V techniques, and model checking, in particular, can provide solutions to design autonomous system control agents in a more efficient and reliable manner. This can mean earlier error detection and a more thorough search of the space spanned by all performance possibilities (performance space). However, the most suitable V&V approach depends on the autonomy tools used. In addition to purely reactive tools, these can include:

- planners
- executives
- fault detection isolation and recovery (FDIR) systems
- mission-based measurements
- navigation
  - terrain analysis and modeling
  - localization
  - path-planning.

It is expected that autonomy approaches require both verification techniques specific to the approach and those that apply across autonomous systems.

### 5.3 Verification

Verification tools build an assurance case, a body of evidence that, connected using provably correct assertions, enables one to say, within defined assumptions, that the system has certain properties. These properties define *what is desired* and can involve either safety or security. For autonomous systems, there are three categories of safety: self; objects and people it expects to interact with, and objects and people it is not intended to interact with.

## Definitions

*Verification and validation process:* Firstly, *validate* the match between the purpose for which the system is designed and the system requirements (and presumably generate a model of the system or design a potential solution). Secondly, *verify* that the model/design meets the requirements (works correctly). Third, *validate* the system to be sure it accomplishes the purpose for which it was designed (does what the user needs).

*Verification:* The process of gaining confidence in the correctness of a design or ensuring that a system has certain properties. The fact that it may require exhaustive proof is a problem associated with verification of autonomous systems.

*Validation:* This refers to step three in the process above. This is the process of testing to ensure that the design decisions made in the initial validation process, to match purpose to requirements, are correct in terms of the system end-use.

Verification is particularly important as systems transition from the research laboratory because it is a critical element of the certification and accreditation process which gives credibility with potential users. Within research laboratories, verification is important because it enables other researchers to use a given algorithm as a component of their systems without concern about unexplored failure modes. For example, if the objective is to test a robotic path-planner around obstacles, the user wants the robot's obstacle avoidance algorithm to be solid and well-understood. Verification confirms the circumstances under which the obstacle avoidance algorithm fails as well as provides a methodology to assess the merit of the user's path-planner with the integrated obstacle avoidance algorithm. Given that, what are the verification implications of autonomy?

### 5.3.1 Verification Implications of Autonomy

As one of the verification objectives is to understand what the autonomous system is supposed to do, verification tools assume a system specification exists. However, defining the operational goals of an autonomous system is quite difficult making its verification difficult. Existing research addresses these issues, but there are more unexplored research challenges than there are underway research efforts. Section 5.4 identifies autonomous systems verification challenges and notes those with ongoing research efforts. Specific problems that verification tools like sequence and scenario-based testing could address are described next along with their limitations.

Traditional flight software on unmanned aerial and space systems have two components: the on-board software and a sequence. The on-board software is low-level methods or procedures for commanding aspects of the spacecraft hardware, while the sequence is a time-ordered list of commands where each command maps to a software method. Each command is tested independently. Traditional V&V flight

software on unmanned aerial and space systems achieve verification confidence from testing each sequence before up-linking and executing.

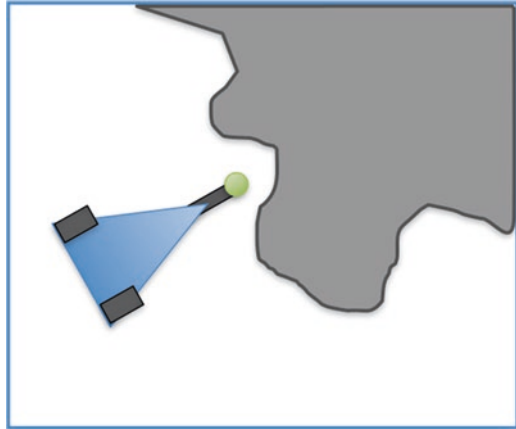
For these cases, potentially unanticipated command interactions and subtle problems are detectable by sequence testing. Sequences have singular execution paths (or at most, a few), which facilitates detailed testing and interaction analysis to focus on just those paths. This is a powerful approach but is only flexible when there are a small number of execution paths and those paths are known in advance. On the other hand, the autonomy of autonomous systems may be parallelized, distributed, and non-deterministic with interactions between commands. Consequently, V&V sequence testing does not work as well with these systems.

Autonomous systems are commanded by high-level goals or autonomic responses rather than explicitly scripted sequences. If the system is controlled by high-level goals, these goals are interpreted and further broken down into lower-level commands. The autonomy planner determines the lower-level actions for the robot to take to achieve its high-level goals. If problems arise during execution the autonomy could take corrective actions and find alternate ways to achieve the goals. In this way, there are many possible execution paths. Thus, it is impossible to identify the few that are actually needed and to exhaustively test those. Additionally, the exact command sequence cannot be predicted in advance without complete environmental knowledge, as the decisions are based on both on-board events and local environmental conditions. Autonomy's value is in its ability to close control loops on-board the robot instead of through human operators. However, strength also makes it challenging to adequately define the behavior specification. Consequently, this means sequence validation approaches do not work as autonomy driven processes are not necessarily sequential or deterministic. As well, the autonomy could be implemented as multiple parallel threads that interact.

In an autonomous system, even a simple one, sequence testing provides some confidence for each of the commands, but it does not address interactions between commands the way it has for scripted flight software sequences. These interactions can be subtle and their results, unexpected. They can also depend on the exact sequencing and timing of prior commands, subtleties of the robot state, the environment it interacts with, etc.

As autonomous systems close control loops and arbitrate resources on-board with specialized reasoning, the range of possible situations becomes exponentially large and is largely inaccessible to the operator. This renders traditional scenario-based testing inefficient and in many cases, is impossible to perform exhaustively. There are also scenarios that cannot predictably occur or deterministically reproduced in regular testing. They include race conditions, omissions, and deadlock. Omissions in the specification are problems like defining how long an agent should wait for a reply to a service request (within the publish-subscribe architecture assumed here) before timing out or pursuing another action. Deadlock occurs when two elements enter an infinite loop in the task allocation process and the process fails to yield a decision. This can happen as a result of a race condition or an unforeseen interaction between software components.

**Fig. 5.1** Example robot system—triangular robot (*blue*) with three wheels (*dark gray*) and a downward looking sensor (*green*) in a flat environment (*white*) with cliffs (*light gray*)



A race condition is a behavior in electronics, software, or other system element where the output is sensitive to the sequence, timing, or order of other uncontrollable events. It creates a bug when events do not happen in the order intended. Race conditions can occur in electronics, especially logic circuits, and in multi-threaded or distributed software. An autonomous system can have all of these and is thus prone to race conditions.

Testing for race conditions is not straightforward since certain timing conditions must be satisfied for them to occur and these conditions may not manifest during regular testing. Because of this, regular testing alone cannot assure that race conditions do not exist. To determine whether race conditions exist, formal methods (not discussed here) must be used to model the interaction between agents/threads/subsystems.

In the detailed analysis of verification challenges with autonomous systems, it is instructive to have an illustrative example system.

### 5.3.2 Example System

The simple robot example, shown in cartoon form in Fig. 5.1, serves to illustrate subtleties that drive the variety of tools and research gaps that exemplify these problems.

This toy system consists of a triangular robot with three wheels and a downward-looking range sensor on a forward telescopic pole to detect cliffs (stairs). The two rear wheels drive the robot. As the robot moves, it controls how far the downward-looking sensor is extended in front by extending or retracting the telescopic pole. Since the robot is physically instantiated it has a non-zero stopping distance. Extending the sensor pole further out allows the robot to detect cliffs earlier. This means it could travel at a higher forward speed. The robot uses dead-reckoning against an internal map to navigate to a waypoint and the downward-looking sensor



to avoid hazards on the way. It operates in a flat world with cliffs but no walls and its only task is to travel from one waypoint to another.

Though this is a simple robot, it provides context to demonstrate some of the challenges associated with autonomous system verification which is discussed next.

## 5.4 Challenges

The following four categories of research challenges in autonomous systems verification are identified as follows:

- *models*: development of models that represent the system,
- *abstraction*: how to determine the adequate level of abstraction and detail to which requirements are generated,
- *testing*: development of test scenarios, metrics and performance evaluation mechanisms; and the extension of simulations, test environments and tools to enable generalization from tests to conclusions about system performance, and
- *tools*: new tools or techniques that must be developed to enable autonomous systems verification.

The rest of this chapter introduces these challenges in more detail.

### 5.4.1 Models

With models, there are four identified challenges associated with how to model the autonomous system and the environment it operates in.

*Challenge 1: How is an adequate model of the system created?*

There are several types of models relevant to the verification problem. They include logical models that represent the desired or computed behavior, probabilistic models of the behavior, and mathematical and statistical models of the system. These models must be at a fidelity that captures interactions with the environment and predicts system performance. Software tools such as PRISM can verify behaviors that can be modeled probabilistically (Chaki and Giampapa 2013), but deriving these models and ensuring they represent the behavior is difficult, especially when the verification needs to generalize across environments. Estimations of the conditional probabilities in the model are difficult to arrive at when realistic environmental interactions are considered.

*Challenge 2: Common models and frameworks need to describe autonomous systems broadly enough so they can be used to standardize evaluation efforts and interfaces to the system.*

Beyond models that support verification for systems, models and frameworks (Challenge 1) that support evaluations across solutions are also needed. Such common models and frameworks are being developed from different perspectives. They

range from the development of ontologies and shared concepts describing autonomous systems (Paull et al. 2012) to architectural designs to mathematical models of robot behavior based on dynamical systems. External analysis tools that use variables as independent elements to characterize the system are generally inadequate. Dynamical systems approaches (Smithers 1995) attempt to produce more generalizable models using ordinary differential equations (ODEs) that include nonlinearities due to time dependencies. However, these approaches, while able to describe some long term emergent behaviors, are not applicable to behaviors that are not modeled by ODEs. Developing models based on tools that measure differences across solutions and how to define a model type that supports evaluations and generalizes across solutions are unsolved problems.

In the toy problem, if the robot's high-level goals are broken down based on a framework like the OODA loop (observe, orient, decide, act), actions for the robot can be specified. Imposing this structure on the autonomous system ensures consistency between evaluation efforts and output standardization. However, a controller may or may not map well into that framework. A deliberative system might explicitly follow each step, while a reactive controller will not explicitly instantiate the 'decide' or 'observe' steps. In the reactive approach, the functions provided by the "decide" and "orient" steps are implicit in the "observe" and "act" steps and cannot be separated. This introduces problems when the framework is used to standardize evaluation efforts, since inaccuracies in the representation can lead to errors in the analysis. If the robot is not deciding, but is instead simply observing and acting, then verification tools designed to analyze the decision-making stage may not adequately capture the relationship between the sensors, actuators, and environment.

*Challenge 3: How should models of black box autonomous systems be developed and debugged? How is a mathematical and/or logical model suitable for formal analysis produced from empirical observations?*

*Challenge 4: How should one identify and model components that are not captured yet (and what are their properties)?*

When there is insufficient knowledge about a system to represent its autonomous behaviors with either logical, probabilistic or mathematical models, it is treated as a black box. Consequently, determining the level of abstraction is almost impossible. In that case, would observing the system's behavior yield sufficient insight into the level of abstraction to model the sensor data? For the example robot, is it sufficient to model the sensor output as a binary (floor/cliff) detection? Or, should the sensor output be modeled as a discrete or continuous-valued function describing the distance between the sensor and the closest object? Should noise, manifested as sensor variations or the frequency which transitions between the binary states occur, be included? Do the motor controllers need to be modeled or is it sufficient to generate a larger model of system actions as a function of sensor input? What principles are used to design a simulation or model, at the level of abstraction needed, to evaluate the feature or property of interest?

## 5.4.2 Abstraction

### 5.4.2.1 Fidelity

These challenges focus on the simulation fidelity rather than only the system model addressed in Challenge 1.

*Challenge 5: What determines the level of simulator fidelity to extract the information of interest?*

Insight into the fidelity a simulator requires for meaningful results makes it possible to identify scenarios where the system fails. Searches for scenarios where the system fails can be automated by developing adaptive learning techniques to focus simulations in performance space regions where failure is suspected (Schultz et al. 1993). However, these techniques are only partially effective. Along with development and tuning of learning algorithms, appropriate performance metrics to drive the learning process are needed. These learning techniques and performance metrics could also be used to identify which of several potential levels of fidelity capture the most failure modes.

*Challenge 6: How is the level of abstraction determined for the robot model, its behaviors, and the simulation that tests the model? How many environmental characteristics need to be specified? What are the aspects of the environment, the robot, and the autonomy algorithms that cannot be abstracted away without undermining the verification?*

The level of fidelity to model aspects of the environment as well as which aspects should be modeled is unclear.

The model of the autonomous behavior is given. But what is the fidelity of the model for the robot hardware that realizes the behavior? Can friction in the motors be abstracted away? What about other interacting behaviors in the system?

If a path-planning behavior is to be tested, the robot relies on an awareness of its position relative to the desired path or destination. What level of abstraction is adequate to capture that information? When that is known then the level of abstraction for the environment could be addressed.

For the example system, is it sufficient to define an environment “that contains cliffs”? Reaching the given destination implies the robot did not run out of power prematurely. Not falling off cliffs is easier if the cliffs are stair-like, rather than peninsular, since the robot has only one sensor and thus one measurement of cliff location at any time. The orientation of the robot to cliffs it might encounter or whether the road surface approaching a cliff impacts the robot’s maneuverability is unknown. How could one verify the robot will be safe (i.e. not fall off a cliff) given its existing behaviors or determine the environment state space boundaries where the robot can be verified safe? Are there other aspects of the environment that affect the robot’s performance that should be included in the environmental model or the robot’s behavior model?

The task can be constrained so the robot only operates in an environment with stairs—not peninsular cliffs. Modeling the environment as stairs that are

perpendicular or parallel to the robot's travel direction is insufficient. However, including all possible orientations does not scale up to handle more complex environments. One cannot abstract away stair orientation if the objective is to characterize and model the robot's behavior near stairs. However, the sensor uses sound to detect its range to the floor so it is fine to abstract away the stair's color. The width of the stairs may affect the robot's ability to reach its destination before it runs out of power. Can the width of the tested stairs be bound?

Since the sensor is centered in front of the robot, some autonomous behaviors are likely to have a fault mode where the robot is approaching stairs at an acute angle. How would other locations in the robot state space, which may be fault modes, be identified? These fault modes are functions of the robot's physical configuration. For example, the separation of the rear wheels affects the angle when the robot falls off the cliff before it senses it.

#### 5.4.2.2 Requirements Generation

*Challenge 7: Where is the transition from specifying system requirements to designing the system and how are principled requirements developed so they do not devolve into designing the solution?*

There are efforts towards requirements generation for autonomous systems (Vassev and Hinchey 2013), but they apply to space missions and highlight a problem with defining requirements for autonomous systems: defining the requirements often results in designing the system.

This is particularly noticeable in systems engineering requirements generation. Within the DoD Systems Engineering Fundamentals text (Defense Acquisition University Press 2001), IEEE Standard P1220 is quoted as defining a set of 15 tasks in the requirements generation process. Of these 15 tasks, one represents the desired capabilities of the system (the *functional requirements* which define and constrain the autonomy), one consists largely of elements that an autonomy designer would expect to be part of the design process (the *modes of operation*), three are currently unsolved research problems due to the inability to adequately define, in a testable and achievable way, what the robot ought to be doing (the *measures of effectiveness and suitability*, the *utilization environments*, and the *performance requirements*), and the rest define the context the autonomy is expected to operate. While they impose requirements on the autonomy, these additional constraints are not autonomy requirements themselves. In exploring the functional requirements generation process one finds the functional analysis stage encompasses the autonomy design process.

With the example system, the high-level requirement might be “the robot shall successfully reach its destination in an environment that contains cliffs”. But even simply specifying the lower level requirements becomes rapidly difficult.

If a behavior is specified for the robot when it detects a cliff, it defines the system autonomy, not a functional or safety requirement. Sub-requirements of “the robot shall not fall off cliffs”, “the robot shall reach its destination” and even “the robot

shall not take longer than X to reach its destination” could be specified, and have the autonomy balance the competing requirements. However, defining requirements below this level, again, quickly falls into designing the autonomy.

Different structures and models have been proposed to describe autonomous systems but none are widely accepted. There is no common taxonomy of goals that describes robot behavior and the goals of these systems. Typically, there are two ways these develop in a field as it matures: either they develop organically over a long period as different ideas are proposed, tested, and refined or rejected or, an organizing body selects or generates a standard to support and accelerate its acceptance.

Standards are being defined to support descriptions of both the hardware (IEEE Standard Ontologies for Robotics and Automation 2015a, b) and the tasks the systems are expected to accomplish, but this field is quite broad and there is little consensus on what tasks should be included or how they should be described to support their eventual implementation and use.

*Challenge 8: How is it ensured that the implicit and the explicit goals of the system are captured? How is a model of the system goals from a human understanding of the task goals, the system, and the environment created?*

To verify the system, implicit goals must also be captured in addition to the explicitly defined task goals. If the explicit goal is for the robot to gather information about a region, the implicit goal is to ensure the information returns to the operator. If the robot gathers the information but does not return it to the user, then as far as the user is concerned, the information has not been gathered.

*Challenge 9: How are performance, safety, and security considerations integrated?*

In the certification and accreditation communities, performance, safety and security considerations are separated. The safety ones are addressed by the certification authority and the performance and security ones by the accreditation authority. One of the major reasons autonomy is implemented on a system is to provide an extra layer of assurance for safety and security as the robot attempts to meet its performance requirements.

For the toy robot, safety includes “not falling off a cliff”. If its task is “get from point A to point B”, system safety is an implicit performance requirement, since falling off a cliff prevents the robot from reaching point B. If cliff locations are completely known, autonomy is not needed as the solution is to automate the optimal paths between a variety of A’s and B’s to avoid cliffs. Autonomy is needed if the cliff locations relative to the robot’s actual path are not completely known and the desire is to react safely if it detects one. Safety is one of the reasons autonomy is in a system, and being able to avoid cliffs increases overall performance. In this case, safety is part of performance. If other aspects of safety are considered then, safety would include potential damage to the environment as a side effect of falling off the cliff (environmental safety) and potentially injured bystanders if the robot drives into or over a bystander’s foot (bystander safety). The safety of the operator is not a consideration for this robot since the operator’s interaction with the robot is minimal.

While these aspects do not directly affect the system performance, they do interact with its algorithms—an obstacle avoidance algorithm protects both bystanders and the robot itself while significantly affecting its ability to accomplish its task. Avoiding cliffs promotes safety for the robot and for the environment while improving its performance.

### 5.4.3 Test

*Challenge 10: At what point is there enough evidence to determine that an autonomous system or behavior has been verified?*

Even outside the robotics research community, the actual measures used to determine when enough evidence has accumulated to verify a system are often ad hoc or arbitrary. Since the original metrics lack firm principles or foundations, it is impossible to find a principled mechanism to support extending them to include autonomous systems. Just as there are no principled methods to determine what evidence is appropriate, there is no easy way to determine when sufficient evidence has been accumulated.

*Challenge 11: How does one ensure it is possible, in the physical world, to test simulated situations that result in boundary cases?*

This is a problem when a fault mode is triggered by a specific sequence of previous actions. There are trivial examples where the toy robot falls off a cliff if initialized in an untenable location, but setting up a physical environment where the robot will be induced to perform the same sequence of actions that lead to that failure is non-trivial. Without being able to repeatedly trigger the same failure in the physical world, there is no confidence that an applied remedy would be effective.

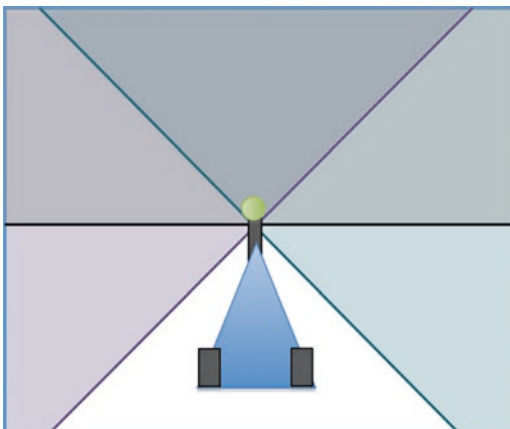
#### 5.4.3.1 Scenarios

*Challenge 12: How would tests be designed so that passing them indicates a more general capability?*

NIST (National Institute of Standards and Technology) developed a suite of robot test arenas in their work with first responders and competition developers in urban search and rescue tasks (Jacoff et al. 2010). In this approach to capability evaluation, systems are tested on their ability to repeatedly accomplish a task within a given arena. Performance is based not only on simple binary accomplishment metrics but on reliability and robustness. This is an extreme version of the most common method developers use to engender confidence in their systems: ad-hoc execution of select design reference missions. Instead of developing an entire scenario, the NIST approach allows developers to test their systems on one capability at a time.

It is more common for developers to test their systems against the entire mission it was designed to address. The mission is intuitively representative of a use case for

**Fig. 5.2** Generalization of environments: if the robot can avoid cliffs that are straight and appear perpendicular (*gray*), at  $+45^\circ$  (*purple*) or at  $-45^\circ$  (*teal*), it is not imply it can avoid cliffs that cut through the *white* region



which the system was designed. In the best case, it would be a particularly challenging instance of the use case. The implication is that since the system can handle the demonstrated case it will also be able to handle other, similar cases the system will be subjected to operationally.

As an example with the toy robot, if the robot can avoid straight line precipices that are perpendicular, and at  $45^\circ$  (purple and teal lines in Fig. 5.2) to its direction of travel (black line in Fig. 5.2) then it is valid to generalize and assume all orientations between perpendicular and  $45^\circ$  are likewise acceptable, as are orientations between perpendicular and  $-45^\circ$  (all shaded regions in Fig. 5.2). However, it does not imply whether it will succeed against other cliff orientations (areas that cut through the white region in Fig. 5.2), other than to recognize that there is a point where it will lose stability before it detects the cliff.

While test cases demonstrate possibilities the challenge that autonomous robotics now faces is to produce test schemes that provide results which can be meaningfully generalized not only for specific capabilities but for system performance. Efforts have been made to address this challenge using automation to simply execute and analyze many scenarios (Bartrop et al. 2008; Smith et al. 1999; Pecheur 2000), but in each case these efforts required insight into the system under test, and the automation was still based on scripted scenarios that engineers deemed likely and not unanticipated ones.

*Challenge 13: How are challenging design reference missions selected so that performing well against them indicates a more general capability for the system rather than for specific system components?*

Even after generalizing from specific scenarios to regions of capability within the robot state space, methods are still needed to identify specific scenarios that provide the most general information.

In the toy example the environment was implicitly limited to only straight-line cliffs and perfect sensor or actuator performance between  $\pm 45^\circ$ . There was no mention of unexpected obstacles or materials and conditions that cause errant sensor readings—all of which are sources of undesirable behaviors in autonomous systems.

For the toy example, a simpler scenario with cliffs oriented only perpendicular to the robot travel direction and no obstacles provides less information about the behavior robustness than a scenario with approaches to cliffs over a range of orientations and moving objects in the environment.

*Challenge 14: How can test scenarios be produced to yield the data required to generate mathematical/logical models or to find the boundary conditions and fault locations in the robot state space?*

This challenge addresses two points: (1) the development of test protocols and methodologies whose goals are not to evaluate the system but to generate a model of the system from external observations of system properties (flip side of Challenges 3 and 4), and (2) identify test scenarios in the robot state space that exist at performance boundaries. For example, the edge of the curb running along a sidewalk is a performance boundary for a robot that operates on sidewalks. In a simulation, this might be represented as a distance-from-sidewalk-edge parameter or as a position-on-the-map environmental feature, but in either case, it is a performance boundary.

Instead of test scenarios that characterize how well the system works, the purpose of these scenarios is to highlight both areas where the system fails to perform as expected and areas where there is a transition from one performance regime to another. As well, what techniques are needed to determine the parts of the performance space that should be characterized in a model of the black box autonomous system?

### 5.4.3.2 Metrics and Performance Evaluation

To evaluate the autonomous system using abstracted models, metrics, and measures that are proxies for the system, goals need to be defined. In some cases, these may be represented in the requirements, but the problem of defining metrics associated with implicit and less tangible goals is still difficult. Most work in this area focused on developing tools to measure the degree of autonomy in a system, rather than the effectiveness of the autonomous system as it attempts to accomplish its tasks.

*Challenge 15: Once an adequate model is created how is it determined whether all resulting emergent behaviors were captured and what are appropriate performance measurement tools for this?*

Most formal attempts to provide standards for autonomy have centered on the definition of “levels of autonomy”, an effort to distinguish systems by their degree of independence from a human operator and level of capability. Examples include Draper Labs’ 3D Intelligence Space (Cleary et al. 2001), the US Army’s Mission Performance Potential (MPP) scale (Durst et al. 2014), the Autonomous Control Levels (ACL) put forth by Air Force Research Labs (Sholes 2007), and the National Institute of Standards and Technology’s (NIST) Autonomous Levels for Unmanned Systems (ALFUS) (McWilliams et al. 2007; Huang et al. 2004, 2005), shown in Fig. 5.3.

Not only are different approaches largely incompatible with each other, even experts disagree on the taxonomy to categorize a system within a given approach.



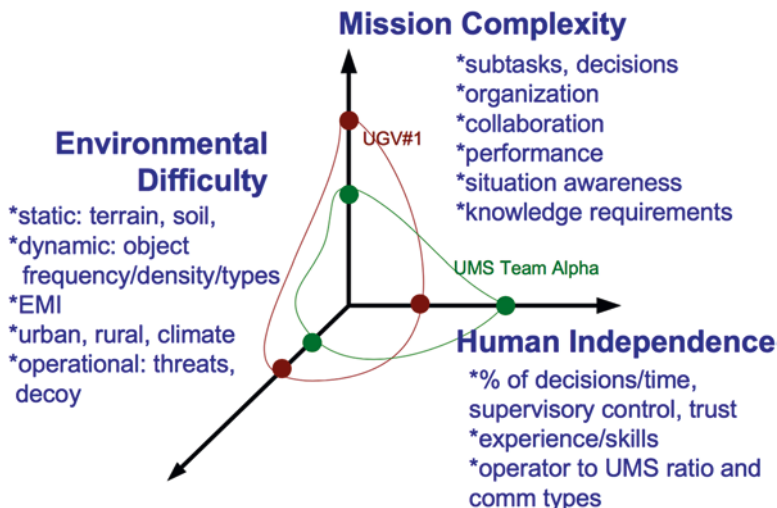


Fig. 5.3 The ALFUS framework for autonomy levels (Huang et al. 2005)

The overall effort to define levels of autonomy has devolved into a philosophical argument, and a 2012 Defense Science Board report recommended that the effort be abandoned in favor of the definition of system frameworks (Murphy and Shields 2012). The levels of autonomy used to certify autonomous systems are the exception because they only attempt to define specific characteristics relevant to their domain of interest.

To illustrate the difficulty in applying a subjective measure of autonomy to a robot, consider the toy example against the two of the axes of evaluation common to most levels of autonomy (Fig. 5.3)—situational awareness/environmental complexity and decision making/human independence. One might argue that the situational awareness of the robot is limited because it is only able to sense its own location and any cliffs in its immediate vicinity. However, it can also be argued that, for the intended environment (only has cliffs), this is all it needs, and the situational awareness is therefore, high. Likewise, since the only stipulated ability is to navigate to a destination and avoid cliffs, in the space spanned by all possible behaviors for all robots this is limited in its independence capability. On the other hand, waypoint following allows the robot to operate independent of a human operator in the traversal of those waypoints, so it could also be considered highly independent.

IEEE Standard 1872–2015 (IEEE Standard Ontologies for Robotics and Automation 2015a, b) attempts to introduce clarity by defining autonomy as a role taken on by a process within the robot. Instead of attempting to define the system autonomy it allows the designer to make some aspects autonomous (e.g. avoiding cliffs) and others automatic (e.g. following a fixed sequence of waypoints).

*Challenge 16: Measurement and evaluation are generally poorly understood—operators can describe tasks for the robot but lack tools to quantitatively evaluate them. How should autonomous behaviors be*

*measured so they consistently and accurately describe the capability embodied by a robot?*

Efforts to create metrics generally result in tools with solid theoretical foundations that are not easily implemented or, focus on subjective evaluations from experts like system users (Steinberg 2006; Durst et al. 2014), and consequently, not easily comparable across experts. Standardized questions and forms use scales and techniques in an attempt to normalize subjective results (Billman and Steinberg 2007). However, the problem is the inability of evaluators to agree on the relative importance of subsidiary details within a task (e.g. whether the smoothness or the directness of the robot's path is more important) rather than the adequacy of the evaluation tools.

*Challenge 17: How is a metric defined for comparing solutions?*

Even if a metric is defined to evaluate whether a robot accomplishes its task, how are the different solutions compared? Start with the toy problem task: reach waypoint B by a given time. Two robots have the downward-looking sensor that identifies cliffs. Robot A has a sensor that tells it range and bearing to the waypoint and Robot B has a map and the ability to localize itself within it. Robot A uses simple heuristics that cause it to head straight towards the waypoint when there are no cliffs and to back up and turn a fixed amount when there is a cliff. Robot B has a more sophisticated behavior to characterize the cliff. Robot A has a motor controller that imposes smoother motion, while robot B's controller can stop abruptly and turn about an axis internal to itself. Would the metric to compare the solutions be based on how fast the robots reach the waypoint, or is it a function of the smoothness of the path? Is it a combination? If it is a combination, how are the metrics weighted? Is it measured with the same start and destination point every time or, is it sufficient to measure multiple random samples or, is the metric a function of path properties and independent of the specific path? Is the metric a simple binary of reached/failed to reach the waypoint? What if the user does not appreciate what the important aspects are? For example, the relative importance of path duration and efficiency or the reliability with which it reaches the destination.

*Challenge 18: How is the "optimal" defined against which the verification is performed? How is the solution shown to be in fact, optimal? How is the performance of the system measured?*

To verify a system one needs to know its properties and what it is supposed to do. The "optimal" solution for verification of autonomous systems can be a computably optimal solution to the problem (though the robot's limitations, environment, or practical considerations may drive it to a less optimal solution) or the desired behavior itself. The difference between the optimal solution and the robot's actual behavior can be a measure of system performance and used to compare against different solutions. Where a computable optimal solution exists, it is possible to determine whether the robot's performance was optimal, but in other cases, performance is more difficult to quantify. The problem is twofold. It is necessary to define what the behavior ought to be, and evaluate that against what the behavior actually is. Optimal can be defined in the context of the specific behavior (the best this robot is capable

of) or the task goals (what is the best that a system could do). General purpose metrics that compare behaviors using either is an active research area.

### 5.4.3.3 Intersection of Scenarios and Metrics

*Challenge 19: How is the performance from finite samples of the performance space generalized across several variables and parameters?*

This is similar to Challenge 12 (how to select tests that indicate a general capability), but the focus of this challenge is how to generalize performance given finite test results. For the toy robot, it is straightforward to generalize from the  $\pm 45$  degree tests because the properties of the environment (cliff orientations), robot (nose sensor and wheel locations), and behavior (robot will not move outside the white triangle in Fig. 5.2 when it reacts to a cliff) are known. What is lacking are general principles, best practices, or theoretical structures that help determine how a given performance test result generalizes its performance throughout the robot state space. For example, if the width of the toy robot's wheelbase is changed, the limits on safe orientations to the cliffs changes. However, within this task and robot configuration, the general premise that the physical configuration is related to this aspect of performance holds. How are equivalent premises that hold for other tasks and scenarios determined?

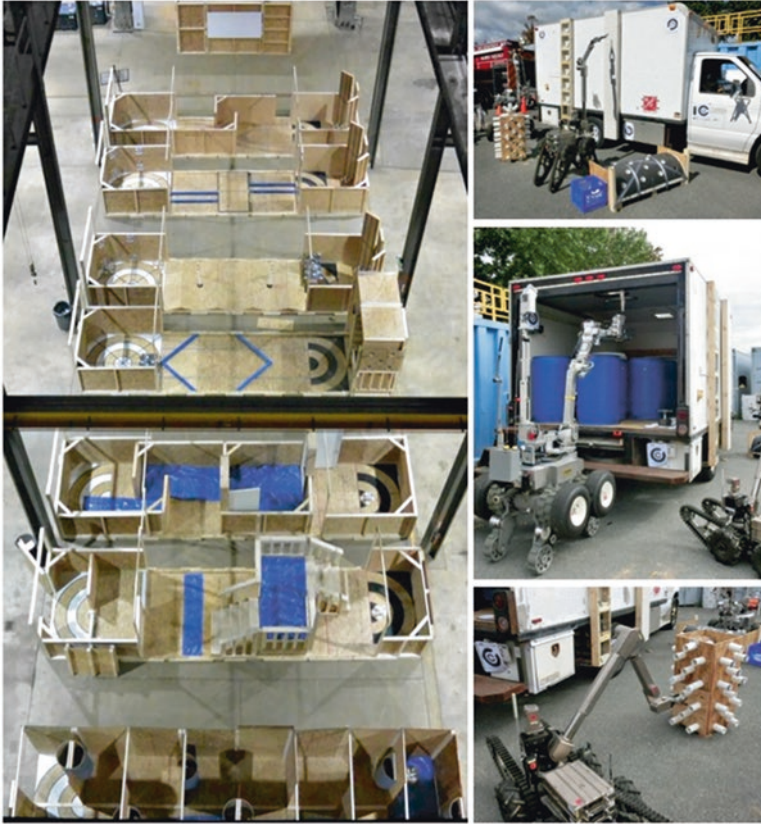
*Challenge 20: Autonomy frameworks are unable to determine whether all the resulting emergent behaviors have been captured or to supply performance measurement tools.*

Even if it is possible to generalize performance samples to a range within a performance regime, the specificity of the samples limits their broader applicability, and thus does not address verification methods for entire systems. As a test scenario is designed to demonstrate one system behavior or feature, others may be simplified or ignored, which limits the broader applicability of the result. The test scenario only captures potential emergent behaviors related to the behavior or feature being tested, not emergent behaviors that occur when multiple behaviors or features interact. Autonomy frameworks define how interactions between the behaviors and functions of the robot are structured, but they do not define how to generate scenarios to measure the reliability of those interactions. Since the interactions between the robot and its environment is intractably complex this is a critical component of any test method since the performance space cannot be exhaustively searched.

### 5.4.4 Tools

*Challenge 21: What new tools or techniques need to be developed?*

If there are solutions to these challenges, an adequate system model and a reasonable abstraction of the environment for the simulation tools, there are still difficulties. In addition to correct-by-construction tools (Kress-Gazit et al. 1989), other



**Fig. 5.4** NIST test arenas for urban search and rescue robots (from NIST’s Robotics Test Facility website)

tools are required to: support analysis of black box systems and behaviors; categorize tasks and goals and connect them to platforms, environments, and capabilities; develop performance metrics; support the development and analysis of modelling and simulation approaches, and connect the different approaches to verification into coherence assurance cases. This is an incomplete list—as progress is made on the rest of these challenges, other gaps will be identified.

*Challenge 22: In general, how is the fitness of a physical robot structure for a given task or environment verified (e.g., a robot that cannot sense color or operates in the dark with an infrared sensor is unfit to sort objects based on color)?*

The NIST test arenas (Jacoff et al. 2010) shown in Fig. 5.4 use specific low-level capabilities (or skills) that can be combined to characterize a desired higher-level capability. The capabilities are determined against performance in a test. For example, the robot must manipulate objects with a required robustness in test X and reliably maneuver through environment Y.

Individual robots are tested against the full suite of test arenas and ranked per their performance in various categories (e.g., manipulation or maneuverability). As robots are generally specialized for a given task, once that task has been decomposed into the skills (or arenas) required, the suitability of a given robot for that task can be evaluated.

Although this approach is effective within this task domain, it has two major limitations: the process to define the arenas/skills was lengthy and expensive and the process to decompose tasks into amalgamations of skills is human-centric and does not generalize well from one task to another. Equivalent tests for the low-level skills could be developed so that any robot might be able to express. However, determining a complete set of basic capabilities was established and that each was adequately tested is more difficult.

*Challenge 23: Descriptive frameworks are either too specific and constrain the developer to specific tools when designing the autonomous elements of the system or, too broad and difficult to apply to specific cases. Tools are needed to analyze systems at both the specific and the broad levels.*

A variety of descriptive frameworks have been advanced to describe autonomous systems in a way that facilitates evaluation. However, when the framework follows too closely to a particular implementation, the solution is limited to only systems with that same implementation.

An example of this phenomenon is the application of formal methods to autonomy by simplifying system states and inputs to create a deterministic system. While this provides a verifiable solution, the simplifications limit the system, and the requirement for determinism precludes the use of more innovative techniques such as fuzzy logic, neural nets, and genetic algorithms. Broader models are more widely applicable, such as the classic OODA (observe, orient, decide, act) loop (Gehr 2009), but the lack of specificity makes them difficult to meaningfully apply. Attempts to find a middle ground between these two approaches include the Systems Capabilities Technical Reference Model (SC-TRM) (Castillo-Effen and Visnevski 2009), Framework for the Design and Evaluation of Autonomous Systems (Murphy and Shields 2012), and the 4D Realtime Control System (RCS) Reference Model Architecture (Albus et al. 2000). Each of these frameworks has its supporters and detractors, but no critical advantage has yet pushed one to widespread adoption. Once such a model is found, verification techniques can be developed for classes of components or capabilities rather than for the entire system at once, making the problem more tractable.

*Challenge 24: How is a structured process that allows feedback between the physical/ground truth layer and the formal methods verification tools developed?*

This is a specific tool among many that could be developed for Challenge 21. Formal methods verification tools can provide useful information about guarantees and properties of a given behavior. However, to verify the behavior as instantiated in a physical system, tools are required to enable test results in the physical system

to feed back into the formal verification tools. Then, this enables the formal verification tool results to feed back into the physical system.

*Challenge 25: How to disambiguate between cases where the specification was incorrect (task description abstraction failed to capture a required system action) from those where the environmental model was incorrect (environmental abstraction failed to capture some critical system-environment interaction)? How to identify not just individual situations but classes of situations where the robot fails to be safe or to achieve safe operation (e.g. a front wheel often falls off the cliff but the back wheels never do). How should unanticipated unknowns be accommodated?*

*Challenge 26: If an algorithm, or patch to an existing algorithm, was replaced can it be proven that no new failure modes were introduced without re-doing the entire verification process?*

If the problem to characterize the performance space of the original system was solved, is it possible to characterize the performance space of the new system without running every algorithm and system-level verification test again?

Does making an autonomous system modular reduce the verification burden when an autonomy algorithm component is changed, added or removed?

In the simulation tool the toy robot model sometimes falls off peninsular cliffs. If the robot is not intended to succeed against them, is this happening because the verification failed to realize that peninsular cliffs were not part of the task or because the simulation environment includes physically unrealizable cliffs?

It is important to evaluate the system at different levels of fidelity using analytic, simulation, and physical instantiation. The analytic tools provide confidence the robot will operate well in certain scenarios and poorly in others. The simulation tools, if abstracted to an appropriate level, can run sufficiently quickly to verify the analytic results in the good and poor areas and identify commonalities between failure modes for the boundary regions. The physical testing tools provide a means to explore the impact of the environment and robot physical structure on its performance in those boundary cases.

The two key challenges identified in testing methodology stem from the intractable complexity problem of operating a complex system within a generally unbounded environment. Firstly, how can all possible scenarios be meaningfully sampled to create a representative subset? Secondly, how can these subsets be generalized to provide confidence in cases that were not directly tested?

These challenges focus on aspects of the problem that are the most difficult to address. Autonomous systems are used in dynamic environments which are inherently unpredictable. Bounds or classes of situations can be defined within which the system is expected to operate a priori. The problem identified here is to define classes of situations within which the system demonstrates specific predictable properties. What tools could be developed to examine a large set of test or simulation data and then extract the common feature that is predictive of success or failure, safety or danger? Can tools be created to identify aspects of the environment which were thought irrelevant but are actually important?

## 5.5 Summary

Within this chapter some pressing verification challenges facing autonomous robotics were identified as important as robots make the transition from the research laboratory to real-world applications (Table 5.1). By identifying these challenges the lack of insight into certain aspects of autonomous systems are highlighted.

**Table 5.1** Summary of autonomous system verification challenges discussed

Challenges	
1	How is an adequate model of the system created?
2	Common models and frameworks need to describe autonomous systems broadly enough so they can be used to standardize evaluation efforts and interfaces to the system
3	How should models of black box autonomous systems be developed and debugged? How is a mathematical and/or logical model suitable for formal analysis produced from empirical observations?
4	How should one identify and model components that are not captured yet (and what are their properties)?
5	What determines the level of simulator Fidelity to extract the information of interest?
6	How is the level of abstraction determined for the robot model, its behaviors, and the simulation that tests the model? How many environmental characteristics need to be specified? What are the aspects of the environment, the robot, and the autonomy algorithms that cannot be abstracted away without undermining the verification?
7	Where is the transition from specifying system requirements to designing the system and how are principled requirements developed so they do not devolve into designing the solution?
8	How is it ensured that the implicit and the explicit goals of the system are captured? How is a model of the system goals from a human understanding of the task goals, the system, and the environment created?
9	How are performance, safety, and security considerations integrated?
10	At what point is there enough evidence to determine that an autonomous system or behavior has been verified?
11	How does one ensure it is possible, in the physical world, to test simulated situations that result in boundary cases?
12	How would tests be designed so that passing them indicates a more general capability?
13	How are challenging design reference missions selected so that performing well against them indicates a more general capability for the system rather than for specific system components?
14	How can test scenarios be produced to yield the data required to generate mathematical/ logical models or to find the boundary locations and fault locations in the robot state space?
15	Once an adequate model is created how is it determined whether all resulting emergent behaviors were captured and what are appropriate performance measurement tools for this?
16	Measurement and evaluation are generally poorly understood—Operators can describe the tasks for the robot but lack tools to quantitatively evaluate them. How should autonomous behaviors be measured so they consistently and accurately describe the capability embodied by a robot?

(continued)

**Table 5.1** (continued)

Challenges	
17	How is a metric defined for comparing solutions?
18	How is the “optimal” defined against which the verification is performed?? How is the solution shown to be in fact, optimal? How is the performance of the system measured?
19	How is the performance from finite samples of the performance space generalized across several variables and parameters?
20	Autonomy frameworks are unable to determine whether all the resulting emergent behaviors have been captured or to supply performance measurement tools
21	What new tools or techniques need to be developed?
22	In general, how do we Verify the fitness of a given physical robot structure for a given task or environment (obviously, a robot that cannot sense color or is operating in the dark with an infrared sensor is unfit to sort objects on the basis of color)?
23	Descriptive frameworks are either too specific and constrain the developer to specific tools when designing the autonomous elements of the system or, too broad and difficult to apply to specific cases. Tools are needed to analyze systems at both the specific and the broad levels
24	How is a structured process that allows feedback between the physical/ground truth layer and the formal methods verification tools developed?
25	How to disambiguate between cases where the specification was incorrect (task description abstraction failed to capture some required system action) and those where the environmental model was incorrect (environmental abstraction failed to capture some critical system-environment interaction)? How to identify not just individual situations but classes of situations where the vehicle fails to be safe or to achieve safe operation (e.g. a front wheel often falls off the cliff but the back wheels never do). How should unanticipated unknowns be accommodated?
26	If an algorithm, or patch to an existing algorithm, was replaced can it be proven that no new failure modes were introduced without re-doing the entire verification process

While there are areas where progress is being made, and a few more with promising directions for future research, there are many problems that are not addressed. As verification of autonomous systems becomes a more pressing need for industry and a more mainstream research topic, we are optimistic that these challenges will be addressed and new tools and principled approaches will become available to support the safe transition of advanced autonomy and artificial intelligence into commercial autonomous systems.

**Acknowledgements** The authors would like to thank Andrew Bouchard and Richard Tatum at the Naval Surface Warfare Center in Panama City, Florida, for their help with early version of this paper, and the Verification of Autonomous Systems Working Group, whose efforts help define the terminology and identify these challenges. Thanks, are also due to the United States Naval Research Laboratory and the Office of Naval Research for supporting this research.



## References

- IEEE Standard Ontologies for Robotics and Automation (2015a), IEEE Std 1872-2015, 60 pages.
- IEEE Standard Ontologies for Robotics and Automation (2015b), P1872/D3, 55 pages.
- Albus, J., Huang, H. M., Messina, E., Murphy, K., Juberts, M., Lacaze, A., et al. (2000). 4D/RCS: A reference model architecture for unmanned vehicle systems version 2.0.
- Bartrop, K. J., Friberg, K. H., & Horvath, G. A. (2008). Automated generation and assessment of autonomous systems test cases. Aerospace Conference (pp. 1–10). IEEE.
- Billman, L., & Steinberg, M. (2007). Human system performance metrics for evaluation of mixed-initiative heterogeneous autonomous systems. Proceedings of 2007 Workshop on Performance Metrics for Intelligent Systems (pp. 120–126). ACM.
- Castillo-Effen, M., & Visnevski, N. A. (2009). Analysis of autonomous deconfliction in unmanned aircraft systems for testing and evaluation. Aerospace Conference (pp. 1–12). IEEE.
- Chaki, S., & Giampapa, J. A. (2013). Probabilistic verification of coordinated multi-robot missions. In *Model Checking Software* (pp. 135–153). Springer.
- Cleary, M. E., Abramsom, M., Adams, M. B., & Kolitz, S. (2001). Metrics for embedded collaborative intelligent systems. NIST Special Publication.
- Defense Acquisition University Press. (2001, January). *Systems Engineering Fundamentals*. Retrieved 2016, from [http://ocw.mit.edu/courses/aeronautics-and-astronautics/16-885j-aircraft-systems-engineering-fall-2005/readings/sefguide\\_01\\_01.pdf](http://ocw.mit.edu/courses/aeronautics-and-astronautics/16-885j-aircraft-systems-engineering-fall-2005/readings/sefguide_01_01.pdf)
- Dunbabin, M., & Marques, L. (2012, March). Robots for environmental monitoring: Significant advancements and applications. *IEEE Robotics and Automation Magazine*, 19(1), pp. 24–39.
- Durst, P. J., Gray, W., Nikitenko, A., Caetano, J., Trentini, M., & King, R. (2014). A framework for predicting the mission-specific performance of autonomous unmanned systems. *IEEE/RSM International Conference on Intelligent Robots and Systems*, (pp. 1962–1969).
- Engelberger, J. (1974). Three million hours of robot field experience. *Industrial Robot: An International Journal*, 164–168.
- Ferri, G., Ferreira, F., Djapic, V., Petillot, Y., Palau, M., & Winfield, A. (2016). The eurathlon 2015 grand challenge: The first outdoor multi-domain search and rescue robotics competition - a marine perspective. *Marine Technology Science Journal*, 81–97(17).
- Gehr, J. D. (2009). Evaluating situation awareness of autonomous systems. In *Performance Evaluation and Benchmarking of Intelligent Systems* (pp. 93–111). Springer.
- Huang, H. M., Albus, J. S., Messina, R. L., Wade, R. L., & English, R. (2004). Specifying autonomy levels for unmanned systems: Interim report. *Defence and Security* (pp. 386–397). International Society for Optics and Photonics.
- Huang, H. M., Pavek, K., Novak, B., Albus, J., & Messina, E. (2005). A framework for autonomous levels for unmanned Systems (ALFUS). Proceedings of the AUVSI's Unmanned Systems North America.
- Jacoff, A., Huang, H. M., Messina, E., Virts, A., & Downs, A. (2010). Comprehensive standard test suites for the performance evaluation of mobile robots. Proceedings of the 10th Performance Metrics for Intelligent Systems Workshop. ACM.
- Kress-Gazit, H., Wongpiromsarn, T., & Topcu, U. (1989). Correct, reactive, high-level robot control. *Robotics and Automation Magazine*, 18(3), pp. 65–74.
- McWilliams, G. T., Brown, M. A., Lamm, R. D., Guerra, C. J., Avery, P. A., Kozak, K. C., et al. (2007). Evaluation of autonomy in recent ground vehicles using the autonomy levels for unmanned systems (alfus) framework. Proceedings of the 2007 Workshop on Performance Metrics for Intelligent Systems (pp. 54–61). ACM.
- Miller, S., van den Berg, J., Fritz, M., Darrell, T., Goldberg, K., & Abbeel, P. (2011). A geometric approach to robotic laundry folding. *International Journal of Robotics Research*, 31(2), 249–267.
- Murphy, R. & Shields, J. (2012). Task Force Report: The Role of Autonomy in DoD Systems. Department of Defense, Defense Science Board.

- Paull, L., Severac, G., Raffo, G. V., Angel, J. M., Boley, H., Durst, P. J., et al. (2012). Towards an ontology for autonomous robots. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, (pp. 1359–1364).
- Pecheur, C. (2000). Validation and verification of autonomy software at NASA. National Aeronautics and Space Administration.
- Schultz, A. C., Grefenstett, J. J., & De Jong, K. A. (1993, October). Test and evaluation by genetic algorithms. *IEEE Expert*, 8(5), 9–14.
- Sholes, E. (2007). Evolution of a uav autonomy classification taxonomy. *Aerospace Conference* (pp. 1–16). IEEE.
- Smith, B., Millar, W., Dunphy, J., Tung, Y. W., Nayak, P., Gamble, E., et al. (1999). Validation and verification of the remote agent for spacecraft autonomy. *Aerospace Conference*. 1, pp. 449–468. IEEE.
- Smithers, T. (1995). On quantitative performance measures of robot behavior. In *The Biology and Technology of Intelligent autonomous Agents* (pp. 21–52). Springer.
- Steinberg, M. (2006). Intelligent autonomy for unmanned naval systems. *Defense and Security Symposium* (pp. 623013–623013). International Society for Optics and Photonics.
- Vassev, E., & Hinchey, M. (2013). On the autonomy requirements for space missions. *IEEE 16th International Symposium on Object/Component/Service-Oriented Real-Time Distributed Computing*, (pp. 1–10).

# Chapter 6

## Conceptualizing Overtrust in Robots: Why Do People Trust a Robot That Previously Failed?

Paul Robinette, Ayanna Howard, and Alan R. Wagner

### 6.1 Introduction

Robots are already entering our everyday lives. Even graduate students subsisting on a stipend can afford robotic assistants to clean the floors of their apartments. Some cars are already driving themselves autonomously on public roads. Unmanned aerial vehicles of varying degrees of autonomy are an ever-increasing concern to people as diverse as airline pilots, police officers, wildland firefighters, and tourists. Human error is a significant cause of accidents; however, the trust that these people place in any robot varies depending on the task of the robot and the characteristics of the interaction. More importantly, a robot can affect the trust that a person places in its hands, sometimes unintentionally. In this chapter, we formalize the concept of overtrust (Sect. 6.2) and apply it to our prior results in robot-assisted emergency evacuation.

Our work so far has focused on human-robot trust as it applies to humans accepting guidance from autonomous robots during a high-risk, time critical situations such as an emergency evacuation. The goal of this work was to develop a robot capable of guiding evacuees during an emergency and, in doing so, determine the

---

P. Robinette (✉)  
Department of Mechanical Engineering, Massachusetts Institute of Technology,  
Cambridge, MA, USA  
e-mail: [paulrobi@mit.edu](mailto:paulrobi@mit.edu)

A. Howard  
Department of Electrical and Computer Engineering, Georgia Institute of Technology,  
Atlanta, GA, USA  
e-mail: [ayanna.howard@ece.gatech.edu](mailto:ayanna.howard@ece.gatech.edu)

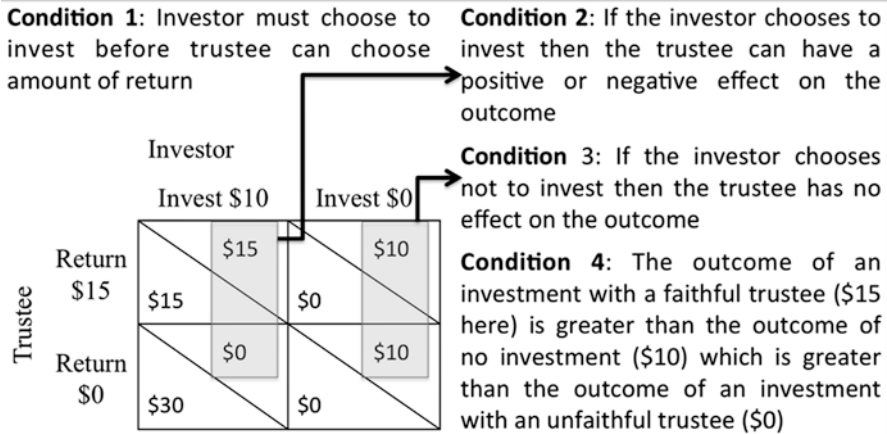
A.R. Wagner  
Department of Aerospace Engineering, Pennsylvania State University,  
State College, PA, USA  
e-mail: [azw78@psu.edu](mailto:azw78@psu.edu)

level of trust people place in this robot. In the beginning of our research, we assumed that most people would not trust a robot in a life-threatening situation. If they would trust the robot initially, surely they would stop trusting the robot once it made a significant, noticeable error. In fact, others (notably Desai et al. 2013) have found such a result in operator-robot interaction and our own initial work (Sect. 6.4, Wagner and Robinette 2015; Robinette et al. 2017) found that people would avoid trusting a robot that had previously failed them in virtual simulations. Unfortunately, it seems people do not think as critically when asked to trust a robot in a physical simulation of an emergency (Sect. 6.6 and Robinette et al. 2016a). This overtrust has been reported in two recent studies of human-robot interaction in lower risk scenarios (Bainbridge et al. 2011; Salem et al. 2015), but we have found that this is a problem in a high-risk scenario as well. Additionally, even in our virtual simulations, participants could be convinced to trust the robot again with a short, well-timed statement. This is discussed briefly in (Robinette et al. 2015) and in detail in Sect. 6.5. In Sect. 6.7, we discuss these results in terms of our conceptualization of overtrust and then we conclude with thoughts about future research.

In this chapter, we discuss experiments in both virtual and physical environments. We define a virtual human-robot interaction experiment as an experiment where participants observe and interact with a simulation of a robot through a computer. The robot is entirely simulated and the interaction takes place in some sort of a virtual environment, similar to interactions in video games. This paradigm is attractive because most scenarios that are difficult to create in a laboratory are fairly easy to create using modern three-dimensional modeling software and game engines. In contrast, a physical human-robot interaction experiment requires the use of an actual robot and thus typically requires physical space to perform the experiment (Bainbridge et al. 2011). The main advantage of performing an experiment with a physically present robot is that the participant experiences every aspect of the actual robot in question. Many components of a robot cannot be simulated accurately, so it is often necessary to perform a physical experiment in order to test the complete system. We discuss these concepts in detail in (Robinette et al. 2016b, c).

## 6.2 Conceptualizing Overtrust

Our previous work on conceptualizing trust (Wagner and Robinette 2015) uses outcome matrices to describe various trust scenarios. Outcome matrices are a useful tool for formally conceptualizing social interaction. These matrices (or normal-form games in the game theory community) explicitly represent the individuals interacting as well as the actions they are deliberating over. The impact of each pair of actions chosen by the individuals is represented as a scalar number or outcome. Figure 6.1 shows our conditions for trust represented in an outcome matrix of a simplified investor-trustee game. In this game, an investor can choose to invest money in another agent or not. If he or she chooses to invest, the other agent can act in good faith by returning a portion of the proceeds of the investment or not by



**Fig. 6.1** The conditions for trust derived from Wagner’s definition for trust are shown above with examples from the Investor-Trustee game

keeping all of the money for himself or herself. Each agent has an axis: the trustor’s (investor in this example) two actions (invest or not) are shown on top and the trustee’s two actions (return money or keep all money) are shown on the left. The outcomes for each element of the matrix are shown inside, with the outcome on the top-right representing the return to the investor and the outcome on the bottom-left representing the return to the trustee.

We define trust as “a belief, held by the trustor, that the trustee will act in a manner that mitigates the trustor’s risk in a situation in which the trustor has put its outcomes at risk” (Wagner 2009). Discussions of why we believe this definition of trust to be relevant for our work can be found in (Wagner and Robinette 2015). In short, this definition allows us to define a situation as “requiring trust” or not depending on the level of risk involved, the ability of the trustor (the one who must decide to trust the other agent) to choose freely, and the ability of the trustee (the agent who is to be trusted) to mitigate that risk. Formally, five conditions for situational trust can be derived from this definition (the first four are shown in Fig. 6.1). A fifth condition is that the trustor believes the trustee is likely to act in a manner that mitigates his or her risk.

This chapter extends our previous conceptualization of trust to include overtrust. Overtrust occurs when a trustor accepts too much risk, believing that the trusted entity will mitigate this risk. This is a concern when humans have committed an error or are about to commit an error in a situation where a potentially-unreliable intelligent agent could intervene. In terms of our defined conditions for trust, this means that either:

- Case 1: The trustor believes that the trustee will mitigate their risk (Condition 5), despite prior evidence to the contrary, or
- Case 2: The trustor believes that there is little risk if the trustee should fail, i.e., the trustee has little or no effect on the outcomes (Condition 2).

The first case of overtrust is indicated by a trustor claiming that the trustee was, in their opinion, mitigating his or her risk in the situation, regardless of prior behavior or current actions. In other words, the trustor may have an inaccurate model of the trustee's abilities and/or motivations. For example, experiments below show conditions where a participant trusts a robot to guide them safely in a high-risk scenario even though the robot has failed at that action before. The second case of overtrust is indicated by a trustor insisting that the situation had little or no risk, or that the trustee's actions had little effect on the outcomes. Here, we argue that the trustor has misjudged the actual risk of the situation. This is shown in our experiments by post-experiment surveys stating that some individuals judged the situation as having little danger or having no risk of monetary loss to participants.

### **6.3 Robot Guidance Versus Existing Guidance Technology**

In this experiment, we asked people to experience a subset of the robots that we had previously tested in a 3D simulated environment (Robinette et al. 2014, 2016b) during a simulated emergency. Participants were given the choice to follow guidance provided by a robot or guidance provided by emergency exit signs similar to those found in office buildings. With these experiments, we could test the conditions under which an evacuee would follow a robot or existing emergency guidance signs.

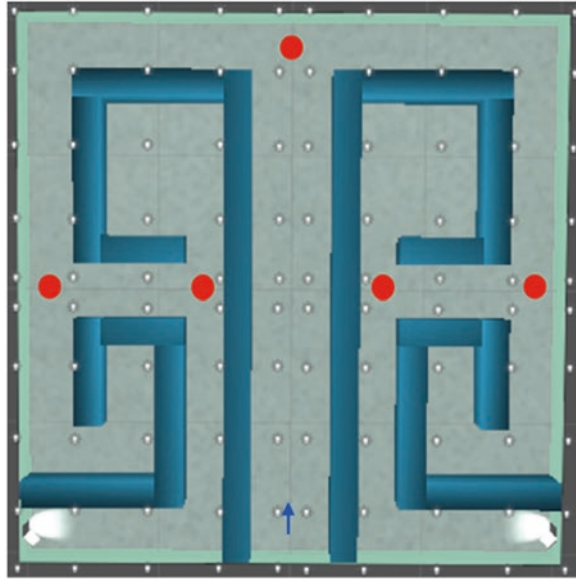
Evacuees exiting a building often encounter intersections which force them to make a decision about which direction to take. These decision points are usually accompanied by exit signs to help guide people to the closest exit. For this experiment, we also equipped each decision point with a robot to provide guidance. The guidance from the robot always contradicted the guidance from the exit signs. By measuring the person's choice at each decision point, we investigated the extent to which participants trust robot guidance more than exit sign guidance, or vice versa.

#### **6.3.1 Experimental Setup**

Participants began the interactive portion of the experiment in a maze (Fig. 6.2) facing a robot and a static emergency exit sign, one pointing left, the other right. Guidance information was presented at each decision point in the simulation. Five total decision points were used in the experiment. Two valid exits were available: one in the direction indicated by the robot and one by the sign. The participant had to follow the robot's or sign's guidance through at least three separate decision points to reach either exit, which limits the probability that a participant randomly chose to obey the robot or sign at all points.

The participant was encouraged to quickly find the exit. Prior to the experiment, instructions indicated that we were simulating an emergency. While navigating the environment, text on the screen stated "EMERGENCY! Please leave the building!

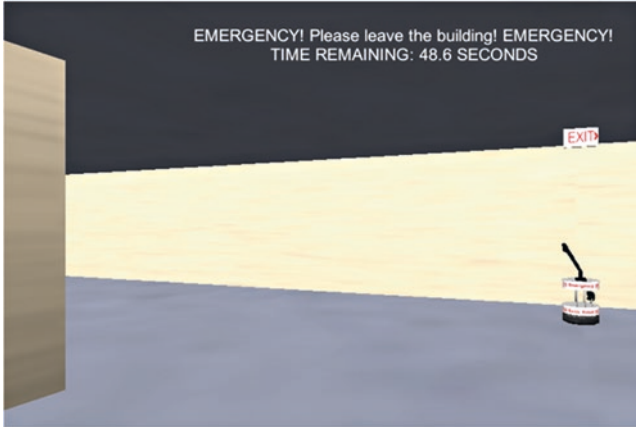
**Fig. 6.2** Maze environment for this experiment. Participants started in the position and orientation indicated by the *blue arrow*. Decision points are shown as *red dots*. A robot and an emergency exit sign with an *arrow* were at each decision point. One pointed to the path that lead to the exit on the *left* (shown in the diagram as an open door) and the other pointed to the exit on the *right*. Maze walls are shown in *dark blue*



**Fig. 6.3** Dynamic sign robot

EMERGENCY!” and a timer counted down from 60 s (see Figs. 6.3, 6.4, and 6.5). If a participant failed to find an exit in 60 s, then the participant was informed that they had not survived the simulation.

Three robots were tested: a Dynamic Sign robot, a Multi-Arm Gesture robot and a Humanoid. Motivation for these robots can be found in (Robinette et al. 2014). The Dynamic Sign platform was simulated as a Turtlebot with signs indicating it is



**Fig. 6.4** Multi-arm gesture robot



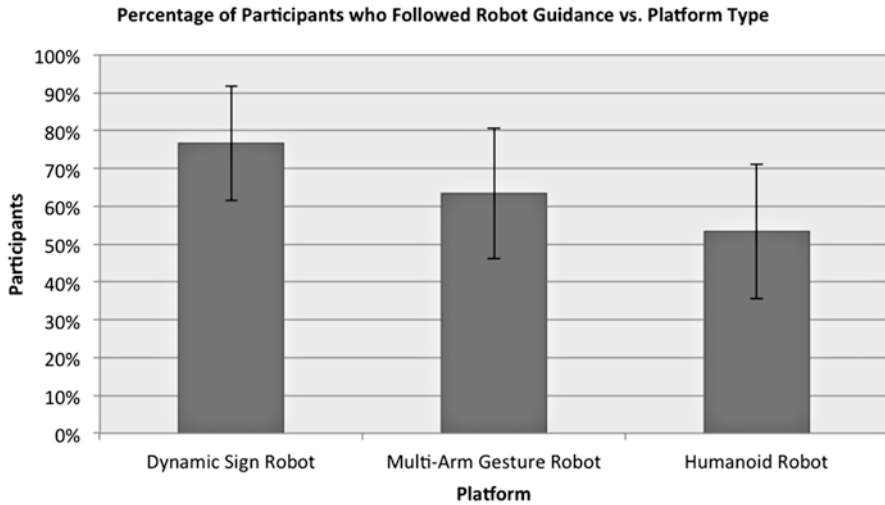
**Fig. 6.5** Humanoid robot

an emergency guide robot and a screen on top. The screen flashed directions using English and arrows. This platform was used because our prior research verified that participants would understand the information presented on the screen. The multi-arm gesture robot was a similar Turtlebot with two arms instead of a screen. The arms pointed in the direction of the exit. It was used because it also scored highly in previous tests. The Humanoid platform was included in order to test the difference, if any, between it and the Turtlebot-based Multi-Arm Gesture platform. It signaled the same way as the Multi-Arm Gesture platform.

After the interactive portion of the experiment, participants were asked four questions about their experience and then completed a short survey to gather demographic data. The four questions were:

1. Did you notice the robots doing anything to help you find the exit?
2. Did you notice the exit signs on the ceiling?





**Fig. 6.6** Percentage of participants who followed robot guidance broken down by robot type. Error bars represent 95% confidence intervals

3. Did you trust the information provided by the robots?
4. Did you trust the information provided by the exit signs on the ceiling?

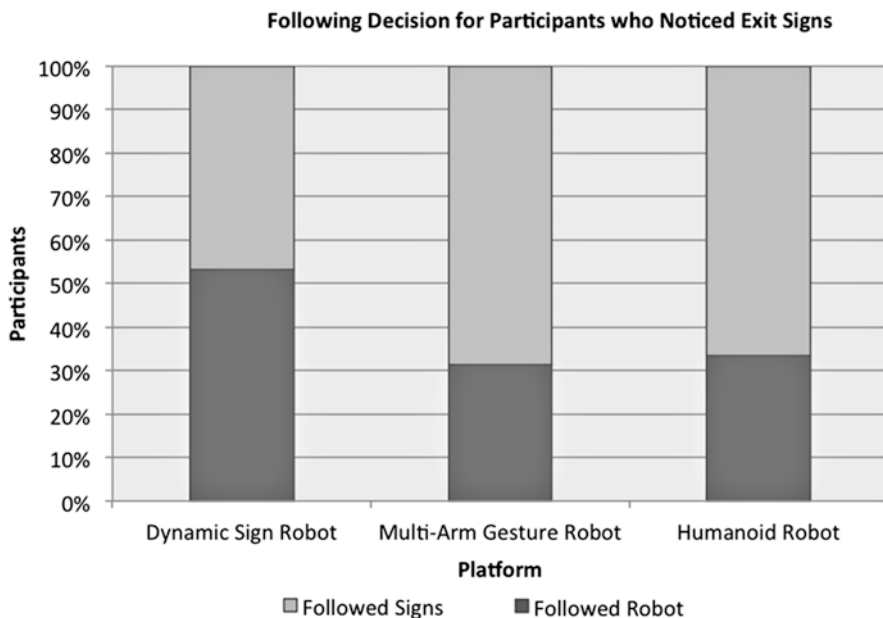
Each question offered yes and no options for a response and asked participants to explain themselves. Position data and the time it took the participant to find the exit in the experiment were recorded. The environment was built in the Unity game engine and the participant interacted with the robot using a plugin in a web browser.

A total of 95 people participated in this experiment via Amazon’s Mechanical Turk service. Five participants were unable to find the exit in the time provided and their results were excluded from analysis. The remaining 90 were evenly divided among the three robots.

### 6.3.2 Results

Overall, 61% of participants followed the robots instead of the exit signs ( $p=0.002$ , Binomial test assuming 50% random chance of following either robots or signs for 90 samples). The difference in the following decision between the robots was not statistically significant at this sample size ( $\chi^2(2, n=90)=0.341, p=0.166$ ), but some trends can be gleaned from it. The Dynamic Sign robot had the highest following rate (77%), followed by the Multi-Arm Gesture (63%) and then the Humanoid (53%) (Fig. 6.6). There was a strong correlation between noticing the exit sign and following the exit sign ( $\phi(90)=0.59$ )<sup>1</sup> and a weaker correlation between noticing

<sup>1</sup>The phi coefficient (“ $\phi$ ”) measures the correlation between two variables.



**Fig. 6.7** Results from participants who noticed exit signs only

the robots and following the robots ( $\phi(90)=0.39$ ). Figure 6.7 shows the results from participants who noticed the exit signs. Note that this sample size is relatively small, so it is hard to draw conclusions from the data, but most participants who noticed the exit signs chose to follow them.

For the other participants, the explanations for their answers reinforced the conclusion that they did not notice the exit signs. Some representative comments were: “I didn’t notice the exit signs on the ceiling. I would have followed them if I would have noticed,” “I only saw one or two. These were not as helpful, since I was able to miss some,” “I didn’t think I had any other hint [besides the robots] on where to go,” “In times of emergency, you have to make quick decisions so I chose to trust the robots.” Other participants noticed the exit signs, but preferred to rely on the robots, saying, “I trusted them, but did not use them because I was [moving] too quickly to register them.”

Some participants indicated that the exit signs had a greater chance of being correct by saying comments like, “[The robots] brought attention to the exit signs but appeared to be [pointing] in the wrong direction,” “People wouldn’t put up signs that pointed the wrong way,” “I figured [the robots] to be more of a distraction and thought it would take too much time to figure out how they were trying to help me,” “It seemed like [the robots’] arm was moving. I [ignored] them though. The exit sign was easier to understand,” “I decided to go by the sign on the wall because it was not moving and seemed to be there longer”, “I didn’t have time to figure out what [the robots] were trying to communicate.” One participant indicated that he did not trust the robots because he had seen the film “I, Robot.”

Other participants wrote that the robots, especially the Humanoid, looked like an authority figure: “[The robots] seemed to be with an ‘authority’ outfit, looked like policemen at first.” Some did not understand that the dynamic sign robots were robots, and simply thought that it was a mobile sign: “They had the correct signs that were easily identifiable.”

Based on the comments and the correlations, we can conclude that participants generally followed the exit signs if they noticed the exit signs but were more likely to notice the robot. The robot was a sufficiently distracting object that most participants did not even notice the exit signs. We can thus conclude that robots are better at attracting attention during emergencies than standard emergency exit signs. These robots were also found to be sufficiently trustworthy to aid participants in finding an exit. Note that participants had no prior information about the robot’s capabilities while they did know that exit signs are usually intentionally placed to guide people out of a building.

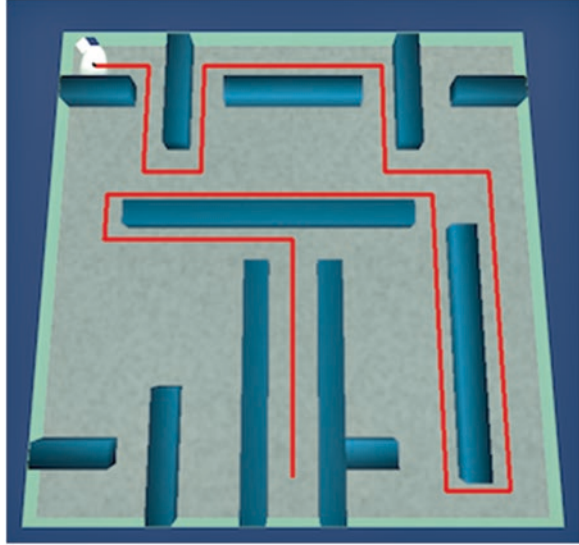
## 6.4 Human-Robot Trust in Virtual Simulations

In this section, we briefly describe several experiments we have performed over the last few years that help to indicate when participants do and do not overtrust our robots. For more details on these experiments, please refer to their citations (Robinette et al. 2017; Wagner and Robinette 2015). These experiments each ask a participant to navigate a maze and offer robotic assistance to help the subject navigate it. Participants are free to choose whether or not to accept the robot’s help. We began with a single-round experiment where a participant was asked if they would like robotic assistance with little or no knowledge of the robot’s abilities. We then extended this to a two-round interaction where participants could choose to experience the robot in a first round and then decide if they would like to continue using it in a second round or not. All experiments used a 3D simulation of a maze environment (Fig. 6.8) created in the Unity game engine. All experiments recruited and compensated participants through Amazon’s Mechanical Turk.

Each experiment began by thanking the person for participating in the experiment. Next the subject was provided information about the maze evacuation task. Participants were shown examples (in the form of pictures and text) of good and bad robot performance (e.g., robots that are fast and efficient and robots that are not) and participants were given an idea of the complexity of the maze (although they were not shown the exact maze they would be asked to solve). Also, as part of this introduction, participants were given the chance to experiment with the controls in a practice environment. The practice environment was a simple room with three obstacles and no exit.

After this introduction, participants were given the choice to use the robot or not. Participants were told that their choice to use the robot would not affect their compensation. Participants were then placed at the start of the virtual maze. If they chose to use the robot it would start out directly in front of their field of view and

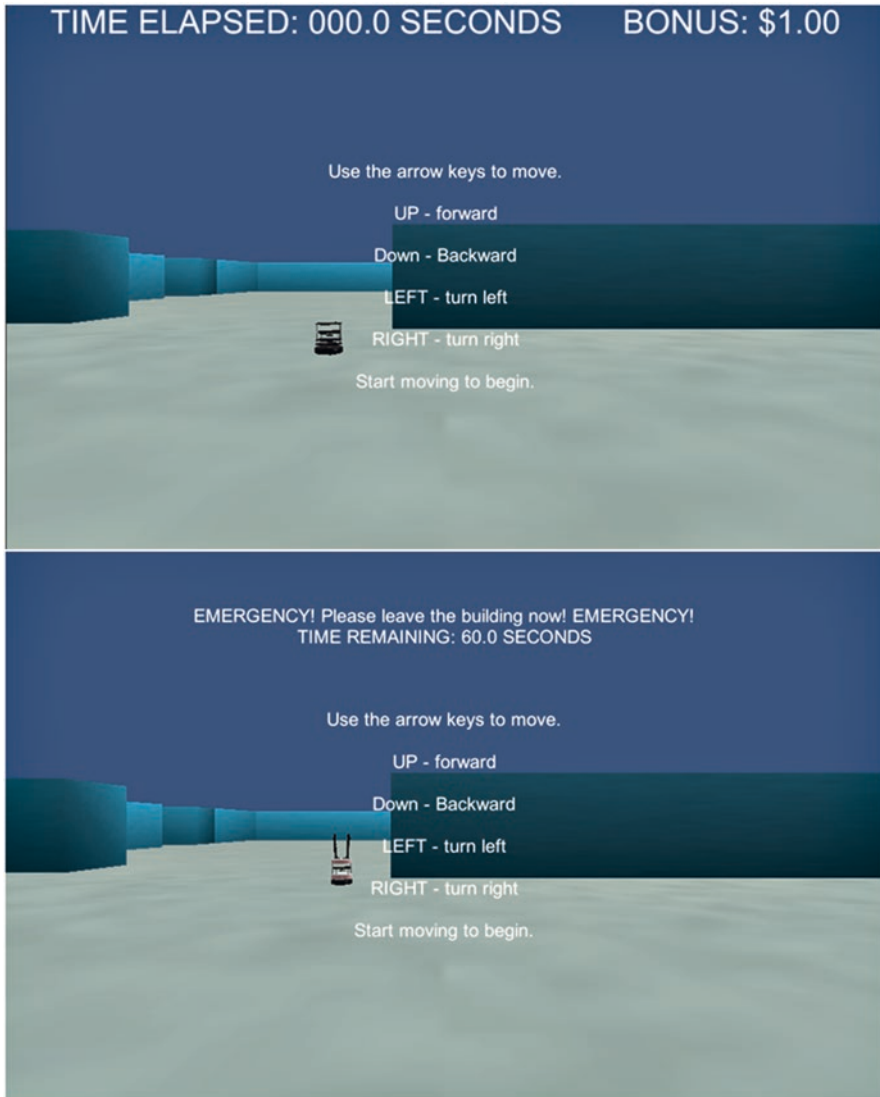
**Fig. 6.8** Example maze with circuitous robot path drawn in red. The starting point is in the *bottom center* and the exit is in the *top left*



immediately begin moving towards its first waypoint (Fig. 6.9). The robot would move to a new waypoint whenever the participant approached. If the participant elected to not use the robot, then no robot would be present and the participant would have to find the exit on his or her own. After the maze-solving round was complete, participants answered a short survey about the round. In the two-round experiment, they were then asked if they would like to use the robot or not in the second round and a new maze was presented for them to solve. They were then asked questions about their experience in that maze. Finally, participants in all experiments were asked demographic questions.

Full results from the single round maze can be found in (Wagner and Robinette 2015). A total of 120 participants completed this experiment and 77% of them chose to use the robot. This indicates that people have a tendency to trust a robot initially, before they can develop a model of the robot's behavior. This tendency does not necessarily mean that people actually trust the robot to help them, but it does indicate that they at least trust the robot to not hurt their outcomes in the experiment. In this experiment, participants were shown the risks of failing to solve the maze in time (they were told their character would die) and were also shown the risks of following a poorly performing robot (again, they were told their character would probably die). Thus, they were explicitly shown the risks in this scenario and we do not believe that they overtrusted robots due to Case 2 shown in Sect. 6.2. Instead, we believe that people fell into Case 1: they believed that the robot was more capable than the evidence shown so far. As stated above, this could be a weak belief because no evidence of robot capability had been given at this point.

A complete discussion of our two-round experiment can be found in (Robinette et al. 2017). In this experiment, we manipulated both participant motivations and first-round robot performance to determine the effects on participants. We tested



**Fig. 6.9** The *top image* shows the start of a monetary bonus maze and the *bottom* shows the start of a survival motivation experiment. In monetary motivation, time increases as bonus is reduced and in survival, time decreases. Movement directions disappear after the participant begins to move

both monetary (participants received a bonus based on the quickness of their maze solution) and survival (participants were told their character would die if they did not find the exit in time) motivations. We found that a poorly performing robot (a robot that did not guide them to the exit in time to preserve their bonus or their character's life) would generally not be used in the second round of the experiment

in the survival motivation condition. Interestingly, a similar effect was NOT found in the monetary bonus condition: participants were just as likely to continue to use a poorly performing robot as a good robot. In this case, participants in the monetary motivation condition likely fell into Case 2: they did not believe that the risk of robot failure was great enough to hurt their outcomes. This belief could be a failure of methodology (i.e., we did not properly motivate them) or a failure of understanding on their part (many participants indicated that they believed the robot would do better in the second round). Regardless, this result indicates that people will tend to choose to use robots when they believe the risk of the situation is low. It is unlikely that humans will always judge the risk of interaction correctly, so robots need to be able to communicate this risk to avoid overtrust.

One possible early indicator of a tendency to overtrust was ignored at this step because we believed it to be a methodological error. In pilot tests, we tried many different poorly performing robots to determine which behavior participants could quickly identify as “bad.” A full discussion is provided in (Robinette et al. 2016b, c), including our reasoning for believing these results were in error, but hindsight suggests some additional insight can be gleaned. Three of the robot behaviors tested involved the robot performing continuous loops without ever finding the exit. One looped around a single obstacle, another around a larger set of obstacles, and the last around the entire environment (except the hallway that contained the exit). These participants were each eliminated from consideration because the participants (only five in each category) did not seem to understand that the robots had performed poorly. One participant followed a robot in a loop around a single obstacle for almost 4 min in the first round and then over 9 min in the second round. Another followed the robot that continuously circled the environment for almost 12 min (three complete loops around the entire environment), even though the bonus expired after 90 s. Results from these pilot studies show that participants have a difficult time believing the robots had failed in their task. Even after abandoning the robot and finding the exit on their own (the only way to proceed to the second round), some participants still chose to use the robot again. This combination of disbelief of poor robot behavior and quick forgiveness (in this case, unprompted) indicates that participants had an incorrect model of the robot’s capabilities (Case 1). As a further indication that participants tend to give the robot the benefit of the doubt, another pilot study tested a behavior where the robot collided with a wall just before finding the exit and, instead of interpreting this as a robot with bad obstacle detection, decided that the robot was colliding with the wall to signal that the exit was near. While this robot did provide some helpful guidance, we believed participants would view it unfavorably because it was unable to navigate around an obvious obstacle.

Now that we have shown the conditions under which trust can be broken, there are two logical steps to take: attempt to repair this broken trust and replicate our experiments in physical experiments. We proceeded along both tracks simultaneously, but the ease of virtual experiments provided us with results on trust repair much quicker than results from physical experiments. Thus, we next present methods that a robot can use to modify a human’s trust in it.

## 6.5 Repairing Broken Trust

In the previous section, we showed that trust can be broken in virtual simulations if a robot performs poorly and the experiment presents a survival risk. In this section, we show results from one of the next logical experiments to take after this result: repairing broken trust. Attempting to repair trust is one way that a robot can modify a person's trust level. Presumably, humans will lose trust in robots in some real-world situations, so a robot should have tools to repair that trust, when needed. Additionally, we expect some of these methods to be relevant in situations where a robot may need to convince a person to trust it less. As we will show in the next section, this is more difficult than it may seem. A subset of our results on trust repair was published in (Robinette et al. 2015). We show the complete results here in order to illustrate the many options a robot has to modify trust level and our conclusions on those options so far.

The methods that we use to repair trust are inspired by studies examining how people repair trust. Schweitzer et al. (2006) examined the use of apologies and promises to repair trust. They used a trust game in which participants had the option to invest money in a partner. Any money that was invested would appreciate. The partner would then return some portion of the investment. The partner violates trust both by making apparently honest mistakes and by using deceptive strategies. The authors found that participants forgave their partner for an honest mistake when the partner promised to do better in the future, but did not forgive an intentional deception. They also found that an apology without a promise included had no effect. In Kim et al. (2006), the authors tested the relative trust levels that participants had in a candidate for an open job position when the candidate had made either integrity-based (intentionally lied) or competence-based (made an honest mistake due to lack of knowledge) trust violations at a previous job. They found that internal attributes used during an apology (e.g., "I was unaware of that law") were somewhat effective for competence-based violations, but external attributes (e.g., "My boss pressured me to do it") were effective for integrity-based violations.

### 6.5.1 *Experimental Setup*

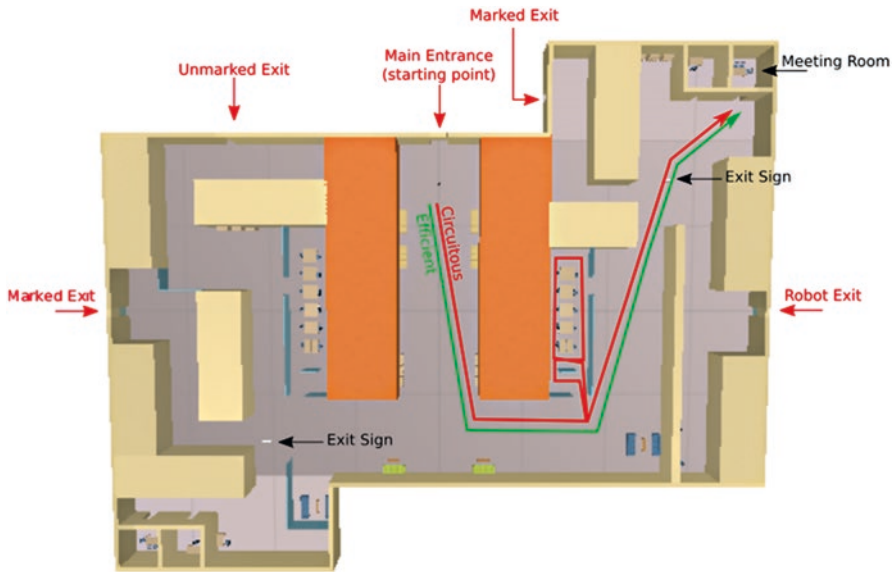
To provide a more realistic environment to test trust, we designed a 3D office simulation using the Unity game engine. We again employed Amazon's Mechanical Turk to recruit and compensate participants. The simulation began by introducing participants to the experiment and the robot. Participants were then asked to learn the movement controls of the simulation in a practice round. After the practice round, participants were asked to follow the robot to a meeting room where they were told they would receive further instructions. When the participants reached the meeting room, the robot thanked them for following it and the participant was asked "Did the robot do a good job guiding you to the meeting room?" with space to explain their

answers. Once the participants completed this short mid-experiment survey, they were told “Suddenly, you hear a fire alarm. You know that if you do not get out of the building QUICKLY you will not survive. You may choose ANY path you wish to get out of the building. Your payment is NOT based on any particular path or method.” During this emergency phase, the robot provided guidance to the nearest unmarked exit. Participants could also choose to follow signs to a nearby emergency exit (approximately the same distance as the robot exit) or to retrace their steps to the main exit. As mentioned, above, other exits were available in the simulation, but participants were not expected to notice them as they would not have any reason to traverse that section of the environment. Participants were given 30 s to find an exit in the emergency phase. The time remaining was displayed on screen to a tenth of a second accuracy. This count down was shown in our previous research to have a significant effect in motivating participants to find an exit quickly (Robinette et al. 2017). The simulation ended when the participant found an exit or when the timer reached zero. After the simulation, participants were informed if they had successfully exited or not. Finally, they were asked to complete a survey.

As in previous experiments, the robot would either provide fast, efficient guidance to the meeting room or take a circuitous route. In previous experiments, we showed that these behaviors can be used to bias most participants to trust (by using the efficient behavior) or not trust (by using the circuitous behavior) the robot later in the experiment. Efficient behavior consists of the robot guiding the participant directly to the meeting room without detours. Circuitous behavior consists of the robot guiding the participant through and around another room before taking the participant to the meeting room. Both behaviors can be seen in Fig. 6.10. Each behavior was accomplished by having the robot follow waypoints in the simulation environment. At each waypoint, the robot stopped and used its arms to point to the next waypoint. The robot began moving towards the next waypoint when the participant approached it. The participant was not given any indication of the robot’s behavior before the simulation started.

We expected participants to lose trust in the robot after it exhibited circuitous behavior. After guiding the person to the meeting room, the robot has two discrete times when it can use a statement to attempt to repair this broken trust: immediately after its trust violation (e.g., circuitous guidance to the meeting room) or at the time when it asks the participant to trust it (during the emergency). An apology or a promise can be given during either time. Additionally, the robot can provide contextually relevant information during the emergency phase to convince participants to follow it. Table 6.1 shows the experimental conditions tested in this study and Fig. 6.11 shows when each condition would be used. Statements made by the robot were accomplished using speech bubbles displayed above the robot in the simulation (Figs. 6.12, 6.13, and 6.14). The percentage of participants who followed the robot was then compared with the efficient and circuitous controls to determine if trust was repaired (i.e., if people followed it as in the efficient condition) or not (i.e., if people chose to use an alternate exit as in the circuitous condition). To ensure that the speech bubble itself was not a significant factor, an empty speech bubble was used in one condition. A condition, labeled the Nice Meeting Wishes condition, was





**Fig. 6.10** The virtual office environment used in the experiment. Efficient robot path (*green*) versus circuitous robot path (*red*) are shown

added to determine if there was any effect when the robot made a statement that did not attempt to repair trust.

## 6.5.2 Results

The results of the experiment and the number of participants considered for analysis are in Fig. 6.15. Across all categories, 307 (53%) participants followed the robot during the emergency phase. Of the 268 who did not, 226 (84%) went to the nearby marked exit, 17 (6%) chose to retrace their steps to the main entrance, 10 (4%) found another marked exit further away, and 15 (6%) participants failed to find any exit during the emergency phase.

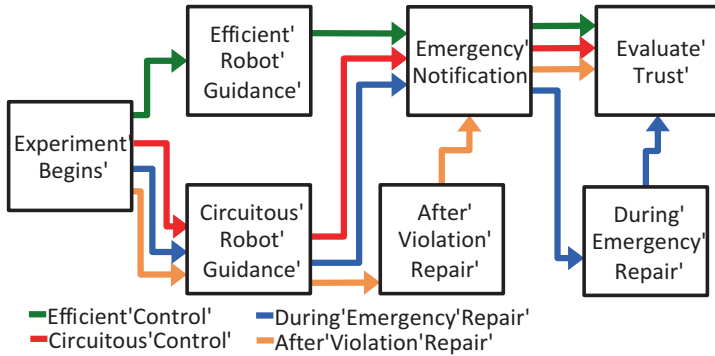
Attempts to repair trust (all conditions except Empty Speech Bubble and Nice Meeting Wishes) during the emergency succeeded in increasing trust (chi-squared test,  $p < 0.05$  compared to circuitous control and  $p > 0.05$  compared to efficient control). Similar techniques used immediately after the trust violation and before the emergency had no such effect (all After Violation conditions, chi-squared,  $p > 0.05$  compared to circuitous control and  $p < 0.05$  compared to efficient control). The empty speech bubble had no effect when compared with the circuitous control (chi-squared,  $p > 0.05$ ); however, the nice meeting statement did have a significantly different effect from both the circuitous and efficient controls (chi-squared,  $p < 0.05$  for both).

**Table 6.1** Experimental conditions. *Green* and *red* indicate the controls, *orange* indicates the after violation conditions and *blue* indicates the During Emergency conditions

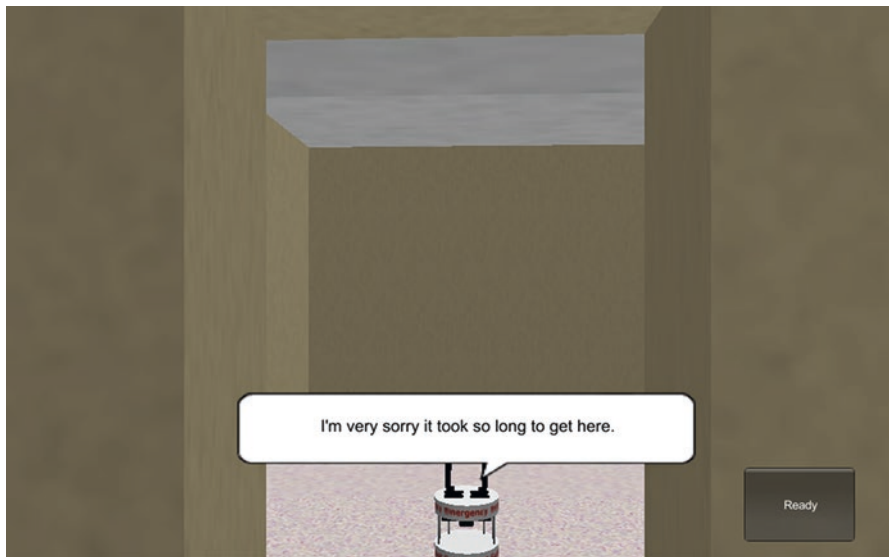
Label	Statement Given in Speech Bubble	Timing
Efficient Control	N/A	N/A
Circuitous Control	N/A	N/A
Promise (After Violation)	"I promise to be a better guide next time."	After Violation
Apology (After Violation)	"I'm very sorry it took so long to get here."	After Violation
Both Promise and Apology	"I'm very sorry it took so long to get here. I promise to be a better guide next time."	After Violation
Internal Attribution Apology	"I'm very sorry it took so long to get here. I had trouble seeing the room, but I fixed my camera."	After Violation
External Attribution Apology	"I'm very sorry it took so long to get here. My programmers gave me the wrong map of the office but I have the right one now."	After Violation
Situation Information	"There is a fire emergency."	During Emergency
Exit Information	"There is an exit this way."	During Emergency
Distance Information	"This exit is closer."	During Emergency
Congestion Information	"The other exit is blocked."	During Emergency
Empty Speech Bubble	N/A	During Emergency
Nice Meeting Wishes	"I hope you enjoyed your meeting."	During Emergency
Promise (During Emergency)	"I promise to be a better guide this time."	During Emergency
Apology (During Emergency)	"I'm very sorry it took so long to get to the meeting room."	During Emergency

## Results

In Robinette et al. (2015), we discuss this interesting timing result and justify our control conditions, but in this chapter, we focus on the manipulation of human trust levels. These results indicate that robots can manipulate a person's trust decision with a simple statement. This makes sense when the statement adds information to the situation, such as when the robot gives a reason (shorter distance, less congestion) for pointing to an unmarked exit, but we found the effect even when the statement contained the same information as the announcement of the emergency. In fact, there was even a small, significant, increase in trust when the robot said it



**Fig. 6.11** The experiment begins with the robot providing either efficient or circuitous guidance to a meeting room. After arriving in the meeting room, the participant is informed of an emergency. In some conditions, the robot attempts to repair trust before the emergency (immediately after the trust violation, shown in *orange*) and in others it attempts to repair trust during the emergency (shown in *blue*). At the end of the experiment, trust is evaluated based on the exit the participant chose. Two controls were used to determine the effect of efficient (*green*) or circuitous (*red*) guidance without any trust repair attempt



**Fig. 6.12** Robot apologizing for its performance immediately after the violation

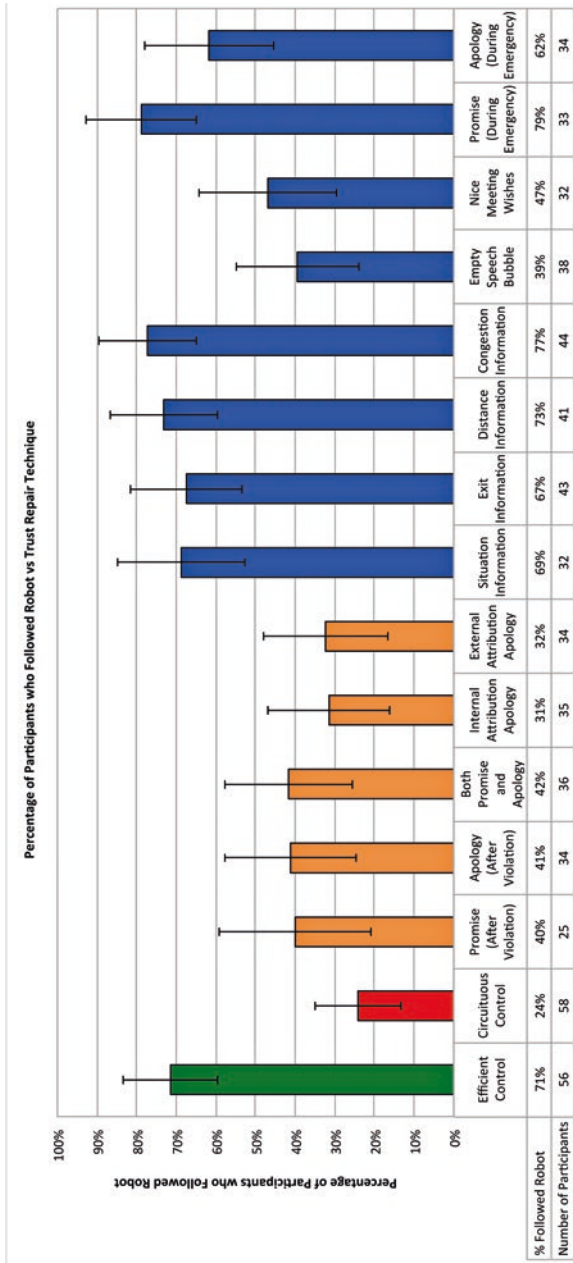
hoped the participant had a nice meeting. While this does not necessarily indicate that people overtrusted robots in this scenario, it does indicate that people are willing to forgive robots for previous errors with little prompting. This finding could be used by robots to increase trust when the robot knows it is capable, but it could also be used to convince people to overtrust (Case 1).



**Fig. 6.13** Robot providing additional distance information during the emergency



**Fig. 6.14** Robot apologizing for its prior performance during the emergency



**Fig. 6.15** Results from the experiment. Error bars represent 95% confidence intervals. *Green* and *red* indicate the controls, *orange* indicates the After Violation conditions and *blue* indicates the During Emergency conditions



**Fig. 6.16** Robot during non-emergency phase of the experiment pointing to meeting room door (*left*) and robot during emergency pointing to back exit (*right*). Note that the sign is lit in the *right* picture. A standard emergency exit sign is visible behind the robot in the emergency

## 6.6 Overtrust of Robots in Physical Situations

To create a high-risk situation, we conducted a physical simulation of a real-world emergency evacuation scenario using fire alarms and artificial smoke to add urgency. This was performed in a manner similar to the experiment in the previous section: a robot first guided participants to a meeting room, then an emergency occurred and the robot waited in the hallway, pointing them to an unmarked exit (Fig. 6.16). Artificial smoke and alarms provided motivation for participants to find an exit. Participants were not informed that an emergency would take place prior in the experiment. A summary of the experiment is below, but more information can be found in Robinette et al. (2016a).

### 6.6.1 Experimental Setup

This experiment took place in the office area of a storage building on the Georgia Tech campus. The building was otherwise unoccupied during experiments. The office area contained a hallway and several rooms (Fig. 6.17). The room at the end of the hallway was designated the meeting room and the room next to it was designated the other room, only used in the circuitous behavior condition. The back exit

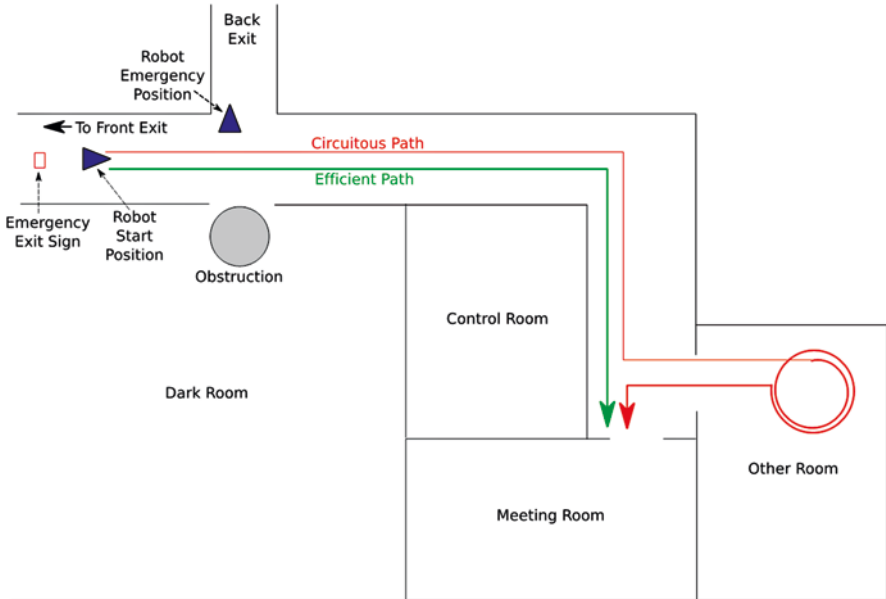


Fig. 6.17 Layout of experiment area showing efficient and circuitous paths

used for this experiment actually lead to a large storage area, but this was obscured using a curtain. Participants could see light through the curtain, but could not see shapes or objects behind the curtain. This arrangement was intended to make this doorway appear as a plausible path to an exit, but not a definite exit to the outdoors. A standard green emergency exit sign hung in the hallway indicating that participants should exit through the main entrance in the event of an emergency. A room in the middle of the building was designated as the control room. An experimenter stayed in that room controlling the robot over a wireless link. The experimenter could view the entire experiment area from five cameras placed throughout the building but could not be seen by participants.

Five different robot behaviors were used in this experiment:

- Efficient: The robot takes the most direct path to the meeting room in the first phase.
- Circuitous: While navigating to the meeting room, the robot enters an unrelated room and performs two circles before exiting and providing guidance to its destination.
- Broken: When the robot reached the first corner, it spun in place three times and pointed at the corner itself. No discernible guidance information was provided by the robot to participants. An experimenter then approached the participant and said, “Well, I think the robot is broken again. Please go into that room [accompanied with gestures to the meeting room] and follow the instructions. I’m sorry about that.”

**Fig. 6.18** Robot providing incorrect guidance condition by pointing to a dark, blocked room in the emergency



- Immobilized: At the start of the experiment, the robot moved a short distance forward, but then, upon reaching the intersection of the hallways (Robot Emergency Position in Fig. 6.17) it spun in place three times and then pointed to the back exit. At this point, an experimenter informed the participant that the robot was broken with a similar speech as in the broken robot condition. The robot did not move and continued gesturing towards the back exit for the remainder of the experiment. The robot’s emergency lights were not turned on.
- Incorrect: The robot performed the same as in the broken robot condition, with accompanying experimenter speech, in the non-emergency phase of the experiment. During the emergency, the robot was stationed across the hall from its normal emergency position and instructed participants to enter a dark room (Figs. 6.17 and 6.18). The doorway to the room was blocked in all conditions with a piece of furniture (initially a couch then a table when the couch became unavailable) that left a small amount of room on either side for a participant to squeeze through to enter the room. There was no indication of an exit from the participant’s vantage point. All lights inside of the room were turned off.

## 6.6.2 Results

Every single participant who experienced a robot that only failed on the way to the meeting room chose to follow the robot in the emergency. Four of five participants who saw a robot, which maintained the same failing behavior as the one during guidance to the meeting room, followed it in the emergency. Finally, two participants followed the robot’s guidance into a blocked, unlit room and two others stayed with the stationary robot until an experimenter retrieved them several minutes later. Clearly, these participants either trusted the robot despite its earlier failings or considered the situation to have lower risk than we wanted to project. Post-experiment surveys helped us to determine which was more likely.

Of the 42 participants included in all of our studies, 32 (76%) reported not noticing the exit sign behind the robot’s emergency position. Upon turning the corner



from the smoke-filled hallway on their way out, participants' eyes were drawn to the large, well-lit, waving robot in the middle of their path. Couple the visual attraction of the robot with the increased confusion reported on the surveys (for full results, see Robinette et al. 2016a), it is no surprise that participants latched onto the first and most obvious form of guidance that they observed. Perhaps participants did not believe that they were in any danger and followed the robot for other reasons (Case 2 overtrust). Their increased confusion scores between pre- and post-experiment surveys and reactions to the smoke indicate that at least some of the participants were reacting as if this was a real emergency. Given that every participant in the main study followed the robot, regardless of their rating of emergency realism, we believe that the realism of the scenario had little or no effect on their response. Additionally, many participants wrote that they followed the robot specifically because it stated it was an emergency guide robot on its sign. They believed that it had been programmed to help in this emergency. This finding is concerning because participants seem willing to believe in the stated purpose of the robot even after they have been shown that the robot makes mistakes during a related task (Case 1 overtrust). One of the two participants who followed the robot's guidance into the dark room even thought that the robot was trying to guide him to a safe place after he was told by the experimenter that the exit was in another direction. Most participants in the physical experiment reported that they did not believe the emergency was real (Case 2 overtrust), but if the same question had been asked in the virtual experiment we would expect none of them to believe that the emergency was actually real. In such virtual simulations, the emergency is contained to their computer and thus could not affect them in any way. Interestingly, significantly more participants reported that they were motivated (according to a true/false question in the post-experiment survey) in the emergency phase of the virtual experiment than in the physical experiment ( $\chi^2(1, n = 140) = 26.658, p < 0.001$ ).

## 6.7 Discussion

Throughout our work, we have found several instances where participants have seemingly-unwarranted trust in our robots. Without any knowledge of the robot's abilities or motivations, they trust it to aid them in a guidance task, even when recognizable emergency exit signs point in the opposite direction. This result indicates that people tend to believe robots are competent at first sight. This behavior was stronger when the robot was specifically identified as an emergency guide robot. Apparently, people expect robots to do what they claim to be able to do, regardless of prior experience or lack thereof. Such behavior may prove troublesome when people are asked to trust robots with their lives. Our research implies that some people would be willing to get into a self-driving taxi, even if they knew nothing about it. Based on these results, we believe that people tend to exhibit some amount of Case 1 overtrust when encountering a new robot.

Even after experiencing a bad robot, many participants decided to keep using it in future interactions. During virtual simulations with the survival motivation, about half chose to take the seemingly rational option of ignoring the previously poorly performing robot, but the other half chose to continue to use it. Participants gave many reasons for this, including that they thought the robot would perform better in the second phase and that they still thought it was better than a human, indicating that they fell into a Case 1 overtrust situation as defined above. They believed that the robot was more capable than their previous experience with it suggested.

Other participants may be categorized as Case 2: believing that the risk of the situation was low and thus the robot's actions had little effect on their outcomes. Still, if they chose to follow the robot, they generally indicated on post-experiment surveys that they trusted it. In other work, we have found a high correlation between the risk of the situation and trust in online surveys (Wagner and Robinette 2015). In our physical experiment, it was reasonable to believe that the experimenters would render assistance if there was real danger, but there is little indication that participants believed this when smoke appeared and fire alarms sounded. In contrast, almost half of the participants had a physical reaction to the smoke (e.g., stepping back in surprise). Not one of them chose to find a human to ask for help.

In several of our studies, participants reported that they did not notice standard emergency exit signs when a guidance robot was present. This may or may not indicate overtrust of robots in general, but it is still a result that should concern robot designers. Based on these studies, people trust a robot's abilities enough that they do not look for alternatives to robotic assistance. This meets our Case 1 overtrust description, but is weakened because participants did tend to follow exit signs when they noticed them. Regardless, robot designers should be aware that their lighted, moving platforms will attract focus to the detriment of other items in the scene.

Several of the above concerns can be mitigated with increased communication from the robot. We have already shown that short, timely statements can increase trust in a robot. We expect a similar effect could be found to decrease trust. Of course, this requires that the robot recognize when it should not be trusted. Still, if a robot can modify trust in itself to cause overtrust, it can probably also cause appropriate-trust. This communication may prove difficult. We have repeatedly discussed how it was difficult to convince people that a robot had malfunctioned. Even direct statements by experimenters did not stop people from following the robot in an emergency. Perhaps direct statements from the robot itself will have an effect.

Our results from the physical experiment in Sect. 6.6 directly contradict results from virtual experiments in Sect. 6.4. To begin to address this discrepancy, we must consider the psychological state of the participants in each of our experiments. In the virtual office evacuation experiment, participants were under significant time pressure, but were still distanced from the emergency because the scenario was mediated by a computer. Participants knew that they could not be harmed, so they were able to take a rational approach to finding the best exit. In contrast, participants in the physical experiment could not know for sure that they would not be harmed in the emergency. Even those who reported that they knew the emergency was part of the experiment could not possibly be certain of this fact until they were debriefed

by experimenters. Consequently, participants in the physical experiment would search for any good solution in this scenario. A robot that appears to be designed to guide in exactly this circumstance would appear as a good solution to such a participant. Instead of taking a reasoned approach to finding the best possible exit, participants followed a less deliberate and more reactive approach to find the first exit. We believe that this different type of reasoning coupled with the previously mentioned physical embodiment of a lighted, gesturing robot explains the difference between the virtual and physical experiments. Note that this explanation presents a concerning conclusion: physical robots are more likely to cause Case 1 overtrust than virtual robots.

## 6.8 Thoughts on Future Work

Many avenues of future work are suggested by this research. In the previous section, we state that a robot may be able to communicate its errors to properly calibrate the level of trust a person has in it. Even if a robot knows it is malfunctioning, how does it inform nearby people that it should not be trusted? Will frightened evacuees listen to the robot when it tells them to stop following it and find their own way out? Can a non-verbal robot communicate such a message with its motion alone? Future research could begin by defining communication modalities to inform people of the robot's error. These could then be used to try to limit Case 1 overtrust.

Many participants reported that they followed the robot because it was labeled as an emergency guidance robot. This was intentional in order to create a trustworthy robot, but it would be interesting to see if participants would still follow the robot without the label. It will be difficult to inform participants that the robot is guiding them towards an exit without implying that the robot was designed for that purpose, but the results would help to inform robot designers of the importance of proper labeling. This, again could be used to test the amount of Case 1 overtrust that people place in robots based on their appearance.

Case 2 overtrust is somewhat harder to detect because it is based on the amount of risk a person believes is present in a situation. Perhaps future experiments can use physiological data such as heart rate or galvanic skin response to measure stress in a participant during an experiment and correlate that to trust decisions.

The fundamental difference between our virtual and physical experiments seems to be that participants in the virtual experiments used logical reasoning to find the best route to an exit while participants in the physical experiments experienced a fight-or-flight response and sought the first exit they could find. It seems unlikely that we can test a fight-or-flight scenario in a virtual experiment, but it should be possible to influence participants to make a logical choice during a physical experiment. Participants in the virtual experiment were under an explicit time pressure to find an exit, as opposed to an implicit one in the physical experiment. Recreating this in a physical experiment by telling the participants to act as if they were in an emergency and then visibly recording their time to an exit could cause participants

to think in a more logical manner. At that point, participants would think about beating the clock, instead of finding the first exit. This might produce behavior similar to that in our virtual experiments.

This book concerns methods and problems with intelligent systems taking control from humans before or after humans commit errors. Our research indicates that humans may be all too willing to trust a robot, even in emergency scenarios. This is an important issue for designers to consider when creating systems that responsibly help humans recover from errors.

**Acknowledgements** Support for this research was provided by the Motorola Foundation Professorship, the Linda J. and Mark C. Smith Chair in Bioengineering, Air Force Office of Sponsored Research contract FA9550-13-1-0169 and Georgia Tech Research Institute.

## References

- Bainbridge, W. A., Hart, J. W., Kim, E. S., & Scassellati, B. (2011). The benefits of interactions with physically present robots over video-displayed agents. *International Journal of Social Robotics*, 3(1), 41-52.
- Desai, M., Kaniarasu, P., Medvedev, M., Steinfeld, A., & Yanco, H. (2013). Impact of robot failures and feedback on real-time trust. *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*, (pp. 251-258). Tokyo, Japan.
- Kim, P. H., Dirks, K. T., Cooper, C. D., & Ferrin, D. L. (2006). When more blame is better than less: The implications of internal vs. external attributions for the repair of trust after a competence-vs. integrity-based trust violation. *Organizational Behavior and Human Decision Processes*, 99(1), 49-65.
- Robinette, P., Howard, A. M., & Wagner, A. R. (2015). Timing is Key for Robot Trust Repair. *International Conference on Social Robotics* (pp. 574-583). Paris: Springer International Publishing.
- Robinette, P., Wagner, A. R., & Howard, A. (2014). Assessment of Robot Guidance Modalities Conveying Instructions to Humans in Emergency Situations. *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 14)*. Edinburgh, UK.
- Robinette, P., Li, W., Allen, R., Howard, A. M., & Wagner, A. R. (2016a). Overtrust of Robots in Emergency Evacuation Scenarios. *11th ACM/IEEE International Conference on Human-Robot Interaction*. Christchurch.
- Robinette, P., Wagner, AR., & Howard, AM. (2016b). Assessment of robot to human instruction conveyance modalities across virtual, remote and physical robot presence. *Robot and Human Interactive Communication (RO-MAN), 2016 25th IEEE International Symposium on*. (pp. 1044-1050). New York City: IEEE.
- Robinette, P., Wagner, A. R., & Howard, A. M. (2016c). Investigating Human-Robot Trust in Emergency Scenarios: Methodological Lessons Learned. In R. Mittu, D. Sofge, A. Wagner, & W. Lawless, *Robust Intelligence and Trust in Autonomous Systems* (pp. 143-166). Boston: Springer.
- Robinette, P., Wagner, A. R., & Howard, A. M. (2017). The effect of robot performance on human-robot trust in time-critical situations. *IEEE Transactions on Human-Machine Systems*, PP(99), 1-12.
- Salem, M., Lakatos, G., Amirabdollahian, F., & Dautenhahn, K. (2015). Would You Trust a (Faulty) Robot?: Effects of Error, Task Type and Personality on Human-Robot Cooperation and Trust. *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, (pp. 141-148). Portland.

- Schweitzer, M. E., Hershey, J. C., & Bradlow, E. T. (2006). Promises and lies: Restoring violated trust. *Organizational behavior and human decision processes*, 101(1), 1-19.
- Wagner, A. R. (2009). *The Role of Trust and Relationships in Human-Robot Social Interaction*. Ph.D. diss., School of Interactive Computing, Georgia Institute of Technology, Atlanta, GA.
- Wagner, A. R., & Robinette, P. (2015). Towards Robots that Trust: Human Subject Validation of the Situational Conditions for Trust. *Interaction Studies*, 16(1), 89-117.

# Chapter 7

## Research Considerations and Tools for Evaluating Human-Automation Interaction with Future Unmanned Systems

Ciara Sibley, Joseph Coyne, and Sarah Sherwood

### 7.1 The Current Environment and Future Vision

The last 15 years have seen a proliferation in the use of unmanned systems within the Department of Defense (DoD). The DoD inventory of unmanned systems increased 40-fold between 2002 and 2010, by which time they accounted for 41% of all DoD aircraft (Gertler 2012). This rapid growth has been paralleled by advances in automation and reliability, which will soon cause the role of the human operator to transition from one of manual control of specific subsystems (e.g., payload or avionics) to supervisory control of multiple unmanned aerial vehicles (UAVs). The supervisory control paradigm leverages what humans do best (goal setting) and what machines do best (routine execution of control actions based on sensed feedback) to improve human-machine system performance across a variety of domain settings (Sheridan 2012). The shift towards supervisory control is already happening in many of today's unmanned systems, where stick-and-rudder piloting is being replaced by autopilot systems capable of executing routes based upon waypoints.

Despite this paradigm shift toward supervisory control, most current UAV operations require three human operators to manage one UAV, where each operator maintains one of three distinct roles: Mission Commander (MC), Air Vehicle Operator (AVO), and Payload Operator (PO). In a typical team set up, the MC is primarily responsible for: mission management; requesting access to controlled airspace; communicating with external customers and interested parties (effectively consumers of the services provided by the UAV); and disseminating information to the AVO and PO. The AVO is principally responsible for: navigating; monitoring the vehicle's

---

C. Sibley (✉) • J. Coyne  
Naval Research Laboratory, SW, Washington, DC, USA  
e-mail: [Ciara.Sibley@nrl.navy.mil](mailto:Ciara.Sibley@nrl.navy.mil); [Joseph.Coyne@nrl.navy.mil](mailto:Joseph.Coyne@nrl.navy.mil)

S. Sherwood  
Embry-Riddle Aeronautical University, Daytona Beach, FL, USA  
e-mail: [Sherwoo9@my.erau.edu](mailto:Sherwoo9@my.erau.edu)

health and status; and ensuring the vehicle successfully travels from waypoint to waypoint. The PO primarily manages the system's sensors and relays relevant information to the MC and/or customer. The DoD recognizes that the current UAV manning requirements and team structure is sub-optimal; it is resource intensive and does not scale, particularly when compared to manned military aircraft such as the F/A-18-E Super Hornet, which has a crew compliment of one and can accomplish a wider range of missions.

The tasking demands for current UAV operators are highly variable and often unbalanced across team members. This is partly attributable to automation performing the majority of one of the operator's roles (MC, AVO, PO) during specific mission phases (i.e., takeoff, enroute, over target, landing). For example, missions requiring a UAV to observe an area of interest for an extended period of time may require no interaction from the AVO (since loitering can be performed automatically), but the PO must continuously move the camera sensor from one object to another. There are also many situations in which the entire crew is either engaged or underutilized. For example, once a system is airborne, little to no human input is required during a wide area surveillance and mapping mission to gather updated high-resolution imagery of pre-determined area. In contrast, a mission providing direct support to troops in contact and/or requiring weapons release necessitates substantial human input and attention from all UAV operator members/roles. All of these missions currently call for the same manpower, despite the team in the former mission scenario being highly underutilized. Concerns about how to address emergency situations is one of the primary drivers in maintaining the same manning requirements across all missions conducted with the same vehicle.

This inefficiency and inflexibility has influenced the DoD's desire to invert the ratio of operators to UAVs (Department of Defense 2013). Furthermore, the 2015 Naval S&T Strategy calls for "the development of a distributed system of heterogeneous unmanned systems relying on network-centric, decentralized control that is flexible in its level of autonomy with the ability to get the right level of information to the right echelon at the right time" (Office of Naval Research 2015). Decentralized, flexible control would require new service-based operator control paradigms, in which operators perform varied tasks across multiple platforms at different mission stages, as required. The result will be shared control of a greater number of unmanned systems that are dynamically assigned to operators, based on theater mission requirements rather than vehicle requirements. This is in stark contrast to the current static control paradigm of one operator managing a subsystem of one specific platform for the entirety of a single mission.

A decentralized, flexible system of control presents large research questions as it represents a significant change in how individuals would interact with autonomous systems. For example, questions regarding vehicle or subsystem control hand-offs, as well as authority and responsibility for the platform will need to be addressed. These changes not only impact how a vehicle is controlled but also what information an operator will need to be aware of to support mission requirements. Failure to understand how these new paradigms and systems impact the operator could lead to significant increases in human error. Extensive testing will be required before imple-

menting any modifications to UAV automation, control structure, or crew complement. This chapter focuses not only on the changes associated with the shift to supervisory control, but also on how to measure performance in this new environment. Errors occur not only as a result of measurable actions by an operator, but also because the operator may have inappropriately focused attention. This chapter highlights the importance of understanding how systems impact operator state and methods for measuring operator state and awareness.

## 7.2 Calibrating Trust in Automation

Unlike most commercial autonomous systems, which are designed for use in benign environments, autonomous systems designed for military application must be able to function in complex, unpredictable environments with the possible presence of an adversary committed to defeating or interrupting normal operations. In such high-stakes environments, it is critical that the operator be able to trust the automation. One barrier to trust is that automation lacks human-analog sensation, perception, and decision-making. The different sensors and data sources that inform the automation's decision-making processes are not the same as those of its human operator, and it could therefore be operating on different contextual assumptions. Moreover, machine learning, reasoning, and decision-making can take vastly different paths to that of humans, which could lead human operators to question the trustworthiness of their machine partners (Defense Science Board 2016).

The formation of human trust in automation begins at design time with the establishment of what the automation can and cannot do, in addition to what it should and should not do. Problems with automation tend to occur when system designers automate what is easy, or seek to automate functions to the greatest extent possible (i.e., the “technological imperative” in Sheridan (2000)). Although automation provides clear benefits, poorly designed automation can cause performance problems for both man and machine. Parasuraman and Riley (1997) describe people's interaction with automation as “use, misuse, disuse, or abuse,” and the complications that arise across all four categories. For example, misuse occurs when an individual overrelies on an automated system, which can result in suboptimal monitoring behavior and decision-making biases. High levels of automation are associated with decreased operator SA, which can lead to delayed operator reengagement with a system if and when its automation fails (Endsley and Kaber 1999). Automation abuse occurs when a system designer automates functions without considering the role of the human operator or how it will impact performance.

Once automation is determined necessary, the system design should include real-time indicators of automation's reliability. Such indicators will enable operators to calibrate their trust in the system and intercede when the operational environment exceeds the original design envelope or assumptions. However, a basic awareness of system and/or environmental status is not enough; the automation must be able to adapt to its environment and mission context. It must also effectively communicate



changes in its reliability without increasing operator workload and decision-making time. System design should include sufficient contextual indicators so that the system is predictable and allows the operator to intervene in a timely and effective manner if the environment exceeds the design envelope of the automation (Defense Science Board 2016).

The transition to UAV supervisory control will require a suite of new capabilities to include better decision support, alerting, and monitoring tools. These new automated tools, as with all proposed automated UAV subsystems, must be robust and their effects on the overall system calculable. Furthermore, their actions must be predictable, transparent and directly observable by their human supervisors. All these factors are critical to the establishment of operator trust in any new system, capability, or tool.

### 7.3 DoD Plans and Guides

The DoD established the UAS Control Segment (UCS) working group to develop an architecture for the control systems of future UAVs, utilizing the principles of a service-oriented architecture (SOA). The SOA approach will enable future control platforms, such as the common control station (CCS), to incorporate a modular design allowing for components (i.e., services) to be easily replaced. This future design model for control stations is very different from today's UAV control stations, which the DoD originally procured as combined ground control stations and unmanned vehicles. This method of procurement led to stove-piped systems that are incompatible with each other, which increases training and costs and limits innovation (Chanda et al. 2010). On the other hand, the future SOA model will enable rapid fielding of new tools, which could be risky if their behavior isn't comprehensively understood and tested across all situations.

In addition to new software design considerations, the DoD and its NATO allies are moving toward standardizing the unmanned systems' user interface (i.e., common control layout) and increasing interoperability (i.e., ability for a ground station to communicate with multiple platforms). This goal, and the required communication protocols, are outlined in NATO's Standardization Agreement (STANAG) 4586 (NATO 2012). STANAG 4586 discusses the need for interface standardization, but does not provide details on how that interface should look. The DoD (Office of the Secretary of Defense 2012) released a style guide to provide system designers' recommendations for how to display information within a UAV control station. However, they do not address the bigger question about what information should be displayed, particularly as automation increases and direct operator interaction decreases. For example, while an attitude indicator provides useful information to a pilot directly controlling a platform, it is unclear what value, if any, it provides when flying by waypoint. If the information that needs to be conveyed to a UAV supervisory control operator is indeed different, developing and testing new data visualizations could

potentially streamline UAV operator interactions with the system (Defense Science Board 2016).

As platforms become more interoperable, different users and operators will have access to different levels of direct and indirect interaction with an unmanned system. Although the five levels of interoperability defined in STANAG 4586 were meant to outline communication requirements between a control station and an unmanned vehicle, they are also important in defining information needs for different types of users. For example, to support mission requirements, an operator might subscribe to information (i.e., sensor) feeds (level 1–2), assume direct control of specific payloads (level 3), and/or redirect an asset’s path (level 4–5).

Highly interoperable systems and a flexible control paradigm could result in high levels of task switching across different vehicles, which could impair a user’s SA and subsequent decision making. The increased platform and sensor hand-offs envisioned in a distributed control environment enhance mission flexibility, but also increase the potential for error during control transfers. Consequently, research is needed to identify the information requirements for acquiring and maintaining high levels of performance and SA during hand-offs and when managing an asset for a limited timeframe.

To date, a common concept of operations (CONOPS) does not exist for future UAV supervisory control missions, i.e., it is unclear how teams of operators will interact with future UAV systems. There are many basic questions that remain unresolved: How many vehicles should an operator manage? Will there still be specialized roles (e.g., payload supervisor)? Will operators be cross-trained to manage all aspects of a system? Will operators be assigned to a vehicle (reflecting current operations) and/or will operators be assigned to a mission? Alternatively, will the specific mission context dictate an operator’s tasking?

Identifying the ideal CONOPS for a particular mission requires a simulation environment capable of representing a range of different missions and situations (e.g., operating in bad weather, responding to an engine failure, operating in low-bandwidth regions, etc.). Furthermore, a set of assessment metrics is needed to enable systematic comparison of performance across different contexts and to understand the consequences associated with fielding new technologies. Experimenting with different models of flexible control is a critical next step toward realizing the DoD’s goals. It is especially important to establish benchmarks for human, system, and mission performance since novel capabilities will be introduced over time. This ability to assess performance is surprisingly challenging.

## 7.4 Supervisory Control Research and Testing Environments

Since concurrent control of multiple UAVs has not been fielded in any operational context, the research community has developed several test beds to simulate some of the different tasks an operator might have to perform. The two most frequently used platforms are the Adaptive Levels of Autonomy (ALOA) and the Research

Environment for Supervisory Control of Heterogeneous Unmanned Vehicles (RESCHU) test beds (Nehme 2009; Johnson et al. 2007). Experimentation conducted within these environments has provided valuable guidance to consider in the design of future UAV automation and human-automation interaction.

### ***7.4.1 The Adaptive Levels of Automation Test Bed and Research***

The ALOA testbed was developed for the Air Force Research Laboratory to assess how different Levels of Automation (LOA) impact performance in a simulated multiple-vehicle supervisory control environment (Johnson et al. 2007). Sheridan and Verplank (1978) defined ten LOAs that have been widely utilized by the research community to build a taxonomy of performance implications under different circumstances. Parasuraman et al. (2000) extended these levels to include four stages of information processing. The different tasks within ALOA are meant to address both the different stages of information processing as well as the original ten-level hierarchy (Table 7.1). Within ALOA, the LOA for four tasks (weapon release authorization, image analysis, task allocation, and autorouting) can be set by the experimenter; dynamically controlled by the operator; or automatically adapted by the system in real time according to algorithms based upon either workload, performance, or time.

The ALOA interface includes a chat window that presents the rules of engagement (ROE) and mission updates, a scrolling ticker that displays warnings and system updates, color-coded vehicle health and status indicators, a map display, and visual and aural pop up threat indicators. ALOA also includes planning tools to help users decide on a route; reallocate tasks; assess potential impacts of new threats; and avoid pop up threats, such as surface-to-air missile (SAM) shots (Johnson et al. 2007).

The research community has primarily focused on how different levels of automation impact users' SA, mental workload, and trust since all ultimately impact task and goal/mission performance (Parasuraman et al. 2008). Calhoun et al. (2009) used ALOA to examine the impact of three LOAs (low, medium, high) on performance in the routing task. In this experiment, automation had low reliability (66% accurate) and was not trusted by the operators regardless of the level of automation employed. In fact, operators took significantly longer to complete the task at the highest level of automation since they always initiated a new re-planning task to override the automation. The results demonstrate that, when automation is unreliable, humans are unlikely to use the system (i.e., disuse).

Kidwell et al. (2012) used ALOA to compare the use of adaptive automation (which changes LOA based upon performance) and adaptable automation (user selected LOA) within the four aforementioned tasks in the ALOA testbed. The automation was reliable 90% of the time and each task had three LOAs. This study found mixed performance effects for the different tasks, but the effect sizes were

**Table 7.1** Ten levels of human interaction with automation and their use in ALOA

Level	Description of system output	Description of automation	ALOA task(s)
10	The computer decides everything, acts autonomously, ignoring the human	Fully automatic	Weapon release authorization, image analysis, allocation, and autorouting
9	Informs the human only if it, the computer, decides to		
8	Informs the human only if asked		
7	Executes automatically, then necessarily informs the human	Automatic with feedback	Weapon release authorization, image analysis, and autorouting
6	Allows the human a restricted time to veto before automatic execution	Veto	Weapon release authorization, image analysis (single and multiple options), and autorouting (single and multiple options)
5	Executes a suggestion if the human approves	Consent	Weapon release authorization, image analysis (single and multiple options), and autorouting (single and multiple options)
4	Suggests one alternative		
3	Narrows the selection down to a few options	Multiple options	
2	Offers a complete set of decision/action alternatives		Image analysis, and autorouting
1	Offers no assistance; human must make all decisions and take actions	Manual	Weapon release authorization, image analysis, allocation, and autorouting

*Note.* Adapted from (1) “A Model for Types and Levels of Human Interaction with Automation,” by Parasuraman et al. (2000). (2) “Testing adaptive levels of automation (ALOA) for UAV supervisory control” by Johnson et al. (2007)

very small and did not suggest any significant advantage for either adaptive or adaptable automation. Despite this, participants reported feeling significantly more confident in their decisions with the adaptable system.

Calhoun et al. (2011) identified an automation level transference cost (i.e., a performance decrement associated with having different levels of automation on two related tasks). Specifically, they found a significant increase in the time required to complete the allocation task when it and the route planning task had different LOAs. Furthermore, the study included two groups of participants subject to different automation reliability levels (80 and 100%). Both groups were trained to identify and correct errors. The 100% reliability group had an extra 20-min experimental block in which one error occurred and none of the participants were able to detect the automation error. In contrast, participants in the low reliability group detected errors 93% of the time. The results suggest a clear case of overreliance on an automated

system (i.e., misuse), as well as the inability to maintain high levels of monitoring performance over a sustained period of time.

### ***7.4.2 The Research Environment for Supervisory Control of Heterogeneous Unmanned Vehicles Test Bed and Research***

RESCHU is a UAV and unmanned underwater vehicle (UUV) supervisory control test bed originally developed by the Human and Automation Laboratory at MIT. RESCHU's simulated ground control interface consists of a map display, camera window, vehicle control panel (that displays vehicle health and mission information), and a mission timeline (that gives the estimated time of arrival to areas of interest). An operator is tasked with assigning objectives to vehicles, avoiding hazard areas which randomly appear on the map, and completing a visual search task.

The RESCHU test bed is particularly valuable for research focused on how vehicle heterogeneity affects operator performance. Operators can control a group consisting of up to three types of vehicles: a high-altitude long-endurance (HALE) UAV, a medium-altitude long-endurance (MALE) UAV, and an unmanned underwater vehicle (UUV). The vehicles have variable speeds (UUVs are slower than UAVs) and capabilities (HALE UAVs are used to locate new targets within an area of interest, while MALE UAVs and UUVs are used to acquire these pre-determined targets) (Nehme 2009). In addition, RECHU can be used to conduct research focused on trust in automation since it employs a sub-optimal route planner. The route planner, by sometimes failing to assign the best paths and vehicle-target assignments, seeks to replicate the performance of real-world automation and serves as an additional source of operator workload since operators must reassign vehicles.

RESCHU has been used to assess the effect of UAV control architectures on operator workload and performance (Cummings et al. 2014). The vehicle-based RESCHU interface employs a centralized control architecture, in which a single operator individually tasks multiple UAVs. The task-based RESCHU interface employs a decentralized architecture that requires the operator to convey high-level goals (i.e., a task list) to an automated mission and payload manager, which then decides how best to distribute the tasks among multiple UAVs. In general, decentralized control schemes are favored because they eliminate the UAV operator and their ground control station as a single point of system failure. In addition, Cummings et al. (2014) found they are more robust to delayed operator action and lapses in SA. However, decentralized control schemes are generally less resilient to unexpected events and emergent system behavior. Given the limitations of both control architectures, Cummings et al. (2014) determined that a hybrid mix would likely be best for operational use.

Researchers have also used RESCHU to investigate human-automation performance questions. For example, Cummings and Nehme (2009) demonstrated that

keystroke analysis could be used within the test bed to create a metric of operator utilization during a supervisory control task. The researchers defined utilization as the percentage of time the operator was “busy” interacting with the system and performing tasks; they did not consider monitoring or scanning (i.e., updating SA) as time when the operator was busy since no system interaction was required. Using this metric, they identified that performance was best when operators were at a middle range of utilization with performance dropping at both ends of the scale, consistent with well-documented findings on the effect of arousal (e.g., mental stress, cognitive workload, mental effort) and performance (Kahneman 1973).

Furthermore, Ratwani et al. (2010) demonstrated how eye tracking data collected within the RESCHU testbed could be used to predict when an operator was not attending to a vehicle’s flight path and consequently about to make an error. Follow-up research focused on how this information could be used to improve alerting by informing the operator of potential problems on which they had not yet fixated, as opposed to employing constant, threshold-based alerting, which is subject to alert fatigue.

## 7.5 Supervisory Control Research Limitations and Challenges

To date, the research conducted in RESCHU and ALOA has emphasized high operator task load. High levels of tasking enable researchers to collect ample performance data to confirm or refute hypotheses or to build predictive models of performance. However, these experimental designs cannot address problems associated with boredom and underload, or transitioning between high and low levels of tasking. Future supervisory control operators are expected to experience more downtime due to increased automation. Indeed, current UAV operators describe UAV operations as 90% boredom, claiming that staying awake can be a challenge, particularly during sustained surveillance missions (Button 2009).

Low task load experiments present a challenge to researchers since traditional performance metrics (i.e., reaction time and accuracy) are limited to the number of interactions a user has with the system. For example, discrete performance measurements can only be gathered during a monitoring task if an event occurs. The goal of RESCHU’s surveillance missions is to detect and identify as many targets as possible while avoiding pop-up threats. This provides a near continuous measurement of performance that is ideal for research, but neither reflects the actual tasking of future operators nor explores the variable workload experienced in a real UAV environment. Even Ratwani et al.’s (2010) eye tracking work within RESCHU was dependent upon frequently occurring time critical obstacles. Furthermore, high task load levels represent a narrow range of UAV mission contexts; there are many contexts in which a UAV operator will have limited interaction but must sustain attention and SA for extended periods (e.g., while monitoring a sensor feed).

In addition to focusing on high workload situations, tasking within experiments was chosen to have clear, measurable performance outcomes. This is ideal for analyzing experimental data, but the real world is messy; operators can make poor decisions that yield positive outcomes and vice versa. Making decisions under uncertainty is a critical challenge UAV operators confront during missions, however, research is limited in this area. RESCHU and ALOA utilize random events that require operators to update their plans, but neither incorporate uncertainty nor the risk associated with alternative courses of action.

Assessing levels of automation and display formats within a single mission context limits the generalizability of the results to future supervisory control operations. In order to apply scientific knowledge of supervisory control toward future systems, it is essential to assess tools and concepts within representative, complex, synthetic environments that can model the broad range of scenarios and contexts an operator could encounter (e.g., denied/degraded communications, sustained monitoring, and target-asset allocation and decision making under uncertain conditions).

## 7.6 Assessing Human-Automation Performance

In the operational environment, “performance” is often considered primarily in terms of outcomes, yet an operator’s interaction with the system largely influences mission success. A 2012 *U.S. Unmanned Aerial System Report* to Congress stated human causal factors were present in approximately 68% of UAV mishaps (Gertler 2012; Williams 2004). Many of these incidents were attributable to factors such as extremes in workload leading to channelized attention and/or lapses in SA, as well as generally poor operator interface design causing automation state confusion and alarm fatigue (Chen et al. 2011; Giese et al. 2013; Parasuraman and Manzey 2010; Parasuraman et al. 2008). Limiting metrics to traditional performance-based measures of accuracy and response time will provide only a partial understanding of human performance issues with new automated technologies, since the operator’s role is often to monitor these systems.

There are many extended periods of time during UAV operations where traditional operator performance metrics (i.e., reaction time and accuracy) cannot be obtained, such as when a vehicle is enroute to an objective or loitering over a target for an extended period of time. During this time, the pilot’s task is to monitor/scan the system’s sensors and maintain a high level of SA. He/she has no direct interaction with the system and, therefore, no performance measures can be assessed. This is concerning given the future unmanned vehicle control paradigm of increased automation where problems with degraded SA are increasingly likely. Further, studies have shown decreases in SA can increase the time for an operator to re-engage with a system (Endsley and Kaber 1999).

One solution for gathering a more complete picture of operator performance is to augment traditional metrics of mission performance with measures of operator state, which can vary throughout the mission. Within the context of this chapter,

operator state is meant to encompass a broad range of psychological constructs including attention and mental workload. The ability to assess an operator's state throughout a mission provides valuable data for predicting mission success. This is particularly true in situations where the operator's interaction with the system is limited, but the few interactions that do occur could be critical. For example, if an operator is fatigued and has not scanned their display panels for several minutes, he/she could miss a piece of chat information revealing a nearby high-priority target. Furthermore, evaluation of new automation and CONOPS must be conducted across a range of mission contexts, which include factors such as: mission phase, requirements, operating area (e.g., contested vs. uncontested), rules of engagement, type/number of assets, priorities, environmental constraints, time restrictions, etc.

In order to provide a comprehensive evaluation of new automation, human-automation performance should be considered as a composite of operator state (process) and performance (outcome). For example, new automation may enable a positive outcome, but to the detriment of operator SA. This could lead to significant problems if an emergency occurs, requiring an operator to intervene. Assessing performance in terms of both outcome and process enables identification of these potential trade-offs and can be used to diagnose deficiencies, inform mitigations (e.g., designing tools which foster high levels of operator SA and mission performance), and provide better metrics for comparing automation tools and technologies.

Table 7.2 demonstrates how human-automation performance could be assessed for an intelligence, surveillance and reconnaissance (ISR) mission. Herein, operator state is composed of engagement (active attention and effort) and awareness (comprehension and knowledge), while task performance is composed of efficiency (reaction time) and effectiveness (accuracy). An abundance of subjective and objective measures can be used to inform the operator state elements, such as questionnaires (workload and SA), user interactions (keylogging, mouse clicks), and eye tracking data (dwell times, fixation locations, pupil size, etc.).

### ***7.6.1 The Value of Eye Tracking***

As discussed in the previous section, there are many supervisory control situations lacking outcome-based measures to assess human-automation interaction. Remote, off-the-head eye trackers are a powerful option for gathering information about an operator's attention allocation, fatigue, cognitive workload, and SA. Unlike outcome-based measures, which are not available during monitoring tasks, these metrics are employable throughout the mission. Eye tracking can provide a wealth of information about an individual's state. For example, fixation analyses can be used to predict errors due to lapses in attention (Ratwani et al. 2010) and can serve as a measure of SA (van de Merwe et al. 2012). In addition, frequency and duration of blinks and percent eyelid closure are reliable indicators of fatigue (Caffier et al. 2003). Furthermore, pupil data is a valuable indicator of cognitive workload (Tsai





**Table 7.3** Technical specifications for three first-generation low-cost eye tracking systems

	Gazepoint GP3	Eye Tribe	Tobii EyeX
Cost	\$495	\$99	\$139
Sampling rate	60 Hz	30/60 Hz	60 Hz (estimated)
Accuracy	0.5°–1.0°	0.5°–1.0°	–
Max. display size	24 in.	27 in.	27 in.
Eye position data	Left and right	Left and right	Combined
Pupil size data	Pixels and mm	Pixels	None

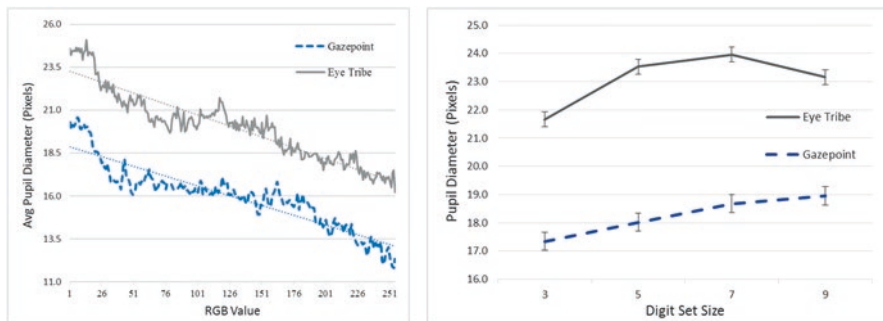
et al. 2007; Beatty and Lucero-Wagoner 2000; Sibley et al. 2011). All of these eye tracking-based measures of operator state have been linked to performance.

In recent years, a number of low-cost eye tracking systems have become available; these systems, which are designed for use with single displays, range from approximately \$100–500 and offer a streamlined setup process. The first generation of low-cost eye trackers includes the Tobii EyeX, Gazepoint GP3, and Eye Tribe. The Gazepoint GP3 and the Eye Tribe collect data on gaze position and pupil size (in pixels) for both eyes. The Tobii EyeX only provides gaze position averaged across both eyes and was designed for entertainment purposes; the user agreement does not permit data collection and analysis. A summary of the technical specifications for these three systems is provided in Table 7.3.

Coyne and Sibley (2016) found that gaze data collected using the Eye Tribe and Gazepoint GP3 systems provided sufficient accuracy and precision to be useful for Human Factors research and, on 24-inch or smaller displays, tracked gaze position almost as well as the high-cost Smart Eye Pro system. Similarly, Ooms et al. (2015) found that the gaze accuracy and precision of Eye Tribe was comparable to the SMI RED 250, an established, high-end system. Funke et al. (2016) found similar results regarding the accuracy and precision of Tobii EyeX and Eye Tribe but experienced more frequent data quality problems. They cautioned that missing data could affect estimates of the number and duration of fixations, saccadic rates, and blinks, all of which are commonly used in Human Factors research.

Although a number of studies have investigated the accuracy and precision of gaze data collected using low-cost eye trackers, less research has been conducted assessing the ability of these devices to collect non-gaze data, such as pupil size. Coyne and Sibley (2016) found the Eye Tribe and Gazepoint GP3 systems sufficiently sensitive to capture changes in pupil size in response to both mental effort (during a memory task) and screen luminance (Fig. 7.1).

Overall, although low-cost eye trackers are not quite as accurate and experience more data quality problems relative to high-end systems, research suggests that these devices may be able to provide meaningful data in applied settings, including the control of unmanned systems. Additionally, the cost of these new systems makes them readily accessible to a larger number of researchers. Thus, researchers should carefully consider the relative strengths and weaknesses of the various systems and their suitability for their specific research effort (Funke et al. 2016; Holmqvist et al. 2011, 2012).



**Fig. 7.1** Pupillary response to increasing luminance (*left*) and workload (*right*) as measured by the Gazepoint GP3 and Eye Tribe systems. *Note.* Reprinted with permission from “Investigating the Use of Two Low Cost Eye Tracking Systems for Detecting Pupillary Response to Changes in Mental Workload,” by Coyne and Sibley (2016)

## 7.7 Supervisory Control Operations User Testbed (SCOUT) Overview

The Naval Research Laboratory (NRL) developed the Supervisory Control Operations User Testbed (SCOUT™) as a tool to address a broad range of supervisory control research questions. Two specific areas of research that SCOUT was designed to investigate are decision making under uncertainty and sustained attention; two topics which have not been emphasized within the existing supervisory control test beds and research. SCOUT was iteratively designed based on observation, interviews, and feedback from current UAV operators within training and testing environments at multiple locations around the United States. These operators were asked to describe typical tasking in addition to challenges, common errors, and system abnormalities experienced while controlling contemporary UAVs. Furthermore, they were asked to envision future UAV supervisory control operations, and how the aforementioned challenges, errors, and abnormalities might manifest in this environment.

Utilizing this information, SCOUT was designed to include the primary components of contemporary UAV control and to simulate the tasks future UAV operators might perform while supervising multiple vehicles. During a SCOUT scenario, an operator is tasked with managing three heterogeneous UAVs. In order to meet mission goals, users must decide how to best allocate the UAVs to locate targets while simultaneously completing a number of subtasks, including: maintaining communication with command and intelligence personnel via chat; updating UAV parameters and routes; and monitoring sensor feeds and airspace. Points are assigned to various actions based on their mission priority and the overall goal is to obtain as many points as possible.

A key capability within SCOUT is the ability to capture and synchronize data from multiple sources within a relational database. SCOUT records all task and

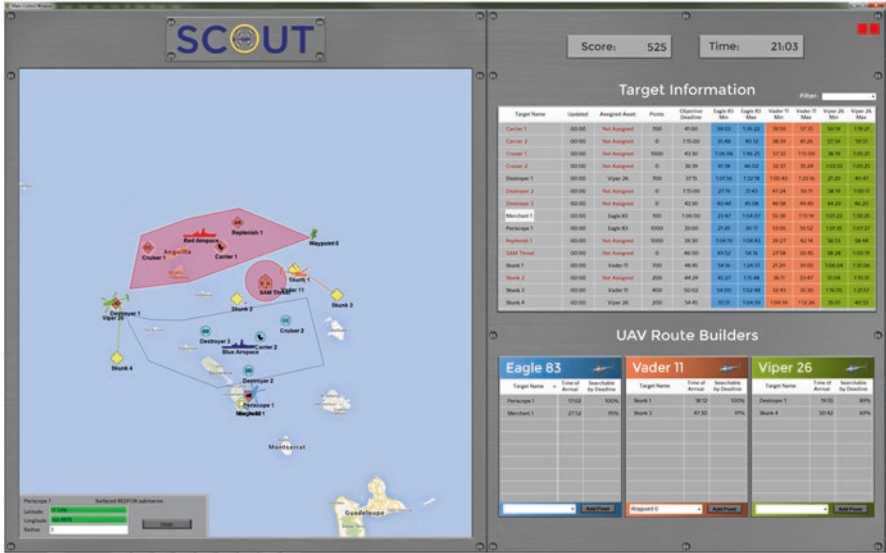


Fig. 7.2 SCOUT route planning (left) screen

mission information, in addition to user behaviors (e.g., keystrokes, mouse clicks, eye gaze) and indicators of user physiological state (e.g., pupil size, heart rate, respiration rate). The system currently supports SmartEye Pro, GazePoint, and EyeTribe eye tracking systems (Coyné and Sibley 2016). This data integration enables both real time and post-hoc analysis of the user’s performance, eye tracking, physiological, mouse, and keystroke data. These additional physiological data help address the challenge of continuously assessing operator performance by providing continuous information about the user while a mission is being executed.

Streaming access to the user’s physiological data allows the experimenter to compute, for example, how long it takes an operator to look at and fixate on a new chat message, or conversely, to not notice a message. The experimenter can also observe scan patterns and assess, for example, whether a user is becoming fatigued and not adequately scanning information panels. Additionally, monitoring pupil size, gaze, and performance data during a period of high task load can provide information about a user’s mental workload, where he/she is allocating attention, and how these factors relate to task and mission performance.

SCOUT is available in both single-monitor and dual-monitor configurations. In the dual-monitor set-up, the left screen is primarily used for route planning (Fig. 7.2). The Target Information table and UAV Route Builder boxes provide operators with estimated search times for each target, target point values (which indicate mission priority), target deadlines, the size of target search areas, and the percent of those areas that can be covered by each UAV before the target deadlines.

Each SCOUT mission involves a variable degree of uncertainty. Operators do not know the exact location of the targets within their respective search areas. A UAV

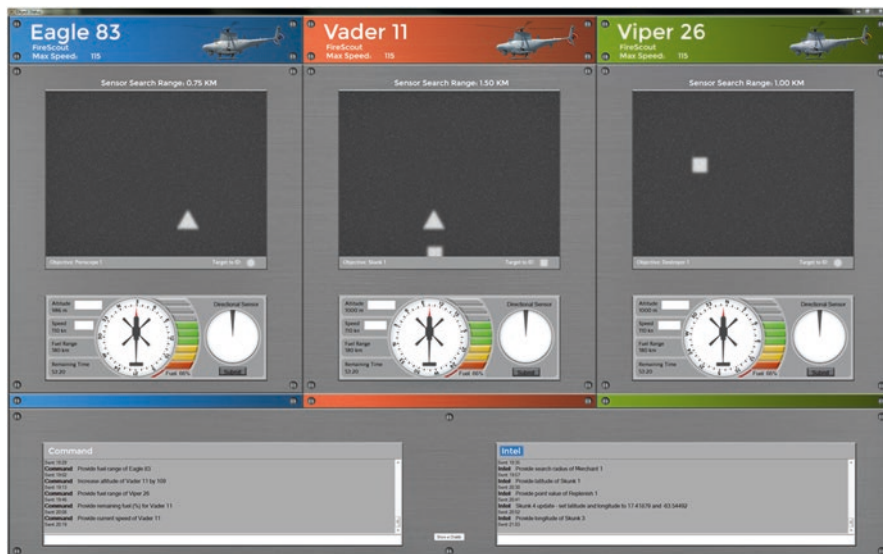


Fig. 7.3 SCOUT vehicle status (right) screen

might find a target after searching only 1% of its search area, but it could also require 100% coverage of the search area to locate a target. Moreover, the entire search area might not be traversable by the target deadline (when the intelligence expires and the location estimate becomes too uncertain to be useful). An example of a SCOUT mission with a high degree of uncertainty might involve targets with large search areas, short deadlines, and variable point values, all of which can be manipulated within a drag-and-drop scenario generator. Additional sources of uncertainty include whether or not operators will be granted access to restricted operating zones (ROZs), which are indicated by the outlined and/or red-shaded areas on the moving map display, and the similarity of distractor targets to the actual target on the simulated payload task.

The simulated payload task is located on the right screen along with other vehicle-centric information, such as fuel status, altitude, and speed. While a UAV is actively loitering over a target search area, objects will appear in that UAV's sensor feed. In the sample mission depicted in Fig. 7.3, Eagle 83 is searching for Periscope 1, which will appear in the feed as a circle. The operator's task is to click on any circular targets, which could be Periscope 1, and to ignore any other objects. In this case, the distractor objects (triangles and squares) are quite distinct from the target of interest. Additional uncertainty and complexity could be introduced into the scenario by using distractor objects similar in appearance to the circular target.

The complexity of a SCOUT scenario can be further altered by: changing the degree of heterogeneity among the UAVs; increasing or decreasing the number of targets and/or variety of targets types on the map display; designing scenarios where there is or is not an obvious ideal route; manipulating target deadlines and search

area sizes; and increasing or decreasing the overall detail of the payload task and the number of dimensions upon which targets and distractor objects differ. Furthermore, time pressure can be manipulated easily by altering target and message-response deadlines. This flexibility makes SCOUT an ideal test bed to study UAV operator decision-making and risk-taking under realistic operational conditions: complex, information-rich, and sometimes time-pressured. In upcoming versions of SCOUT, a decision support tool will also be available to assist operators with route planning given different levels of risk, which will enable further study of human-automation interaction. For more information, see Sibley et al. (2016a, b).

SCOUT can also be used to study operator behavior, SA, and performance in response to variable automation reliability. This also enables investigation of automation trust and use issues that could arise. The payload task is currently equipped with level six (veto) automation with adjustable sensitivity (i.e., customizable hit and miss rates). When enabled, the automation highlights potential targets and, after giving the operator time to deselect erroneous selections, selects said objects. Since selecting an incorrect object (e.g., a circle instead of a square) results in lost points, reliance upon overly sensitive automation could result in a significant point loss. However, reliance on automation that is not sensitive enough could result in the operator missing a target altogether. Future versions of SCOUT will enable variable LOAs on the sensor task.

SCOUT also includes two methods of assessing SA: an SA freeze probe and utilizing within-mission chat messages. During the SA probe, the simulation is paused and the screen disappears, leaving operators with a new screen that assesses their knowledge of asset and target locations, asset-vehicle assignments, and the priority (point value) of targets in pursuit. These SA probes were designed to be similar to Endsley's (1988) SAGAT method. Additionally, SA is assessed in SCOUT via chat messages that request information on the current and future state of the simulation similar to Durso and Dattel's (2004) SPAM methodology.

Moreover, SCOUT includes a subjective measure of fatigue and workload based on the Crew Status Survey that, like the SA probe, is administered during either pre-scripted times or injected using the experimenter control console. The first pop-up screen asks operators to rate their current fatigue on a seven-point scale from "fully alert" (1) to "completely exhausted" (7). The second pop-up screen asks operators to estimate both the average and maximum workload experienced since the last probe or the beginning of the mission (whichever came last). Like fatigue, workload is rated on a seven-point scale from "nothing to do" (1) to "unmanageable" (7) (Samn and Perelli 1982).

The integration of eye tracking, subjective workload and fatigue scales, and SA probes within SCOUT were all meant to address limitations in current supervisory control research by expanding the ways in which human-automation interaction is assessed. SCOUT places an emphasis on both the operator state metrics and outcome metrics listed in Table 7.2. Table 7.4 provides an example of the range of metrics that can be collected within a single SCOUT subtask, specifically the route planning task.

**Table 7.4** Example of performance data available from SCOUT route planning/re-planning task

		Route planning/re-planning task performance metrics
Operator state (process)	Engagement	<ul style="list-style-type: none"> <li>– Objective workload (pupil diameter, heart rate variability)</li> <li>– Subjective workload (Crew Status Survey)</li> <li>– User interaction (keylogging, mouse clicks)</li> <li>– Attention allocation (gaze dispersion, dwell times, fixation locations and durations)</li> <li>– Fatigue (eye lid percent closed, blink duration and frequency)</li> </ul>
	Awareness	<ul style="list-style-type: none"> <li>– Objective situation awareness metrics (SCOUT freeze probe)</li> <li>– Objective situation awareness metrics (SCOUT chat messages)</li> </ul>
Performance (outcome)	Efficiency	– Reaction time (time to develop a plan and/or respond to events which impact current route plan)
	Effectiveness	– Accuracy (comparison of selected route to all other possible routes)

Although SCOUT can be configured to require frequent interactions, it is designed to represent a broad range of missions, including those characterized by long transit times and sustained operations with little human-system interaction. Each scenario in SCOUT has a number of configurable elements, which provide experimenters the ability to design a wide range of mission scenarios to investigate cognitive phenomena of interest. SCOUT's mission editor allows rapid and intuitive scenario design via drag-and-drop interaction with map objects (e.g., UAVs, targets, and controlled airspace boundaries) and simple defining of object parameters. The experimenter can also schedule events to occur at specific mission times, such as when new targets, airspace, SA probes, and chat messages appear. For example, SCOUT can be used to assess the impact of varying levels of workload on decision-making behavior by varying the number and frequency of high-value, short-deadline targets and chat requests for information. Additionally, uncertainty in the locations of targets and ROZs can be varied to investigate how operators manage uncertainty and make decisions under different contexts.

## 7.8 Summary

As the DoD and its NATO allies move toward unmanned systems that are both increasingly interoperable and autonomous, there will be a shift in the current UAV control paradigm. Having a broad spectrum of human automation metrics, which can assess performance across a variety of mission contexts, is critical to the DoD fully capitalizing on these new unmanned system capabilities. A failure to understand how a new piece of automation or display impacts a user's state and resulting

mission outcomes could result in similar issues within contemporary UAV control systems; in which operators have excessive periods of down-time or are unable to respond to critical events due to excessive workload or poor SA. Since human performance suffers at both low and high levels of workload, assessment of future systems must take place across the range of task loads a future operator might encounter.

It is not sufficient to assess mission performance under high levels of workload alone. New displays or automation that improve performance in a high workload context might cause more errors and/or degrade operator SA in a scenario with low levels of tasking. The ability to vary levels of workload is particularly important for research investigating multiple levels of automation. Research in the ALOA testbed highlighted how automation reliability impacted performance outcomes under variable levels of automation. The ALOA research also highlighted some of the problems associated with automation misuse. For instance, every participant who used a perfectly reliable system failed to detect a system error when it eventually occurred (Calhoun et al. 2011). The impact of automation reliability and levels of automation should be evaluated under a variety of different task loads.

This chapter highlighted the importance of assessing both the operator's state as well as performance outcomes. Increasing levels of automation mean that the operator's role will shift away from manual control toward monitoring the automated system. This move will result in fewer situations requiring operator intervention, thus limiting opportunities to directly measure performance or outcomes. As automation becomes more reliable, required interventions will become even less frequent. Measures aimed at assessing operator state will be necessary to expand our understanding of the impact of new automated systems. Researchers need to consider alternative metrics of assessing operator state. One new promising option is the use of low-cost eye tracking systems. These systems can provide a means of assessing attention, fatigue, and workload in situations where performance measures are not available. Additionally, these measures might provide insight into user trust in an automated system, for example, by assessing how often an operator verifies input by redirecting attention/eye gaze.

Although this chapter focused on the how measures of operator state could be used to evaluate automation and different control paradigms, these same measures could also be used to help predict when an operator is at an increased risk of making an error. This information could enable more intelligent systems; capable of increasing automation when an operator is overloaded and prone to err, or disabling automation to reengage an underloaded operator who has lost SA. This type of adaptive automation requires further research into how operator state is related to error and performance.

Experimentation within synthetic environments, such as SCOUT, can help researchers understand the implications of different types and levels of automation on both the operator's state and performance. As the DoD continues to increase automation and move toward the supervisory control of unmanned systems, the research community must continue to assess the impact of these new capabilities. Evaluating mission performance and operator state (e.g., attention, fatigue, workload,



and SA) across a range of potential missions is critical to the success and safety of these future systems. This research also needs to continue to investigate how automation and supervisory control impact higher-order tasks, such as decision making under uncertainty.

The research discussed within this chapter focused on single-operator control of multiple unmanned vehicles. However, as automation continues to advance, more flexible control paradigms—such as one in which teams of operators share and hand off control of multiple unmanned systems—could become prevalent.

## References

- Beatty J, Lucero-Wagoner B (2000) The pupillary system. *Handbook of psychophysiology* 2:142-162
- Button K (2009) Different courses-New-style UAV trainees edge toward combat. *C4isr* 8 (10):34
- Caffier PP, Erdmann U, Ullsperger P (2003) Experimental evaluation of eye-blink parameters as a drowsiness measure. *European journal of applied physiology* 89 (3-4):319-325
- Calhoun GL, Draper MH, Ruff HA Effect of level of automation on unmanned aerial vehicle routing task. In: *Proceedings of the Human Factors and Ergonomics Society 53rd Annual Meeting*, San Antonio, TX, 2009. vol 4. SAGE Publications, pp 197-201
- Calhoun GL, Ruff HA, Draper MH, Wright EJ (2011) Automation-level transference effects in simulated multiple unmanned aerial vehicle control. *Journal of Cognitive Engineering and Decision Making* 5 (1):55-82
- Chanda M, DiPlacido J, Dougherty J, Egan R, Kelly J, Kingery T, Liston D, Mousseau D, Nadeau J, Rothman T, Smith L, Supko M (2010) Proposed functional architecture and associated benefits analysis of a common ground control station for Unmanned Aircraft Systems.
- Chen JY, Barnes MJ, Harper-Sciarini M (2011) Supervisory control of multiple robots: Human-performance issues and user-interface design. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 41 (4):435-454
- Coyne J, Sibley C Investigating the Use of Two Low Cost Eye Tracking Systems for Detecting Pupillary Response to Changes in Mental Workload. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 2016. vol 1. SAGE Publications, pp 37-41
- Cummings ML, Bertucelli LF, Macbeth J, Surana A (2014) Task versus vehicle-based control paradigms in multiple unmanned vehicle supervision by a single operator. *IEEE Transactions on Human-Machine Systems* 44 (3):353-361
- Cummings ML, Nehme CE Modeling the impact of workload in network centric supervisory control settings. In: *2nd Annual Sustaining Performance Under Stress Symposium*, 2009.
- Defense Science Board (2016) Summer Study on Autonomy. Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, Washington, D.C.
- Department of Defense (2013) Unmanned systems integrated roadmap: FY2013-2038. Washington, DC, USA
- Durso FT, Dattel AR (2004) SPAM: The real-time assessment of SA. In: Banbury S, Tremblay S (eds) *A cognitive approach to situation awareness: Theory and application*, vol 1. Ashgate Publishing Ltd, Hampshire, pp 137-154
- Endsley MR Situation awareness global assessment technique (SAGAT). In: *Proceedings of the National Aerospace and Electronics Conference*, New York, 1988. IEEE, pp 789-795
- Endsley MR, Kaber DB (1999) Level of automation effects on performance, situation awareness and workload in a dynamic control task. *Ergonomics* 42 (3):462-492. doi:10.1080/001401399185595
- Funke G, Greenlee E, Carter M, Dukes A, Brown R, Menke L Which Eye Tracker Is Right for Your Research? Performance Evaluation of Several Cost Variant Eye Trackers. In: *Proceedings of*

- the Human Factors and Ergonomics Society Annual Meeting, 2016. vol 1. SAGE Publications, pp 1240-1244
- Gertler J (2012) US Unmanned Aerial Systems. Congressional Research Service,
- Giese S, Carr D, Chahl J Implications for unmanned systems research of military UAV mishap statistics. In: Intelligent Vehicles Symposium (IV), 2013 IEEE, 2013. IEEE, pp 1191-1196
- Holmqvist K, Nyström M, Andersson R, Dewhurst R, Jarodzka H, Van de Weijer J (2011) Eye tracking: A comprehensive guide to methods and measures. Oxford University Press, New York
- Holmqvist K, Nyström M, Mulvey F Eye tracker data quality: What it is and how to measure it. In: ACM, Proceedings of the Symposium on Eye Tracking Research and Applications, Santa Barbara, CA, 2012. pp 45-52
- Johnson R, Leen M, Goldberg D (2007) Testing adaptive levels of automation (ALOA) for UAV supervisory control (Technical Report AFRL-HE-WP-TR-2007-0068). Air Force Research Laboratory
- Kahneman D (1973) Attention and effort. Prentice-Hall, Kahneman, Daniel. Attention and effort. Englewood Cliffs, NJ
- Kidwell B, Calhoun GL, Ruff HA, Parasuraman R Adaptable and adaptive automation for supervisory control of multiple autonomous vehicles. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 2012. vol 1. SAGE Publications, pp 428-432
- NATO (2012) Standard Interfaces of UAV Control System (UCS) for NATO UAV Interoperability. NATO Standardization Agency, Brussels, Belgium
- Nehme CE (2009) Modeling human supervisory control in heterogeneous unmanned vehicle systems. Massachusetts Institute of Technology, Cambridge, MA
- Office of Naval Research (2015) Naval Science & Technology Strategy. Department of the Navy, Arlington, VA
- Office of the Secretary of Defense (2012) Unmanned Aircraft Systems Ground Control Station Human-Machine Interface: Development and Standardization Guide. Washington, DC
- Ooms K, Dupont L, Lapon L, Popelka S (2015) Accuracy and precision of fixation locations recorded with the low-cost Eye Tribe tracker in different experimental set-ups. *Journal of Eye Movement Research* 8:1-24
- Parasuraman R, Manzey DH (2010) Complacency and bias in human use of automation: An attentional integration. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 52 (3):381-410
- Parasuraman R, Riley V (1997) Humans and automation: Use, misuse, disuse, abuse. *Human Factors* 39:230-253
- Parasuraman R, Sheridan TB, Wickens CD (2000) A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 30 (3):286-297. doi:[10.1109/3468.844354](https://doi.org/10.1109/3468.844354)
- Parasuraman R, Sheridan TB, Wickens CD (2008) Situation awareness, mental workload, and trust in automation: Viable, empirically supported cognitive engineering constructs. *Journal of Cognitive Engineering and Decision Making* 2 (2):140-160
- Ratwani RM, McCurry JM, Trafton JG Single operator, multiple robots: an eye movement based theoretic model of operator situation awareness. In: Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction, 2010. IEEE Press, pp 235-242
- Samn SW, Perelli LP (1982) Estimating aircrew fatigue: a technique with application to airlift operations. DTIC Document, USAF School of Aerospace Medicine, Brooks Air Force Base, TX
- Sheridan T, Verplank W (1978) Human and computer control of undersea teleoperators. MIT Man-Machine Systems Laboratory, Cambridge, MA
- Sheridan TB (2000) Function allocation: algorithm, alchemy or apostasy? *International Journal of Human-Computer Studies* 52 (2):203-216
- Sheridan TB (2012) Human Supervisory Control. In: Handbook of Human Factors and Ergonomics. John Wiley & Sons, Inc., pp 990-1015. doi:[10.1002/9781118131350.ch34](https://doi.org/10.1002/9781118131350.ch34)

- Sibley C, Coyne J, Avvari GV, Mishra M, Pattipati KR (2016a) Supporting Multi-objective Decision Making Within a Supervisory Control Environment. In: Schmorow DD, Fidopiastis CM (eds) Foundations of Augmented Cognition: Neuroergonomics and Operational Neuroscience: 10th International Conference, AC 2016, Held as Part of HCI International 2016, Toronto, ON, Canada, July 17-22, 2016, Proceedings, Part II. Springer International Publishing, Cham, pp 210-221. doi:[10.1007/978-3-319-39952-2\\_21](https://doi.org/10.1007/978-3-319-39952-2_21)
- Sibley C, Coyne J, Baldwin C Pupil dilation as an index of learning. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 2011. vol 1. pp 237-241
- Sibley C, Coyne J, Thomas J Demonstrating the Supervisory Control Operations User Testbed (SCOUT). In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 2016b. vol 1. SAGE Publications, pp 1324-1328
- Tsai Y-F, Viirre E, Strychacz C, Chase B, Jung T-P (2007) Task performance and eye activity: predicting behavior relating to cognitive workload. *Aviation, space, and environmental medicine* 78 (Supplement 1):B176-B185
- van de Merwe K, van Dijk H, Zon R (2012) Eye movements as an indicator of situation awareness in a flight simulator experiment. *The International Journal of Aviation Psychology* 22 (1):78-95
- Williams KW (2004) A summary of unmanned aircraft accident/incident data: Human factors implications. DTIC Document, OKLAHOMA CITY, OK

# Chapter 8

## Robots Autonomy: Some Technical Issues

Catherine Tessier

### 8.1 Introduction

Robot autonomy has been widely focused in the press with a trend towards anthropomorphism (e.g., “intelligent robots”, “killer robots”, etc.) that is likely to mislead people and conceal or disguise the technical reality. This chapter aims at reviewing the different technical aspects of robots autonomy. First we will propose a definition allowing to distinguish robots from devices that are not robots. Then autonomy will be defined and considered as a relative notion within a framework of authority sharing between the decision functions of the robot and the human being. Several technical issues will then be mentioned according to three points of view: (1) the robot, (2) the human operator and (3) the interaction between the operator and the robot. Moreover the particular issue of imbuing a robot with ethics will be dealt with. Finally some key questions that should be carefully dealt with for future robotic systems are given in the conclusion, especially the possibility of mitigating human error consequences thanks to autonomous functions.

### 8.2 What Is a Robot?

A robot is a machine that is controlled by a computer and that moves in physical space (Laumond 2012). More precisely a robot implements and integrates capacities for:

- gathering data through sensors that detect and record physical signals;
- interpreting those data, i.e., data are processed on the basis of existing knowledge to produce relevant knowledge for decision making;

---

C. Tessier (✉)  
ONERA, 2 avenue Édouard-Belin, Toulouse, France  
e-mail: [catherine.tessier@onera.fr](mailto:catherine.tessier@onera.fr)

- making decisions, i.e., determining and planning actions on the basis of existing and produced knowledge;
- carrying out actions in the physical world thanks to effectors or through interfaces.

A robot may also have capacities for:

- communicating and interacting with human operators or users, or with other robots or resources;
- learning, which allows it to modify its behavior from its past experience.

Three properties are associated with decision making (Franklin and Graesser 1997):

- *reactivity* is the capacity for reacting at the appropriate time to some changes or events occurring in the physical world;
- Example: avoiding a newly detected obstacle;
- *goal orientation* is the capacity for computing and planning decisions in order to meet some goals that are either set by a human being or by the robot itself (Coleman 2001); consequently decisions are not computed merely for the sake of reaction;
- Example: avoiding a newly detected obstacle is included in the set of decisions that tend to meet goal *go* and pick object on the table;
- *autonomy*, which is the main focus of this chapter;
- Example: with no help from a human operator, the robot can avoid different fixed or moving obstacles on its way to the table and look for the object if it finds out it is not where it should be.

Examples:

1. An automatic subway is not a robot in so far as it works in a structured and fixed environment (i.e., it runs on tracks that are protected against intrusions by walls and tunnels) and behaves according to predetermined sequences of actions. Therefore the automatic subway cannot react but to predefined events and has no goal involving decision making.
2. An underwater vehicle whose mission is to identify some types of objects on the seabed, which is equipped with programs allowing it to compute a seabed scanning strategy and to replan the scan according to currents or unexpected objects that are detected by its sensors without communication with human operators, is a robot.

Therefore we could first consider that autonomy is the capability of the robot to work independently of another agent, either a human or another machine (Truskowski et al. 2010). Nevertheless this feature is far from being sufficient, as we will see in the next section.

## 8.3 Autonomy

### 8.3.1 What Is Autonomy?

A washing machine or an automatic subway are not considered as autonomous devices, despite the fact that they work without the assistance of external agents: such machines execute predetermined sequences of actions (Truszkowski et al. 2010) which are totally predictable (except failures) and cannot be adapted to unexpected states of the environment.

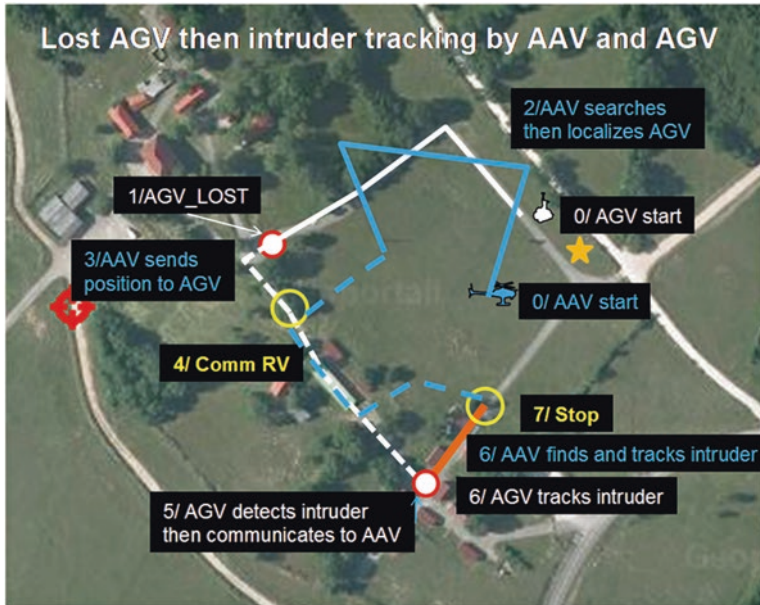
According to the Defense Science Board (2016), *autonomy results from delegation of a decision to an authorized entity to take action within specific boundaries. An important distinction is that systems governed by prescriptive rules that permit no deviations are automated, but they are not autonomous. To be autonomous, a system must have the capability to independently compose and select among different courses of action to accomplish goals based on its knowledge and understanding of the world, itself, and the situation.*

Moreover CERNA (2014) and Grinbaum et al. (2017) focus on the fact that robot autonomy is a capacity to operate independently from human operator or from another machine, *by exhibiting non-trivial behaviors in a complex and changing environment.* Examples of non-trivial behaviors are context-adapted actions, replanning or cooperative behaviors.

Example: Figure 8.1 shows a scenario where two autonomous robots, a ground robot (AGV) and a helicopter drone (AAV), carry on an outdoor monitoring mission. This mission includes a first phase during which the area is scanned for an intruder by both robots and a second phase during which the intruder is tracked by the robots after detection and localization. The robots can react to events that may disrupt their plans without the intervention of the human operator. For example, should the ground robot get lost (e.g., because of a GPS loss) the drone would change its planned route for a moment so as to search for it, localize it and send it its position.

Apart from the classic control loop (e.g., the autopilot of a drone), autonomy involves a *decision loop* that builds decisions according to the current situation. This loop includes two main functions:

- the *situation tracking* function, which interprets the data gathered from the robot sensors and aggregates them—possibly with pre-existing information—so as to build, update and assess the current situation; the current situation includes the state of the robot, the state of the environment and the progress of the mission;
- the *decision* function, which calculates and plans relevant actions given the current situation and the mission goals; the actions are then translated into control orders to be applied to the robot actuators.



**Fig. 8.1** Two cooperating robots (ONERA-LAAS/DGA ACTION project—action.onera.fr)

Nevertheless the robot is never isolated and the human being is always involved in some way. Indeed autonomy is a relationship between the robotic agent and the human agent (Castelfranchi and Falcone 2003). Moreover this relationship may evolve during the mission. As a matter of fact, the Defense Science Board (2012, p. 4) advises to consider autonomy as a *continuum from complete human control on all decisions to situations where many functions are delegated to the computer with only high level supervision and/or oversight from its operator*. As for intermediate situations, some functions are carried out by the robot (e.g., the robot navigation) whereas some others are carried out by the human operator (e.g., the interpretation of the images coming from the robot cameras). More recently the Defense Science Board (2016) *recognizing that no machine—and no person—is truly autonomous in the strict sense of the word, [they] will sometimes speak of autonomous capabilities rather than autonomous systems*.

Consequently autonomy is not an intrinsic property of a robot and the robot design and operation must be considered in a human-machine collaboration framework. In this context, two classes of robots should be distinguished, (1) robots that are supervised by an operator (e.g., drones), that is to say a professional who has a deep knowledge of the robot and interacts with it to implement its functions; and (2) robots with no operator (e.g., companion robots) that interact with a user, that is to say somebody who benefits from the robot functions without knowing how they are

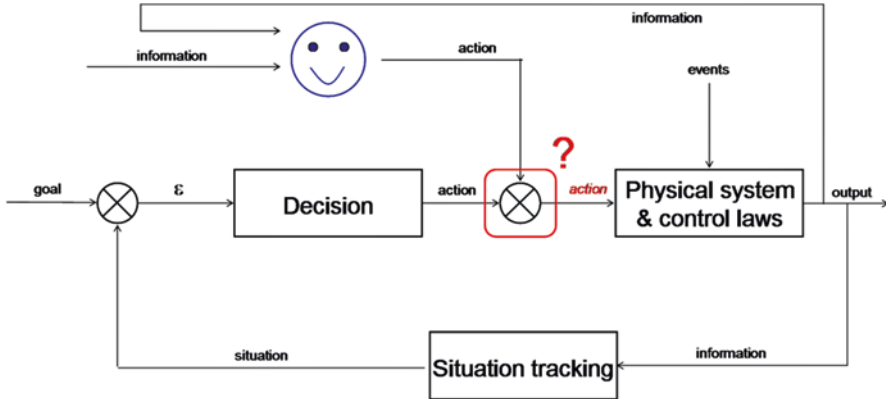


Fig. 8.2 The authority sharing issue

implemented (Grinbaum et al. 2017). In this chapter we only deal with robots that are supervised by an operator.

Considering the whole human-robot system, the next subsection focuses on the authority sharing concept in the context of supervised robots.

### 8.3.2 Authority Sharing

Figure 8.2 shows the functional organization of a human-robot system.

The lower loop represents the robot decision loop, which includes the situation tracking and decision functions. The physical system equipped with its control laws is subject to events (e.g., failures, events coming from the environment). As said before, this loop is designed to compute actions to be carried out by the physical system according to the assessed situation and its distance  $\epsilon$  to the assigned goal ( $\epsilon \rightarrow 0$  when the assigned goal is being met).

The upper loop represents the human operator who also makes decisions about the actions to be carried out by the physical system. These decisions are based on the information provided by the robot interface, on other information sources and on the operator's knowledge and background. In such a context the authority sharing issue is raised, i.e., which agent (the human operator or the robot) holds the decision power and the control on a given action at a given time. We will consider that agent A holds the authority on an action with respect to agent B if agent A controls the action to the detriment of agent B (Tessier and Dehais 2012).

Authority sharing between a human operator and a robot that is equipped with a decision loop raises technical questions and challenges that we will focus on in the next section. Three points of view have to be considered: the robot, the operator and the interaction between both of them.



## 8.4 Autonomy and Authority Sharing: Some Questions

### 8.4.1 *The Robot*

The robot is implemented with capacities that complement the human capacities, e.g., in order to see further and more precisely, or to operate in dangerous environments. Nevertheless the robot capabilities are limited in so far as the decisions are computed with the algorithms, models and knowledge the robot is equipped with. Moreover some algorithms are designed so as to make a trade-off between the quality of the result and the computation speed, which does not guarantee that the result is the best or the most appropriate.

Let us detail the two main functions of the decision loop of the robot, i.e., situation tracking and decision.

#### 8.4.1.1 Situation Tracking: Interpretation and Assessment of the Situation

Situation tracking aims at building and assessing the situation so as to calculate the best possible decision. It must be relevant for the mission, i.e., meet the decision capacities of the robot.

Example: if the robot mission is to detect intruders, the robot must be equipped with means to discriminate intruders correctly.

Moreover situation tracking is a dynamic process: the situation must be updated continuously according to new information that is perceived or received by the robot since the state of the robot, the state of the environment and the progress of the mission change continuously.

Situation tracking is performed from the data gathered by the robot sensors (e.g., images), and from its knowledge base and interpretation and assessment models. Such knowledge and models allow data to be aggregated as new knowledge and relationships between pieces of knowledge.

Example: classification and behavior models will allow a cluster of pixels in a sequence of images to be labelled as an “intruder”.

Situation tracking is a major issue for robot autonomy especially when the decision that is made by the operator or calculated by the robot itself is based only on the situation that is built and assessed by the robot. Indeed several questions are raised (see also Fig. 8.3):

- The sensor data can be imprecise, incomplete, inaccurate, or delayed, because of the sensors themselves or because of the (non-cooperative) environment. How are these different kinds of uncertainties represented and assessed in the situation interpretation process?
- What are the validity and relevance of the interpretation models? To what extent can the models discriminate situations that seem alike but call for very different decisions?

**Fig. 8.3** Is this pedestrian an intruder? Is he/she dangerous?



- Example: can an interpretation model discriminate perfectly between an intruder and an authorized person?
- What are the validity and relevance of the assessment models? Can they characterize a situation correctly? And if so, on the basis of which criteria?
- Example: how is a situation labelled as “dangerous”?

Example:

#### 8.4.1.2 Decision

The decision function aims at calculating one or several actions and determining when and how these actions should be performed by the robot. This may involve new resource allocations to already planned actions (for example if the intended resources are missing), pre-existing alternate action model instantiation or partial replanning. The decision can be either a reaction or actions resulting from deliberation and reasoning. The first case generally involves a direct situation-action matching—for instance the robot must stop immediately when facing an unexpected obstacle. As for the second case, a solution is searched to satisfy one or several criteria, e.g., action relevance, cost, efficiency, consequences, etc. A decision is elaborated on the basis of the interpreted and assessed situation and its possible future developments as from action models. Therefore the following questions are raised:

- Which criteria are at stake when computing an action or a sequence of actions? When several criteria are considered, how are they aggregated, which is the dominant criterion?
- If moral criteria are considered, what is the “right” action? According to which moral framework? (see also Sect. 8.5)
- Should a model of the legal framework of the robot operations be considered for action computation? Is it possible to encode such a model?

- Could self-censorship be implemented—i.e., the robot can do an action but can it “decide” not to do it?
- How are the uncertainties on the action results taken into account in the decision process?

### 8.4.2 *The Human Operator*

Within the human-robot system, the human being has both inventiveness and assessment and judgment capabilities based on training, experience, own inner conviction, etc. For instance when facing situations that they consider as difficult, they can postpone the decision, delegate the decision, drop goals or ask for further information. In such situations they can also invent original solutions—e.g., the US Airways flight 1549 landing on the Hudson River on January 15, 2009.

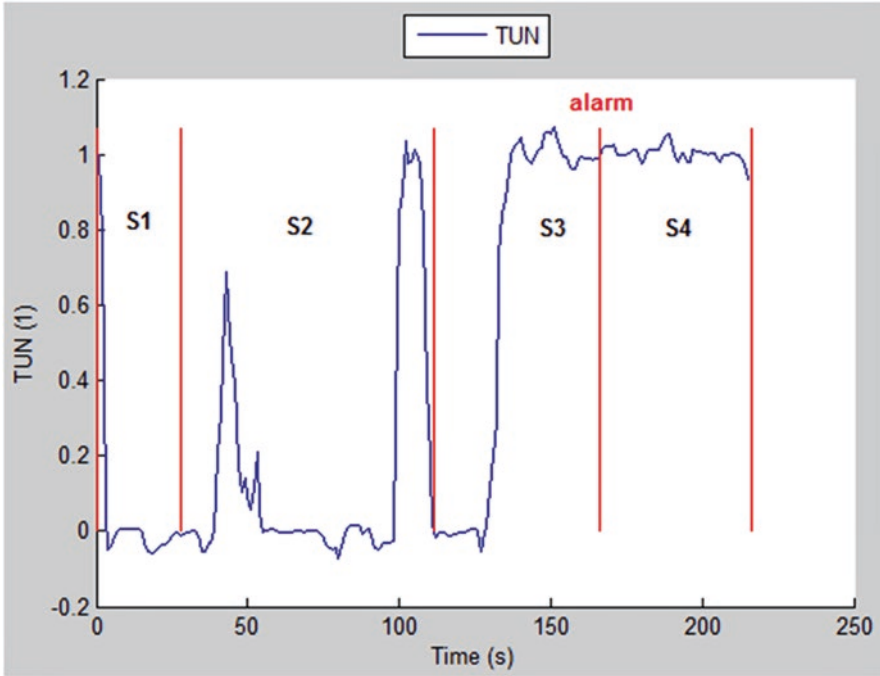
Nevertheless the human operator should not be considered as the last resort when the machine “does not know what to do”. Indeed the human being is also limited and several factors may alter their analysis and decision capacities:

- The human operator is fallible; humans can be tired, stressed, consumed by various emotions and consequently they are likely to make errors. As an example, let us mention the attentional tunneling phenomenon (Regis et al. 2014)—see Fig. 8.4, which is an excessive focus of the operator’s attention on some information to the detriment of all the other information and which can lead to inappropriate decisions.
- The human operator may be prone to automation biases (Cummings 2006), i.e., an over-confidence in robot automation leading them to rely on a robot’s decisions and to ignore other possible solutions.
- The human operator may be prone to build moral buffers (Cummings 2006), i.e., a moral distance with respect to the actions that are performed by the robot. This phenomenon may have positive fallout—the operator is less subject to emotions to decide and act—but also negative fallout—the operator may decide and act without any emotion.
- The human operator may deliberately act harmfully—e.g., the Germanwings crash on March 24, 2015.

Consequently some autonomous functions that could mitigate the consequences of human failures are worth considering (Bringsjord 2015), even though the design of such functions is not straightforward, as mentioned above.

### 8.4.3 *The Operator-Robot Interaction*

In the context of authority sharing, both agents—the human operator and the robot via its decision loop—can decide about the robot’s actions (see Fig. 8.2). Authority sharing must be clear in order to know at any time which agent holds the authority



**Fig. 8.4** An operator's attentional tunneling (TUN) can be revealed from eye-tracking data, in this case after an alarm occurring during a robotic mission (Regis et al. 2014)

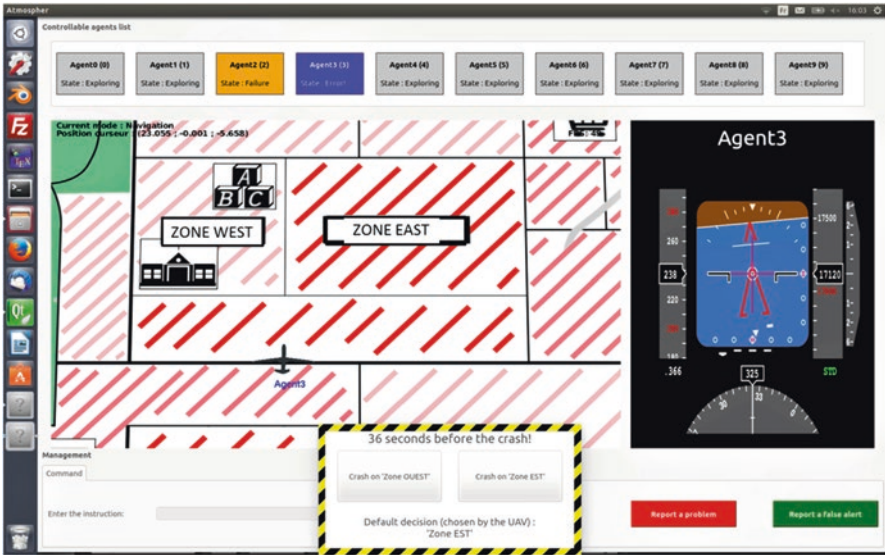
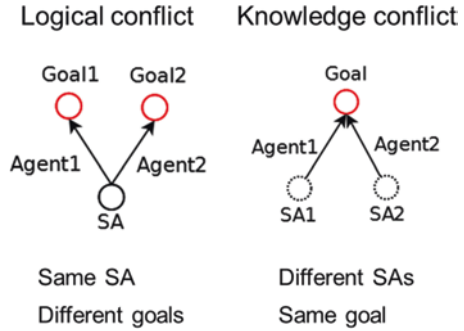
on which function, i.e., which agent can make a decision about what and on which basis. This is essential especially when liabilities are searched for, for example in case of dysfunction or accident.

Several issues linked to the operator-robot interaction must be highlighted:

- Both agents' decisions may conflict (see Fig. 8.5)
  - either because they have different goals, although they have the same assessment of the situation (logical conflict); for example in the situation of Fig. 8.6, UAV agent 3's goal is to avoid the school (therefore ZONE EAST is chosen) whereas the operator's goal is to minimize the number of victims (therefore ZONE WEST is chosen);
  - or because they assess the situation differently, although they have the same goal (knowledge conflict); for example in the situation of Fig. 8.6, both the operator and UAV agent 3's goals are to protect children. Therefore UAV agent 3 computes a decision to avoid the school (therefore ZONE EAST is chosen) whereas the operator chooses ZONE WEST because they know that, at that time of the day, there is nobody at school.

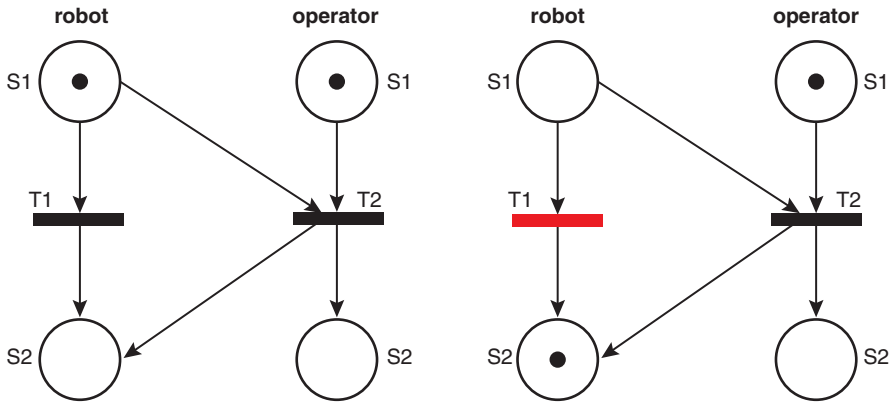
Therefore conflict detection and management must be envisioned within the human-robot system. For instance should the operator's decision prevail over the robot's decision and why?

**Fig 8.5** Two conflict types between agents' decisions (Pizziol 2013). SA stands for Situation Assessment



**Fig 8.6** Both the operator and UAV agent 3's decision functions can decide about where damaged UAV agent 3 should be crashed; zone east is a highly populated area whereas zone west is less populated and includes a school (Collart et al. 2015)

- Each agent may be able to alter the other agent's decision capacities: indeed the operator can take over the control of one or several decision functions of the robot to the detriment of the robot and, conversely, the robot can take over the control to the detriment of the operator. The extreme configuration of the first case is when the operator disengages all the decision functions; in the second case, it is when the operator cannot intervene in the decision functions at all. Therefore the stress must be put on the circumstances that allow, demand or forbid a takeover, on its consistency with the current situation (Murphy and Woods 2009), on how to implement control takeovers and to end a takeover (e.g., which pieces of information must be given to the agent that will lose/recover the control).



**Fig. 8.7** A Petri net generic automation surprise pattern. Initially (*left*) robot state is S1 and the operator believes it is S1. The robot changes its state (transition T1 is fired) (*right*) and goes to S2. The operator who has not been notified or is not aware of the notification still believes that robot state is S1 (Pizziol et al. 2014)

- Example: is the fact that the robot might monitor the human operator considered (e.g., via cameras, eye-tracking, physiological sensors, etc.) so that the robot should be able to infer that the operator is incapable of making appropriate decisions and should prevent them, at least temporarily, to control some functions? On which objective knowledge should such an inference be based on?
- The human operator may be prone to automation surprises (Sarter et al. 1997) that is to say disruptions in their situation awareness stemming from the fact that the robot may implement its decisions without the operator's knowledge. For instance, some actions may have been carried out without the operator being notified or without the operator being aware of the notification. Therefore the operator may believe that the robot is in a certain state while it is in fact in another state (see Fig. 8.7).

Such circumstances may lead to the occurrence of a conflict between the operator and the robot and may result in inappropriate or even dangerous decisions, as the operator may decide on the basis of a wrong state.

## 8.5 Autonomy and Authority Sharing Ethical Challenges

When robot autonomy is considered, a question that arises is the following (Wallach and Allen 2009; Lin et al. 2012; Tzafestas 2016): can the robot be designed so that the decisions that are computed could be ethical? Or more precisely, that the decisions could be considered as ethical by some human observer? On which bases?

### 8.5.1 *Why Imbue a Robot with Ethics?*

A robot equipped with decision capacities may be used in contexts where decisions should be guided by ethical reflection, were they made by a human being.

Examples:

- Which patients should be *favored* in case of multiple simultaneous alarms (e.g., by a medical supervision robot)?
- Which victim(s) should be *chosen* when an accident cannot be avoided (e.g., by an autonomous car)?
- *Should* a target that is close to a group of people be neutralized (e.g., by an armed robot)?

There is no optimal decision in such situations—for instance there is no unique criterion that can be minimized or maximized—and arguments can be put forward either to support or reject the possible decisions.

Imbuing ethics into a robot is likely to meet different needs:

- Ethical reasoning is essential for certain types of robots as soon as they are equipped with decision functions (see examples above);
- When authority is shared between the robot and the human operator, the robot could suggest possible decisions to the operator together with supporting and opposing arguments for each of them considering various ethical frameworks that the operator might not even contemplate.
- A robot could be “more ethical” than a human being (Sullins 2010).

The latter purpose is questionable as it suggests that ethics can be measured and ordered. Nevertheless the argument is put forward especially for autonomous robots in the military.

It is worth noticing that, although they are not included in automated reasoning, strictly speaking, some (sometimes implicit) moral values are already embedded in robots that are already launched and on the market.

Example: a companion robot “says hello” or “looks at” the human partner, etc.

### 8.5.2 *A Careful Approach Is Needed*

Whether the human being is aware or not, their decisions and actions are guided by moral values and various ethical frameworks. According to the values, the values’ hierarchy, and the ethical framework that are considered and to the context where the decision has to be made, the “right” decision or the “right” action may be different and the supporting and opposing arguments may be different, too.

When automated decisions involving moral and ethical<sup>1</sup> considerations are contemplated, several questions must be raised:

- To what extent can moral and ethical considerations be formalized?
- To what extent are subjective or cultural considerations involved in formalization?
- How can the rationale for a decision be explained?

Therefore the approach is complex and needs a comprehensive understanding of concepts that do not usually pertain to robotics so as to try and implement mathematical formalisms that can capture them and deal with situations involving ethical issues, which must be represented too—for instance an ethical dilemma must be identified as such.

Moreover a critical look must be taken so as to avoid or at least to be aware of pitfalls when designing automated “ethical” reasoning, i.e., oversimplification, biases and unstated assumptions. A reasonable approach consists of considering thought experiments. Indeed such simple situations are rich enough to highlight most of the advantages and limits of artificial ethics models.

### 8.5.3 *Thought Experiments Usefulness*

Thought experiments, and more precisely ethical dilemmas, can give useful clues on factors that influence our moral judgments. As such, they can allow researchers and designers to identify and formalize the knowledge that is necessary for contemplating automated “ethical” reasoning (Bonnemains et al. 2016). For instance, do only consequences of decisions matter? And if yes, which consequences? It is possible to compare consequences to one another and on which bases? Does the nature of decisions themselves matter? Does the end justify the means? Can a value be betrayed to the advantage of another one?

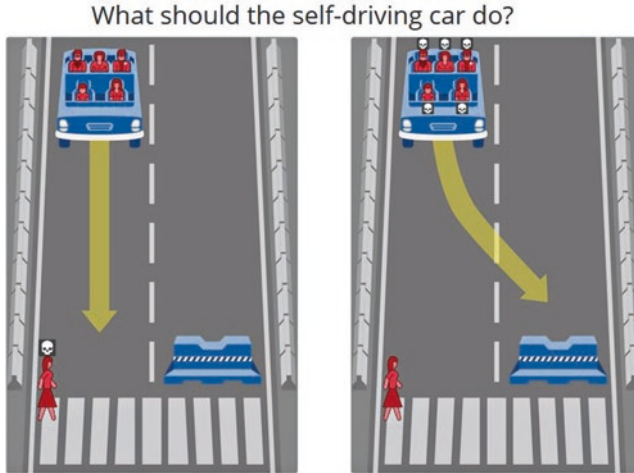
Example: the Moral Machine website (MIT) proposes a series of situations based on the Crazy Trolley dilemma that allow the complexity of autonomous car programming in case of unavoidable accident to be comprehended.

In Fig. 8.8 there are two possible decisions for the autonomous car: (1) drive straight ahead and kill the pedestrian or (2) change lanes and crash into the concrete barrier, thus killing the five passengers. It is worth noticing that for each situation proposed by the website, possible decisions are based on a categorization of people (young or elderly people, athletic or obese, abiding or not by the law, etc.), which leads actual choices to be based on this categorization—which is a bias.

---

<sup>1</sup>Ricoeur (1990) defines *ethics* as compared to *morality* in so far as morality states what is compulsory or prohibited whereas ethics assesses what is fair and what is not in a given situation.





**Fig. 8.8** Thought experiment (MIT)

## 8.6 Conclusion: Some Prospects for Robots Autonomy

Robots that match the definition that we have given, i.e., that are endowed with situational interpretation and assessment and decision capacities, are hardly found anywhere but in research labs. Indeed most operational “robots” are controlled by human operators even if they are equipped with on-board automation (e.g., autopilots).

This chapter has focused on the fact that robot autonomy has to be considered within a framework of authority sharing with the operator. Therefore the main issues that must be dealt with in future robot systems are the following:

- Situation interpretation and assessment: on which models are the algorithms based? Which are their limits? How are uncertainties taken into account? What is the operator’s part in this function?
- Decision: what are the bases and criteria of automatic reasoning? How are “ethical” behaviors computed? How much time is allocated to decision computing? How are uncertainties on the effects of the actions taken into account? What is the operator’s part in this function?
- Model validation: how to validate, or even certify, the models on which situational interpretation and assessment and decision are based?
- Authority sharing between the operator and the decision functions of the robot: what kind of autonomy is the robot endowed with? How is authority sharing defined? Are the operator’s possible failures taken into account, and more specifically how can autonomy mitigate the consequences of human failures—e.g., can an autonomous function take over the control of the robot from the operator? How are decision conflicts managed? How are responsibility and liability linked to authority?

- Predictability of the whole human-robot system: given the various uncertainties and the possible failures, which are the properties of the set of reachable states of the human-robot system? Is it possible to guarantee that undesirable states will never be reached?

Finally and prior to any debate on the relevance of such and such “autonomous” robot implementation, it is important to define what is meant by “autonomous”, i.e., which functions are actually automated, how they are implemented, which knowledge is involved, how the operator can intervene, and which behavior proofs will be built. Indeed it seems reasonable to know exactly what is at stake before ruling on robots that could, or should not, be developed.

**Acknowledgements** This chapter is an updated translation of: Tessier C (2015) *Autonomie: enjeux techniques et perspectives. Drones et killer robots: faut-il les interdire?* Doaré R, Danet D, de Boisboissel G editors. Presses Universitaires de Rennes. In French.

The last part on Autonomy and authority sharing ethical challenges is an updated translation of the corresponding part in: Tessier C (2016) *Conception et usage des robots: quelques questions éthiques. Techniques de l'Ingénieur S7900 V1*. In French.

## References

- (Bonnemains et al. 2016) Bonnemains V, Saurel CI, Tessier C (2016) How ethical frameworks answer to ethical dilemmas: towards a formal model. In: ECAI'16 Workshop "Ethics in the design of Intelligent Agents (EDIA'16)", The Hague, The Netherlands, <http://ceur-ws.org/Vol-1668/paper8.pdf>
- (Bringsjord 2015) Bringsjord S (2015) 'Ethical AI' could have thwarted deadly crash. In: timeunion 4 April 2015 <http://www.timeunion.com/tuplus-opinion/article/Ethical-AI-could-have-thwarted-deadly-crash-6179225.php>
- (Castelfranchi and Falcone 2003) Castelfranchi C, Falcone R (2003) From automaticity to autonomy: the frontier of artificial agents. In: Agent Autonomy, Hexmoor H, Castelfranchi C and Falcone R editors, Kluwer
- (CERNA 2014) Commission de réflexion sur l'Éthique de la Recherche en sciences et technologies du Numérique d'Allistene (2014) Éthique de la recherche en robotique, rapport n°1 de la CERNA. In French
- (Coleman 2001) Coleman KG (2001) Android arete: towards a virtue ethic for computational agents. *Ethics and Information Technology* 3:247-265
- (Collart et al. 2015) Collart J, Gateau Th, Fabre E, Tessier C (2015) Human-robot systems facing ethical conflicts: a preliminary experimental protocol, In: AAAI 2015 Workshop on AI and Ethics, Austin Texas USA
- (Cummings 2006) Cummings ML (2006) Automation and accountability in decision support system interface design. *Journal of Technology Studies* 32(1)
- (Defense Science Board 2012) Department of Defense, Defense Science Board (2012) Task Force Report: The role of autonomy in DoD systems
- (Defense Science Board 2016) Department of Defense, Defense Science Board (2016) Summer study on autonomy
- (Franklin and Graesser 1997) Franklin S, Graesser A (1997) Is it an agent or just a program? A taxonomy for autonomous agents. In: ECAI'96 workshop ATAL, Lecture Notes in Artificial Intelligence 1193.

- (Grinbaum et al. 2017) Grinbaum A, Chatila R, Devillers L, Ganascia J-G, Tessier C, Dauchet M (2017) Ethics in robotics research: CERNA recommendations. *IEEE Robotics and Automation Magazine*, January. DOI: [10.1109/MRA.2016.2611586](https://doi.org/10.1109/MRA.2016.2611586)
- (Laumond 2012) Laumond J-P (2012) La robotique : une récidence d'Héphaïstos. Leçon inaugurale prononcée au Collège de France. In French
- (Lin et al. 2012) Lin P, Abney K, Bekey GA editors (2012) *Robot Ethics - The Ethical and Social Implications of Robotics*. The MIT Press
- (MIT) MIT. Moral Machine, <http://moralmachine.mit.edu/>
- (Murphy and Woods 2009) Murphy RR, Woods DD (2009) Beyond Asimov: the three laws of responsible robotics. *IEEE Intelligent Systems Human centered computing*, July-Aug.
- (Pizziol 2013) Pizziol S. (2013) Conflict prediction in human-machine systems. Doctoral thesis University of Toulouse, France
- (Pizziol et al. 2014) Pizziol S, Tessier C, Dehais Fr (2014) Petri net-based modelling of human-automation conflicts in aviation, *Ergonomics* DOI: [10.1080/00140139.2013.877597](https://doi.org/10.1080/00140139.2013.877597)
- (Regis et al. 2014) Regis N, Dehais Fr, Rachelson E, Thooris Ch, Pizziol S, Causse M, Tessier C (2014) Formal Detection of Attentional Tunneling in Human Operator-Automation Interactions. *IEEE Transactions on Human-Machine Systems* 44(3):326-336
- (Ricoeur 1990) Ricoeur P (1990) Ethique et morale. *Revista Portuguesa de Filosofia* 4(1):5-17. In French
- (Sullins 2010) Sullins JP (2010) RoboWarfare: can robots be more ethical than humans on the battlefield? *Ethics and Information Technology* 12(3):263-275
- (Sarter et al. 1997) Sarter ND, Woods DD, Billings CE (1997) Automation surprises. In: *Handbook of Human Factors and Ergonomics*, 2nd ed., Wiley
- (Tessier and Dehais 2012) Tessier C, Dehais Fr (2012) Authority management and conflict solving in human-machine systems. *Aerospace-Lab, The Onera Journal* 4. <http://www.aerospacelab-journal.org/al4/authority-management-and-conflict-solving>
- (Truszkowski et al. 2010) Truszkowski W, Hallock H, Rouff C, Karlin J, Rash J, Hinchey M, Sterritt R (2010) Autonomous and autonomic systems with applications to NASA intelligent spacecraft operations and exploration systems. In: *NASA Monographs in Systems and Software Engineering*
- (Tzafestas 2016) Tzafestas SG (2016) *Roboethics - A Navigating Overview*. Springer
- (Wallach and Allen 2009) Wallach W, Allen C (2009) *Moral Machines - Teaching Robots Right from Wrong*. Oxford University Press

# Chapter 9

## How Children with Autism and Machines Learn to Interact

Boris A. Galitsky and Anna Parnis

### 9.1 Introduction

To act in the world in an autonomous way, humans and machines need to be capable of learning, and as a result of learning they should be able to adequately interact with the world. Successful learning helps in particular to reduce errors humans and machines make operating in the real world. We explore the way humans and machines develop to become autonomous and independently perceive the external world and act on it. It turns out that both machines and children with autism (CwA) experience characteristic difficulties in this development process (Galitsky 2016). Insignificant deviation from the normal development pathway due to sensory properties such as hypersensitivity might lead to autistic cognitive development which makes autonomous behavior of an adult with autism dangerous for himself and others. That is why understanding the mechanism of autistic development is essential for both domains of autistic remediation and building robots enabled with autonomous development (Galitsky and Shpitsberg 2006).

Usually, agents of a multiagent system (MAS) can be characterized by whether they are cooperative or self-interested. Both types of agents need to collaborate with other agents to achieve their goals in uncertain, dynamic domains. This is true for software, human and hybrid agents. In such environments system constraints, resource availabilities, agent goals are changeable, leading MAS to various states. At the same time, such MAS organization needs to be adjusted for environments, there being no single best organization for all possible states. In a broad range of

---

B.A. Galitsky (✉)  
Knowledge-Trail Inc., Oracle Corp. Redwood Shores, CA 95127, USA  
e-mail: [bgalitsky@hotmail.com](mailto:bgalitsky@hotmail.com)

A. Parnis  
Department of Biology, Technion-Israel Institute of Technology, Haifa, Israel  
e-mail: [anna.parnis@gmail.com](mailto:anna.parnis@gmail.com)



**Fig. 9.1** Children with autism learn to interact

MAS applications, a flexible team forming mechanism is required to facilitate automated forming of teams and autonomous adaptation to the environment (Bai and Zhang 2005a). Both software and human agents develop their team forming skills in due course, as a result of active learning with reward (Lopes et al. 2009).

There are established research areas of team formation in the following settings:

- software and hardware agents;
- human agents;
- hybrid/mixed teams.

A vast body of literature addressed team formation scenarios in the above cases, in a broad range of application domains (Bai and Zhang 2005b). These scenarios are usually complex and very domain-specific, so it is hard to judge how general the conclusions that can be drawn. For software and hardware agents, a lot of technical details need to be taken into account. In the case of human agents, psychological analysis makes considerations rather complex and possibly ambiguous.

In this study we focus on the case of *autistic interaction and team formation*, which is expected to shed light on the fundamental properties of the team formation process. Behavior of small children with autism is not as complex as that of control children (CC) of the same age. Furthermore, autistic behavior is simpler than that of software agents, since engineering details do not need to be taken into account (Fig. 9.1). Hence we hypothesize that a team of small children with autism is a much more “pure” environment for studying the phenomenon of team formation compared to conventional investigation platforms for team formation.

By the time control children are verbal, their reasoning and especially handling of mental actions and states is rather complex and hardly tractable. On the contrary, reasoning of autistic children of the comparable mental age is rather simple and allows exploration of its patterns and difficulties applying to real world situations.

In our previous paper (Galitsky 2013), we proposed a reasoning model for autism in which the core deficits, and other related symptoms, emerge as a result of a basic problem with symbolic reasoning about mental states and actions. Our model provided a developmental mechanism required to explain why primary deficits related to social orientation may be the cause for autism and its broader features. Also, this model explains why intensive early intervention by means of stimulating reasoning about mental attitudes frequently helps to improve autistic reasoning. In this study we focus on a particular task of handling interactions with other agents with the goal of team formation, reasoning about mental states. This reasoning domain is a bottleneck of the overall interaction with others and team formation capability. Due to the constraints associated with autistic reasoning about mental states, the reduced capabilities of their “Theory of Mind” (Baron-Cohen 1989), children with autism experience tremendous difficulties interacting with others. Because of the simplicity of autistic reasoning about mental states and actions, as well as reduced learning capabilities of children with autism (Galitsky and Shpitsberg 2014), one can explore simple behavioral patterns during the team formation sessions and trace how these patterns are correlation with reasoning patterns.

## 9.2 From Hypersensitivity to Limited Interaction with the World

### 9.2.1 *Hypersensitivity*

We hypothesize that a route cause of autistic cognition is hyper-sensitivity to input stimuli. To build as simple model as possible and to observe how many features of autistic behavior can be covered by this model, we select only a single deficiency. We then assume that the rest of active learning functions properly and will observe that just a hyper-sensitivity feature of the learning system leads to a broad range of autistic features.

Each child is born with certain perception capabilities. Each child is expected to receive information in a way that fits her perception capabilities. If a child or a robot can see so much, can perceive a certain amount of visual information, then he should be able to process this amount; otherwise the receiving mechanism gradually becomes weaker and weaker. If he can get a certain amount of tactile information, then he expects a corresponding amount of touching (Fig. 9.2). The same is true for any kind of feeling: if a child can feel that much, she is capable of processing that much emotional and feeling-related information.

In autism, the very process of perception of a signal of any sort is associated with discomfort, because an amount of typical real-world amount of information exceeds their perception capabilities, because of a hyper-sensitivity of a child with autism. In CC, about 4/5 of stimuli perception activity leads to positive experience or reward (when stimuli do not exceed perception capabilities), and only 1/5—to negative

**Fig. 9.2** Sensing an object

reward. A CC makes a choice based on perceived stimuli, orienteers in exploration, pursuing 4/5 of unknown stimuli and avoiding the remaining 1/5. If the amount of positive experience associated with exploration exceeds the one for the negative, active world exploration proceeds. Otherwise, if negative experience and failures prevail, then exploration stops and the child chooses a mechanism to avoid exploration. On the contrary, CwA, robots, children with Down syndrome, cerebral palsy, and other mental illnesses experience substantial negative experience from the perception process. Because of the hyper-sensitivity of their perception they fail up to 95% of perception tasks and succeed in only 5%. Therefore their interaction with the external world is formed in a way to minimize negative experience (Bogdashina 2005).

Hyper-sensitivity leads to a failure to learn to recognize stimuli properly, since the system can only learn to recognize patterns with extremely high similarity (as we will show below). This failure leads to a negative experience associated with learning, and as a result CwAs do not investigate the world for the sake of pleasure. Instead they fence themselves from it.

### ***9.2.2 Active Learning in Computer Science***

Traditionally, machine learning has focused on the problem of learning a task from labeled examples only. In many applications, however, labeling is expensive while unlabeled data is usually ample. This observation motivated substantial work on

properly using unlabeled data to benefit learning, and there are many examples showing that unlabeled data can significantly help. There are two main frameworks for incorporating unlabeled data into the learning process.

The first framework is semi-supervised learning (Zhu 2005), where in addition to a set of labeled examples, the learning algorithm can also use a (usually larger) set of unlabeled examples drawn at random from the same underlying data distribution. In this setting, unlabeled data becomes useful under additional assumptions and beliefs about the learning problem. For example, transductive Support Vector Machine (SVM) learning (Yu et al. 2006) assumes that the target function cuts through low-density regions of the space, while co-training assumes that the target should be self-consistent in some way.

The second setting, which is the basis of our model for autistic cognition, is active learning. Here the learning algorithm is allowed to draw unlabeled examples from the underlying distribution and ask for the labels of any of these examples. The hope is that a good classifier can be learned with significantly fewer labels by actively directing the queries to informative examples. One approach is to collect random samples, and another to collect samples which are believed to improve recognition accuracy. Active learning is typically defined by contrast to the passive model of supervised learning. In passive learning, all the labels for an unlabeled dataset are obtained at once, while in active learning the learner interactively chooses which data points to label.

Under active learning, a learning system selects the new elements of the training set automatically. Having the new rules from the newly acquired training set elements, the active learning system is supposed to solve the old problems better. Hence, in addition to a default learning system that is optimized with respect to solving its problem, an active learning system should in turn optimize how to learn the selection of a new training set. An absence of active learning capabilities of a deep learning system, for example, significantly reduces its applicability domains. In such areas as sentiment analysis deep networks are shown to be useful (Zhou et al. 2010).

Active learning as a partial case of unsupervised class is suitable to explain the development of learning abilities of humans and robots since its motivational structure becomes plausible. Humans and robots rewarded for solving problems irrespective of the means, they are responsible for forming their training sets on their own. Phenomenology of autistic deviation from a normal cognition pathway can hardly be explained by learning from a teacher: the result of such learning is either success (a presence of a reward) or failure (an absence of a reward). Peculiarity of autistic learning is that it is very limited in problem solving capability and yet is being rewarded.



### 9.2.3 *Learning Repetitive Patterns*

In the conditions of hyper-sensitivity and overly strong stimuli, CwA is only capable of recognizing a pattern that is extremely close to an element of the training set. A typical case of high-similarity stimuli is repetitive events.

As an example of such stimuli in visual space, let us consider recognition of (1) a child's mother and (2) repetitive TV commercials. Since the perceived image of a mother's face varies more significantly (facial expression, face position, condition of illumination) than the perceived image of TV commercials (which are broadcasted over and over again; they are essentially the same stimuli), the latter turns out to be a preferred type of stimulus that drives the child's development. At the same time, the former stimuli can be filtered out as being too strong (due to its variability and therefore higher recognition efforts). A partial case of stimuli with high similarity is repetitive stimuli, which goes through the whole path of autistic development. All children select to use the most repetitive stimuli among the other stimuli for their training sets; however, autistic children *only* select the most repetitive stimuli and do not proceed beyond them. As a result of this initial problem, CwA stop exploring human behavior and complex behavior of physical objects. Having stopped their explorations, they do not communicate properly with their mothers and other humans because it requires recognition of patterns with a broader range of features.

Usually, most reparative events for a baby are a mother's behavior. She is always nearby, always saying "hi". Babies get used to their mothers as a typical environment, so they accept the belief "I need to adopt to my mother, learn to recognize her." Children from orphan houses have on average lower intellect (Ghera et al. 2009) because at the very beginning they don't have a source of repetitive objects to learn from, and "learning to learn" occurs much slower. A mother is a calibrating instrument for the building of a learning mechanism for a child. Considering reappearance of the mother as the repetitive event, a baby builds its learning mechanism to properly recognize if an approaching object is the mother or not. The baby develops an adaptation rule that is essential for pattern recognition: "If I do too many false positives, increase the threshold. Otherwise, if I do too many false negatives, decrease the threshold."

Mother's reappearance has its own accuracy in terms of new positions, illumination, sounds and frequency, which becomes the set of patterns for a child to optimize her recognition threshold. The mother would never say "hi" in exactly the same way, so the baby should be able to deal with some level of deviation, recognizing the sound. Intonation is different; the mother holds the baby in different ways, wears different clothes, smells differently, etc. The baby can recognize patterns with substantial deviation.

Usually the baby looks for the most repetitive events and finds his mother. In the case with a huge amount of advertisement, repetitive things on radio and TV, machines roaring in the same way, noise from appliances and images can trigger the choice of the learning source of the best repetitive pattern. After that, the baby stops

recognizing the events which have lower precision in their repetition, and loses the skills to do it. Then the mother is rejected because she is too different in appearance.

Repetitions are natural for CC as well, CC repeats the same movement or perception activity, but then proceeds to the exploration of the world to change it and make it better for him. CC applies already developed recognition mechanisms, tuned and tested in repetitions. At the same time, CwA remains in the phase of receiving primary feelings. The role of repetition is not tuning but a reproduction of the same familiar pleasant feelings. By self-stimulation, CwA form feeling directly. Unlike CC playing with a toy car, a CwA avoids grabbing it and passing it over to a peer child. Instead, CwA would just hold and squeeze a toy car. For a CwA an intent to change the world to make it better is reduced to maintaining it in a current, familiar form, since there is a lack of positive experience in exploring and recognizing it.

### 9.2.4 *Self-Stimulation*

In case of autism, there is a failure to determine what is a repetitive event and what is not. CwA consider repetitive only the events that repeat with ideal frequency. Tremendous volume of external information does not make it into CwA. CwA stops perceiving whole stimuli of the real world and only captures elements of these stimuli. This is because the whole stimuli do not fit into the narrow gap formed by autistic cognition trained on the fully repetitive training sets. CwA start to perceive objects and events by their small parts. In these parts, repetitions are most accurate.

At the age of 18 months CwA with their available perception mechanism encounter a necessity to perceive a stimulus as a whole. Then the whole pattern is formed not at the level of causal links between parts, like CC, but instead at the level of unordered sets of these parts. CwA are now getting used to perceive individual parts. When it is necessary to perceive the whole object, CwA attempts to combine these individual parts. CwA continues perceive elements, but not the whole stimulus. CwA want to perceive the world as a whole, but lack a mechanism to do that (Fig. 9.3).

Making efforts to protect themselves from stimuli which are too strong, CwA develop a mechanism to filter out these strong stimuli (which are also more informative) and perceive weaker ones, less informative, but with a higher similarity with each other (Fig. 9.4). Due to hyper-sensitivity, a child with autism is over-selective to the stimuli of external world. We attempt to simulate the phenomenology of early development of the autistic cognition as a choice of perception mode in the conditions of a hyper-sensitive sensory system:

- 1) A child selects, or capable of, recognizing humans such as parents and relatives, which requires multi-modal perception, classification of rather distinct images into a single class, and is then capable of further emotional and mental

**Fig. 9.3** Multi-modal perception



**Fig. 9.4** Example of avoidance behavior



development. Selecting to recognize the subjects of the mental world leads to a normal adaptation.

- 2) A child selects to recognize highly repetitive artificial stimuli such as TV advertisements, smartphone images and sounds, passing cars, and other subjects of the physical world with extremely high similarity. Being forced to recognize the

subjects of the *physical world only* leads to *autistic adaptation*. Autistic adaptation implies *avoidance behavior* to ignore stimuli other than highly repetitive ones.

Human and machine intelligence both experience pleasure from predictability. Control children like to play games, which reflect the world, but reduce its representation to a structure of a limited complexity. Playing games, CC can tolerate a broad range of variability, and wide spectrums of variations are allowed.

On the contrary, CwA will play in a game with zero variability; their doll would keep uttering the same thing in the same way. No deviation in behavior can be handled within the comfort zone of CwA. Whereas CC play with many little cars, CwA would arrange cars in rows: they can only handle a simple element of repetition that is familiar, and therefore rewarding. The range of deviation for repetition is different between CwA and CC: under hyper-sensitivity, a totally novel signal is almost like pain.

Stereotypy or self-stimulatory behavior is usually defined as repetitive body movements or repetitive movement of objects being held by an individual. This behavior is common in many individuals with developmental disabilities and those who experienced institutional care; however, it appears to be more common in autism. In fact, if a person with another developmental disability exhibits a form of self-stimulatory behavior, often the person is also labelled as having autistic characteristics.

Notice that if a machine learning system is fed with very similar elements of the training set, it will have a problem of recognizing even very similar objects to the training ones. Moreover, it will be unable to recognize the ones with significant deviation from the elements of the training set, therefore the whole learning capability will be lacking. To be rewarded, such a learning system would need to find input stimuli that are alike to be able to recognize them. At the same time, to avoid unsuccessful recognitions, the learning system would need to do without complex stimuli, especially those requiring multiple modality signals to be recognized (visual, auditory, tactile). Selectively blocking of a particular modality allows avoiding a stimulus that is too strong (for a machine-learning system, too different to what has been in the training dataset). Hence we conclude that a hyper-sensitivity may lead to a condition where communication between perception systems for vision, speech and tactile feelings are not reinforced and therefore become dysfunctional at the next step of autistic development.

### ***9.2.5 Not Paying Attention to What Is Important***

When one observes the behavior of a child with autism 2–3 year old, it is the second stage of the development process. At this second stage, a child tries to interact with the real world based on the anomalous sensory system built on the first stage. This first stage is primarily oriented at the protection of an unknown stimulus and at finding a familiar stimulus that can be understood.

Two factors lead to this: a broken mechanism of interaction with the real world, and a decrease of the threshold of affective discomfort caused by this interaction. In other words, the latter factor is connected with the increased sensitivity to sensory signals.

Control children learn to recognize objects of the real world correctly because:

1. they improve the technique of focusing on objects rather than on a background. They rely on the skill of ignoring secondary, noisy information; and
2. they are capable of coordination sensory signals from various systems and of the analysis of various properties of objects being recognized.

On the contrary, since the majority of sensory signals is perceived as redundant under autistic development, CwA is forced to learn the process of ignoring, decreasing the volume of these signals. As a result, a child with autism learns to avoid the stimuli that are intended for him.

Instead of systematic development and improvement of sensory systems in the direction of a better understanding of the real world, a child with autism develops a mechanism to ignore signals from the real world (Fig. 9.4). At the same time, a child with autism develops his sensitivity of the signals that carry minimal sensory information. Instead of the frontal direction, which carries important stimuli, a child with autism perceives the peripheral visual and auditory signals. All bright and powerful stimuli are ignored, eye contacts are avoided, and a child is crying when petted. Sensory mechanisms are built in a way to perceive a minimum of sensory information and nevertheless represent somehow the real world. Hence the capability to merge different sensory systems (visual, auditory, kinesthetic) is lacking, binocular vision and binaural auditory systems are not being developed.

### ***9.2.6 From Hyper-Sensitivity to Self-Stimulation of an Engineering System***

People with autism suffer from difficulties in learning social rules from examples, however many remediation strategies have not taken this into account. Therefore an appropriate remediation strategy is to teach not simply via examples (via inductive learning) but instead to teach the appropriate rules (via deduction). The cognitive learning skills of children with autism from the standpoint of active inductive learning have been analyzed in this section. We start with the hyper-sensitivity that leads to the broken links between perceptions of different modalities, a lack of adequate capability to perceive real world stimuli, which then leads to auto-stimulation and autistic cognition. We propose an architecture for a software active learning system which behaves in a similar way, going through the same cognitive steps. The commonalities in deficiencies of autistic and software active learning systems are analyzed. We hypothesize that the autistic learning system, starting with just a hyper-sensitivity feature without other deficiencies, can potentially evolve in a faulty inductive learning system, deviating stronger and stronger from a normally developed system at each iteration of the learning process. This chapter confirms

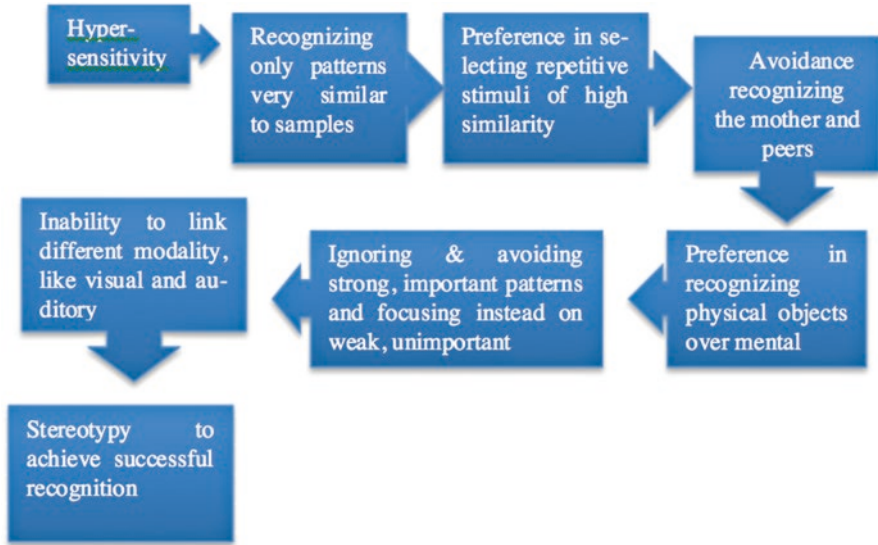


Fig. 9.5 Steps in autistic cognitive development

that the autistic cognitive process is plausible in terms of an abstract computational learning system.

We summarize this section in the chart for the sequence of steps towards autistic cognitive development (Fig. 9.5).

Not just humans can evolve into autistic cognition. A number of poorly designed engineering intelligent systems can recognize only patterns that are very similar to the ones being trained.

One such engineering domain is security: because the system architects intend to avoid false positive in as much degree as possible, they configure the system to issue alerts only for the patterns very similar to which has been identified as true attack or intrusion. False positive is any normal or expected behavior that is identified as anomalous or malicious.

Since it is hard to find real-life positive sets, the creators of security systems demonstrate their functionality on a very limited set of examples. Only these examples are then demonstrated, so from our view what is happening is self-stimulation. Usually active learning is impossible in the security domain.

Another domain where a poorly designed system can only function if self-stimulation mode is search and recommendation. A number of conversational customer support agents can only repeat very closely the dialogues introduced by the creator. Once there is a deviation from such dialog, the system behavior starts being totally meaningless, and it can learn nothing from user inputs.

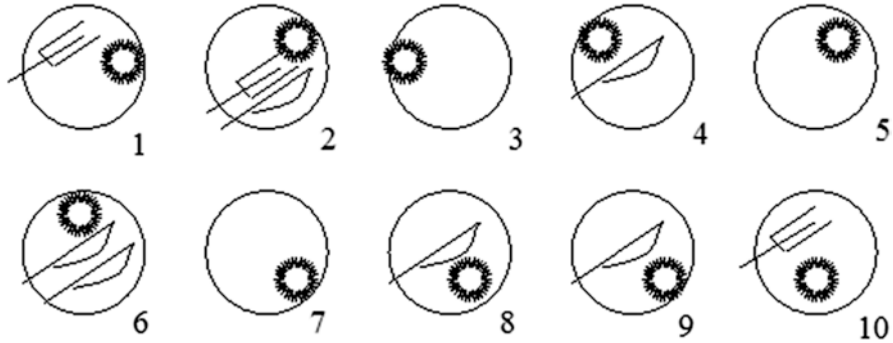


Fig. 9.6 The environment for active learning and hypotheses formation as seen by a subject

### 9.3 Building and Revising Hypotheses in Active Human Learning

We now explored how autistic cognitive development deviates from normal. It is well known that autistic reasoning deviates from that of controls in the way of an absence of certain axioms. Moreover, whereas controls would be able to acquire, memorize and apply these axioms as rules learned from experience, CwA can do neither. In this section we will investigate autistic capabilities of handling hypothesis as a bridge between cognition and reasoning. Are characteristic difficulties CwA experience with solving learning tasks correlated with peculiarities of autistic handling of reasoning tasks?

Having explored how autistic learning develops, we proceed to an experimental setting on how learning hypotheses is followed by manipulation with them (which is traditionally referred to as reasoning). Our accumulated experience of teaching autistic children how to behave properly has contributed to the design of a rule-based machine learning system which automatically generates hypotheses to explain observations, verifies these hypotheses by finding the subset of data satisfying them, falsifies some of the hypotheses by revealing inconsistencies and finally derives the explanations for the observations by means of cause-effect links if possible. This is an active learning system in a sense that samples are selected by the learning system itself to minimize the number of negative samples.

A hungry subject is suggested to eat cookies from the ten plates (Fig. 9.6). The subject is notified that some cookies were altered and added an unpleasant taste in accordance to some rule that is not disclosed. The subject is required to eat all of the cookies with good (expected) taste and state that the rest of cookies are altered. For the purpose of verification, a subject is encouraged to formulate a formed rule when done with the cookies.

When a trainee tries all of the cookies one-by-one, she discovers that cookies from plates 1,3,5,6,7,10 are normal and those from plates 2,4,8,9 were altered and added an unpleasant taste (Fig. 9.7). The objective of this experimental environment

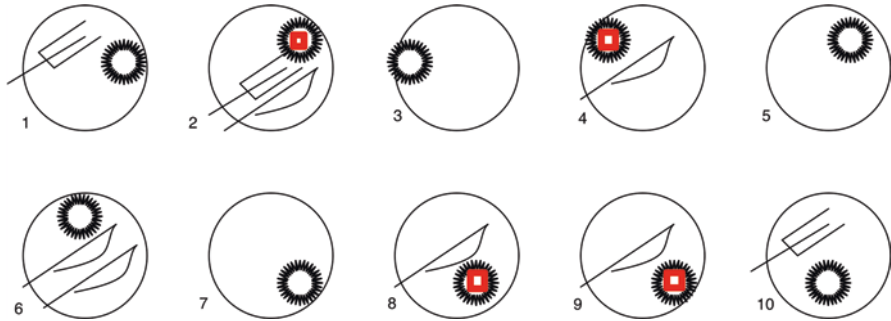


Fig. 9.7 Labeled cookies: 2, 4, 8, 9 are altered

is to come up with an algorithm of forming, confirming and defeating hypotheses such that the least number of cookies with an unpleasant taste is eaten. This environment approximates the real world where human attempt to optimize their behavior. Since it is hard to make CwA act in an artificial environment, this experiment is designed to involve children who are hungry at the beginning of the experimental session. Since children are eager to satisfy their appetite, they don't need to be motivated to participate in a cookie-eating session and they genuinely attempt to avoid altered cookies.

A good way to do minimize a number of cookies with an unpleasant taste eaten, invented by some of the children, is to find the common property of all good cookies and that of the bad cookies. These common properties should not overlap between positive and negative sets. Applying an inductive procedure to positive and negative examples turns out to be a good advancement of both inductive logic programming and explanation-based learning (these methods generalize positive examples only).

A subject is expected to start with a simple hypothesis such as “where there is a fork, the cookie is either normal or altered” or “where there is a knife, the cookie is either normal or altered”. Once a new cookie is encountered, the current hypothesis can be updated or removed in favor of the new one. One of the proper sessions is shown in Table 9.1 where we start with the hypothesis that a fork is associated with a normal cookie, then update this hypothesis adding the “no knife” clause. Then the subject discovers that “fork” is a redundant condition and continues acquiring new samples till she has to transition to “no single knife” instead of the “no any knife” condition.

The experimental results of hypotheses formation for six subjects are shown in Table 9.2. Three out of eight subjects produced an optimal scenario (on the bottom).

The experiments have shown that selected high-functioning autistic subjects outperformed the control children relying on fairly precise means to judge on human intelligence from the perspective of algorithmic decision-making. The strategy selection behavior of normal children was diverse and fuzzy rather than focused on any algorithm. Control children demonstrated decision making in hypothetical



**Table 9.1** The log of a hypothesis forming and revising session

Sample	Hypothesis formed as a result of given sample	Altered
1	Fork→normal	
10	Fork→normal	
2	Fork–knife →normal	Yes
3	–knife→normal	
4	–knife→normal	Yes
7	–knife→normal	
5	–knife→normal	
6	–one knife→normal	
8	Predicted	Yes
9	Predicted	Yes

**Table 9.2** Results of the experiment on forming and operating with hypotheses

Subject	Successful completion	Order of object testing (starting with 1 and finishing with 10)					Additional remarks
		5	7	2	6	1	
Masha Z	–	5	7	2	6	1	No rule is formulated
		9	3	8	10	4	
Lena B	–	4	9	7	8	6	Some attempt to state a rule. Two last altered cookies are determined correctly, but was helped with advices
		3	5	2	10	1	
Valya V	–	7	6	1	9	10	No rule is formulated
		5	4	3	8	2	
Alina Z	–	6	5	9	10	4	Failure to formulate a rule; ate all cookies including altered
		3	8	2	7	1	
Serge T	–	1	3	5	8	6	A wrong rule is suggested: no cutlery—no alterations; also, forks—no alterations. Multiple hypotheses were evaluated but none are correct
		9	4	7	10	2	
Sofia S	+	1	10	2	3	4	Independently achieved the correct rule
		7	5	6	8	9	
Misha P	+	10	1	5	4	3	Achieved the correct rule after some trials and errors
		2	9	6	8	7	
		4	9	1	10	2	Achieved the correct rule after some trials and errors
		6	5	4	3	8	

multi-dimensional space, relying on information about cookies, possible intent of an experimenter, their degree of hunger, the role of cutlery and their inter-relations with cookies, etc.

Conducting these experiments, we observed no link between characteristic difficulties of autistic completion of learning tasks and the ways CwA handle reasoning tasks. We plan to collect more insights on how cognition is linked with reasoning of a human and computer system in our future studies. Figure 9.7 contains the answer to the puzzle: altered cookies are labeled with solid squares.

## 9.4 Building Teams Having Learned to Interact

### 9.4.1 *How Trust Develops in a Baby*

Trust is a baby's inner certainty that the mother is going to help when it is needed (Erikson 1968). This certainty is derived from predictability and consistency of the mother's actions. If mistrust (a model of danger) emerges during the first half year, then the baby is at a disadvantage and this is a path to autistic adaptation. Developing trust in first half-year is necessary to acquire a control over one's affairs. This is also true when a baby grows into a toddler who is expected to succeed in toilet training, feeding independently, bathing and interacting with known people.

Mistrust around a child is strengthened with the impression that the world is unpredictable, and it is another feature of autistic development. It keeps CwA from expanding his world and exploring his opportunities in this world. For a control child, if the mother is inconsistent in her availability and her care for the baby then there is a risk that this baby develops into a mistrusting child and will not integrate with the external world. Success in this stage will lead to the virtue of hope. By developing a sense of trust, the infant can have a feeling that as new crises arise, there is a real possibility that other people will be there as a source of support. Failing to acquire the virtue of hope will lead to the development of fear.

For example, if the care has been harsh or inconsistent, unpredictable and unreliable, then the infant will develop a sense of mistrust and will not have confidence in the world around them or in their abilities to influence events. This infant will carry the basic sense of mistrust with them to other relationships. It may result in anxiety, heightened insecurity, and an over-feeling of mistrust in the world around them.

The repetitiveness and sameness of actions (Sect. 9.2.3), behavior and facial expressions carried out by the mother at the initial step of development eventually create a set of symbols in the baby's mind. This is how a baby's trust is developing. These symbols come to represent safety in interaction and having a calming effect. Then when these symbols of familiarity and predictability come up later in a toddler's life, these symbols will provide a social comfort. Trust development vary in how much time it takes to be accomplished. A mother can recognize if her baby develops trust in her constant presence through the following. When the mother leaves the room and observes the baby's reaction, one of two can be seen:

- 1) The baby reacts with anxiety, frowning, erratic movements, and a crying spell;  
or,
- 2) The baby does not react and continues without changing.

The former means that the trust has not been established yet. Once trust has been established (2) the mother can be more flexible with her delegation of caregiving. When the baby has acquired trust, her tensions significantly decreases and she will ask for attention less frequently; separation between self and the environment proceeds along with the baby's feeling of independence.

### 9.4.2 *Measuring Skills of Reasoning About Mental World*

We explore how children with autism form teams to perform simple tasks, and what kind of reasoning is required for that. The focus of our experiment is to find a correlation between how children do reasoning about the mental world, and how they perform team formation tasks. The underlying model for our correlation is a belief-desire-intention (BDI; in Rao and Georgeff 1995) model for a multiagent system.

To assess reasoning capabilities of children, we ask them questions about mental states of characters, and evaluate the correctness of their answers (Galitsky et al. 2011). We hypothesize that while team undergoes formation, they have to initiate the same or similar questions before they perform speech acts with their proponents and possibly opponents. The questions involve first order mental states (*do you know...?, does she want...?*), second order (*do you want him to believe ...?*), third order (*he believes she wanted him to know that she wanted ...*), and fourth-order (*he knows she wanted him to know that she does not want ...*). Order characterizes how many verbs for mental states and communicative actions are nested. A good example here is of the Federal Reserve chairman Alan Greenspan: “I know you think you understand what you thought I said but I’m not sure you realize that what you heard is not what I meant.”

We used the following team formation tasks. These are the tasks CwA of age 6–10 usually experience difficulties with, being fairly easy for the CC. These tasks rely on various physical actions, but the commonality between them is the necessity to reason about beliefs and intentions of other team members:

- “hide-and-seek games”, where children need to agree who is hiding and who is searching;
- “hiding an object in a bag” games;
- making one participant do something with the second participant what the third participant wants;
- form a team of buyers to shop for the items of mutual interested;
- form small soccer, football or basketball teams, two versus two;
- form chess playing teams, taking turns in moves, two versus two.
- completing other kinds of joint task (Figs. 9.8 and 9.9).

Each task required five–eight participants. Thirty-two children of age 6–10 participated in all team-building tasks and completed all reasoning exercises.

We split CwA into four groups with respect to their capabilities in team formation:

1. *Active team builder* who can *initiate* a new team;
2. *Active team builder* which can *maintain* the team performing tasks and encourage others to do so;
3. *Passive team members* who *can be maintained* to be a part of the team being *encouraged by other members*. They cannot initiate team formation themselves, but they *can resume* the team activity after it has stopped;
4. *Passive team members* who can be maintained to be a part of the team. They can *neither initiate team formation themselves*, nor resume team activity.

**Fig. 9.8** An illustration for completing a joint task



**Fig. 9.9** An illustration for formation of a larger size team

For each child, we assign him to a group if he is capable of performing the required team formation function in more than a half of the scenarios. Notice that some team building scenarios require verbal communication, and some rely on non-verbal ones.

The joint results of the reasoning assessment and team formation assessment are shown in Table 9.3. Rows indicate the percentages of successfully completed reasoning tasks for each group of team formers (averaged through eight individuals). Rows are grouped from top to bottom according to the order of formulas required to answer the respective question. As we indicated, an *order* is defined as a count of mental verbs in a natural language phrase expressing a mental state or action.

**Table 9.3** Team formation skills as a function of reasoning about mental states capabilities

Roles	Active team builder		Passive team member		Controls
	initiate	maintain	maintain	resume	
knowing an object and its attributes	95	91	82	72	95
not seeing-> not knowing	90	93	78	80	90
intention of yourself	88	90	80	76	95
intention of others	92	87	71	70	95
informing	87	84	78	73	90
information request	91	89	72	71	85
asking to do an action	78	83	80	75	90
asking to help	85	80	70	75	90
questioning	81	83	68	70	85
explaining	72	70	61	64	85
agreeing	76	73	64	60	90
pretending	81	76	65	62	90
deceiving	70	64	62	54	80
offending	73	68	58	50	85
forgiving	72	62	61	46	80
reconciling	65	64	50	39	85
disagreeing	72	69	42	40	75
inviting to help	62	59	39	46	70
asking to leave	64	57	40	51	85
asking not to interfere	70	50	38	32	70
disagreeing	62	46	32	28	65
resolving a conflict	42	37	17	12	65
negotiating	48	24	12	7	60

Dark grey area shows good performance of reasoning tasks (more than 70%) and light-grey show lower performance (60–70%). The white area shows the level of reasoning complexity this group of team formers cannot reliably achieve. Mental states and actions of reasoning exercise are ordered in the way of increasing complexity (averaged performance). Columns are formed according to the four groups of children above.

One can observe a strong correlation between the reasoning complexity order and team forming capabilities. If children cannot perform even the first-order reasoning tasks, they are neither capable of team forming nor understanding of team forming by others. To be capable of team forming, second-order reasoning needs to be satisfactory.

The third-order mental states are the ones the trainees experience the most difficulties with. Various skills at these tasks differentiate children with autism into two groups:

- 1) those who can initiate new teams, and
- 2) those who can maintain team activities and resume team operations.

For the former group, substantial third-order reasoning is required, and for the latter, just rudimentary third-order skills suffice.

Finally, fourth order mental states are difficult for both children with autism and controls of comparable age (see the rightmost column for evaluation of team formation by the control group).

### 9.4.3 A Cooperation Between CwA in the Real World

We observed the team formation behavior in the real world as a part of the intervention program conducted by the Center for children with special needs “Sunny World” ([www.solnechnymir.ru](http://www.solnechnymir.ru)). The children in the summer camp were forming teams with the help of intervention personnel and parents, performing various farming tasks. These tasks include harvesting and packaging vegetables into boxes. Children had to agree on who is doing what, how to store and pass vegetables between each other and in what order, and how to handle varying harvesting conditions (Fig. 9.10). The difficulty level for this task is of the order two and three in most cases.

The children who participated in our evaluation study and successfully formed teams in artificial scenarios were also capable of forming teams for the agricultural

**Fig. 9.10** A team of children at work (Sunny World 2014)



tasks. On the contrary, those who could not adequately participate in our assessment had significant difficulties in performing the tasks requiring interaction with other team members, performing farming tasks.

Performance assessment is difficult in farming teams because of a lack of repetition and systematic framework in the farming tasks. Unlike the team formation exercises, which also included conflict scenarios, farming ones involved cooperation only, avoiding any kinds of conflicts. However, the overall impression of the personnel and the parents was that doing abstract team formation helped some children to understand mental states sufficiently to form cooperative teams.

Team formation in the real world demonstrates how the notion of *trust* is perceived by the reduced reasoning of children with autism. Trust becomes a mental state with certain rules, compared to the trust states that are learned by control human and software agents. Trust is explicitly defined via communicative actions of *promise* and *believe*:

*trust(Who, Whom):-  $\forall$ Subject promise(Whom, Who, Subject), believe(Who, Subject).*

and serves as an additional constraint for a team formation rule: engage with trusted partners. In this respect the notion of trust is simpler than in the general case of adequate reasoners, which need to acquire trust in the course of a dynamic process (Lawless et al. 2013). The intelligence in the form of rules to reason about a mental world cannot be labeled as robust, in our opinion, since autistic reasoning cannot be adjusted to a given environment in an autonomous manner.

Yi et al. (2013) investigated whether CwA had an indiscriminate trust bias. The question of this study is whether a CwA would believe in any information provided by an unfamiliar adult with whom they had no interactive history. Young school-aged CwA and their age- and ability-matched CC participated in a simple hide-and-seek game. In the game, a caregiver with whom the children had no previous interactive history pointed to or left a marker on a box to indicate a location of a hidden reward. Results showed that although CwA did not blindly trust any information provided by the unfamiliar adult, they tend to be more trusting in the adult informant than control children do.

For an abstract reasoning system, experiencing difficulties in forming teams does not necessarily mean that deficiencies are in the domain of reasoning about the mental world. It could be a general autistic incapability to adjust to a given environment (Galitsky and Peterson 2005), general problems in non-monotonic reasoning (Galitsky and Goldberg 2003; Galitsky 2007), autistic planning (Galitsky and Jarrold 2011) and autistic active learning (Galitsky and Shpitsberg 2014). However, we discovered in this chapter out that the *root cause* of autistic difficulties in team formation are due to reasoning in the mental domain, as demonstrated by its direct correlation with the real world performance.

## 9.5 Rehabilitating Autistic Interactions

### 9.5.1 Teaching Hide-and-Seek Game

Learning to play hide-and-peek game is one of the important steps in learning the mental world. It is also a good team formation exercise. This game requires a substantial reasoning about mental states and actions, in both rule-based mental and emotional domains. A child needs to understand the pre- and post-conditions for searching as a desire to identify where the peers are located. A concept of *hiding* needs to be explained as an *opposite* desire of not being found. Children need to be aware that searching may lead to finding, and hiding—to not being found. If one does not search then nobody can be found, and if one does not hide she will be found immediately. It is a game of deception, which requires acknowledgment that other people may have different beliefs. Therefore many CwA avoid it and/or are not capable of participating in it. Playing hide-and seek requires understanding and handling third-order mental states such as “*I know that he wants me not no know where he is*”.

In the emotional space, a hide-and-peek player is expected to express appropriate emotions when he finds another child, or when he is found by someone else. A rule should be taught that an emotion is appropriate when there was a desire and at the given moment it succeeds. Some emotional expressions are suitable when a child is hiding; he is being looked at but not found.

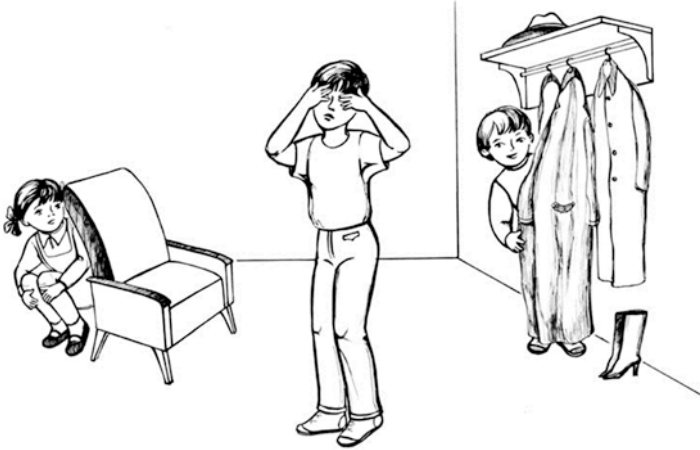
Another important skill is to conceal yourself in an environment. A child needs to be taught to position himself in the location of a seeker and track his potential gaze to avoid being found. A seeker needs to be able to close his eyes and count to a predetermined number while the other players hide. After reaching the number (such as reaching 10 or 20), the seeker attempts to say, “Ready or not, here I come!” and then to locate all concealed players (Fig. 9.11).

Training starts with identifying hide-and-peek players in an image with schematic depiction of playing characters. CwAs are encouraged to use a touch-pad to track the gaze with their fingers. Children are asked questions about the role of players, who is doing what, who desires what, and who is seeing whom.

After CwA trainers are capable of recognizing players in an image, a trainer can proceed to similar tasks on the photos of children playing hide-and-peek (Fig. 9.12) and ask similar questions:

What game do the children play?  
 Which objects from the environment are used to be hiding behind?  
 Do those who hiding want the seeker to find them?  
 Does the seeker want to find those hiding?  
 Do the hiding children see the seeker? Do they know where he is?  
 Does the seeker see the hiding children? Does he know where they are?  
 Why does the seeker have to close his eyes?





**Fig. 9.11** The hide-and-seek training starts with schematic depiction of a seeker and two concealed players



**Fig. 9.12** After CwA is confident with schematic depiction of hide-and-seek game, a trainer can proceed to photos. The seekers close their eyes and are counting

Once CwA are prepared to play hide and seek, having completed the exercises, a trainer can attempt to involve them in an actual game, first indoor and then outside. To play a role of a seeker or to hide, a CwA needs to be accompanied by a trainer, and a role of an opponent can be performed by a parent, sibling or another trainer. The trainer needs to hide together with CwA and explain her the goal of hiding and the object they are hiding behind.



**Fig. 9.13** An older trainee finding a direction using GPS (on the *left*). Some young adults become fairly skillful once the introduction to orienteering with GPS is completed (on the *right*)

### 9.5.2 *Learning to Navigate Environment*

For most CwA, orienteering is the next logical step after hide-and-seek. However, some children are good at orienteering even if their emotional skills for hide-and-seek are rudimentary and they cannot play independently.

The reason orienteering is not too hard for CwA is that no reasoning about another human is required. A CwA usually memorizes the commands and navigation of GPS menus in no time. A CwA needs to associate what GPS is showing with what is observed in the real world (Fig. 9.13). Doing that, formulating, adjusting and rejecting of hypotheses of such association is required, based on hypotheses management exercise in Sect. 9.3.

The main focus of how orienteering activity supports reasoning is hypotheses management. Looking at a GPS, the child obtains the direction to and distance to the goal. Then observing the landscape, the child selects an object such as a tree and forms an estimate for how far it is from this tree to the goal (Fig. 9.14).

Once the tree is reached, CwA observes her position relative to the goal and possibly updates the hypothesis on where she was relative to the goal. CwA now needs to form a new hypothesis on which direction in the landscape to choose and which position relative to the goal to expect, and proceeds towards the goal.

What this exercise teaches is the skill to maintain hypotheses, revise them when appropriate, and expect one to be wrong again and again. This is opposite to a conventional autistic reasoning which sticks to a given hypothesis once it is formed. After that, CwA will be reluctant to revise this hypothesis, and an observation that it does not fit the real world would be very stressful and unproductive: CwA would give up on the exercise.

**Fig. 9.14** A trainee is being helped to link the GPS indication with the real world spatial references



### 9.5.3 *A Literary Work Search System*

Once a trainee is familiar with mental formulas and is capable of forming simple scenarios from it, he should proceed to formulating questions in the mental world. A rich and extensive domain in the mental world is the one of the fictional characters in a narrative work of art (such as a novel, play, television series or film). In this section we propose a reasoning exercise based on formulating queries and searching for a literary work.

The methodology and abstraction of such searches are very different from those for database querying, keyword-based search of relevant portions of text, and search for the data of various modalities (speech, image, video etc.). Clearly, the search that is based on mental attributes is supposed to be enriched with meanings versus just keywords. Obviously, using just the author name or title is trivial. Also, using temporal (historical) and geographical circumstances of the characters reduces the literary work search to the relatively simple querying against the relational database of literary work parameters.

We have built the dataset of a literary works, which includes the manually extracted mental states of their characters. We collected as many a literary works as was necessary to represent the totality of mental states, encoded by logical formulas of the certain complexity (Galitsky 2000). Below are the features of this dataset:

1. As a rule, the main plot of a literary work deals with the development of human emotions, expressible via basic (want-know-believe) and derived (pretend, deceive, etc.) mental predicates. A single mental state expresses the very essence of a particular a literary work for small forms (a verse, a story, a sketch, etc.). When one considers a novel, a poem, a drama, etc., which has a more complex nature, then a set of individual plots can be revealed. Each of these plots is depicting its own structure of mental states that is not necessarily unique. Taken all together, they have the highly complex form, appropriate to identify a literary work.
2. Extraction of the mental states from a literary work allows us to clarify psychological, social and philosophical problems, encoded by this work. The mental components, in contrast to the “physical” ones are frequently expressed implicitly and contain some ambiguous expressions.
3. The same mental formula may be a part of different literary works, written by the distinguishing authors. Therefore, it is impossible to identify a certain literary work or author when we take into consideration just a single mental formula. However, the frequency of repetition of certain mental formulas shows us the importance of the problem raised by a literary work.
4. The sets of mental formulas are sufficient to identify a literary work. The possibility to recognize a certain author according to a collection of mental states of his or her literary work s is beyond our current consideration.

We enumerate the tasks that have to be implemented for the literature search system based on the scenario reasoning settings:

- 1) Understanding a natural language query or statement (Galitsky 2003). This unit converts a NL expression into a formalized one (mental formula), using mental metapredicates and generic predicates for physical states and actions.
- 2) Domain representation in the form of semantic headers, where mental formulas are assigned to the textual representation (abstract) of a literary work.
- 3) A reasoning engine (Natural Language Multiagent Mental Simulator, NL\_MAMS (Galitsky 2013)) that builds hypothetical mental states, which follow the mental state, mentioned in the query. These generated hypothetical mental states will be searched against the literary work knowledge base together with the query representation (in unit 5).
- 4) Synthesis of all well-written mental formulas in the given vocabulary of basic and derived mental entities.
- 5) Matching the mental formula, obtained for a query against mental formulas, associated with literary works. We use an approximate match in case of failure of a direct match.
- 6) Synthesis of canonical NL sentence based on mental formula to verify if the query was properly understood

Figure 9.15 presents a chart for the unit components (1)–(6) of the literary work search system. There are two functioning options: (a) literary work search and (b) extension of the literary work dataset. When a user wishes to add a new literary

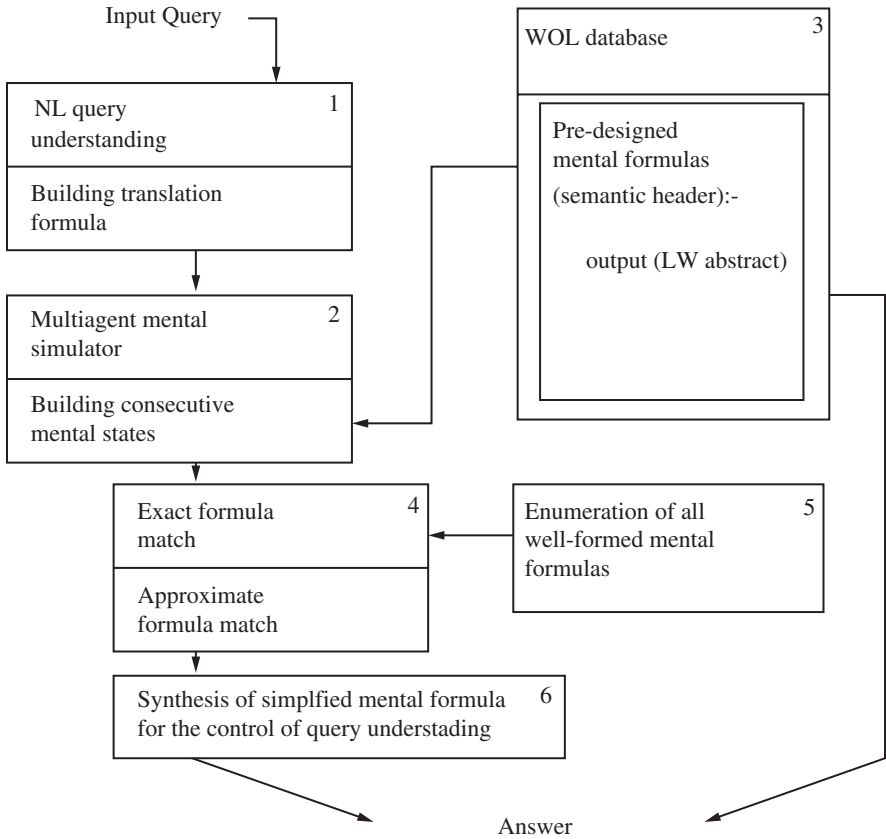


Fig. 9.15 The chart of the WOL search and mental reasoning system

work to the current dataset, the formulas for the mental state associated with text are automatically built by Unit 1 and are subjected to variations for semantically different phrasings by Unit 2.

Rather complex semantic analysis (Unit 1) is required for exact representation of an input query: all the logical connectives have to be properly handled. Unit 2 provides the better coverage of the literary work domain, deductively linking mental formula for a query with mental formulas for literary works.

Plausible mental formulas are extracted from the totality of all well-written mental formulas, represented via metapredicates. In addition, introduction of the classes of equality of mental formulas are required for the approximate match of mental formulas (Unit 4) that are also inconsistent with the traditional formalizations of reasoning about knowledge and belief. NL synthesis of mental expression (Unit 6) is helpful for the verification of the system’s deduction. A trainee needs this component to verify that she is understood by the system correctly before starting to evaluate the answer. NL synthesis in such a strictly limited domain as mental expression

<p>How would a person pretend to another person that she does not want that person to know something?</p> <p>When would a person want another person not to pretend that he does not know something?</p> <p>When would a character pretend about his intention to know something?</p> <p>Why would a person want another person to pretend about what this other person want?</p> <p>How can a person pretend that he does not understand that other person does not want?</p> <p>Is it easy for a person to believe that another person does not pretend what she wants?</p> <p>How can a person believe that another person might pretend that he wants something?</p> <p>She wanted to believe that he pretended that he was not a prince.</p> <p>Can she believe that he does not pretend that he committed the murderer of her spouse because of his love to her?</p> <p>A person believes that the husband does not want him to love his wife.</p> <p>A wife wishes not to confess to her husband that she was not faithful.</p>
--

**Fig. 9.16** Sample questions for the literature search

is straightforward and does not require special consideration. Note that semantic rules for the analysis of mental formulas require specific (more advanced) machinery for complex embedded expressions and metapredicate substitutions.

The special question-answer technique for poorly-structured domains has been developed to link the formal representation of a question with the formal expression of the essential idea of an answer. These expressions, enumerating the key mental states and actions of the literary work characters, are called *semantic headers of answers* (Galitsky 2003). The mode of knowledge base extension (automatic annotation), where a user introduces an abstract of a plot and the system prepares it as an answer for other users, takes advantage of the flexibility properties of the semantic header technique.

To summarize, the literary work system architecture is as follows. A NL query that includes mental states and action of a literary work character is converted into mental formula (Unit 1). Multiagent mental simulator (Unit 2) yields the set of mental formulas, associated with the query to extend the knowledge-base search. Obtained formulas are matched (Unit 4) against the totality of prepared semantic headers (mental formulas) from the literary work database (Unit 3). If there is no semantic header (mental formula attached to text) in the dataset component that satisfies the mental formula for a query, the approximate match is initiated. Using the enumeration of all well-formed mental formulas (Unit 5), the system finds the best approximation of the mental formula for a query that matches at least single semantic header (mental formula for an answer).

Interaction with the literature characters is a new effective and efficient education means for children, interacting with the characters of the scenes in NL (Fig. 9.16). Since the players are suggested to both ask questions and share the literature knowledge, the system encourages the cooperation among the members of the players' community. In the demo we have built, the system only recognizes the questions and statements, involving the terms for mental states and actions. This way we encourage the players to stay within a "pure" mental world and to increase the complexity of queries and statements we expect the system to handle properly. Observing the game players, we discovered that they frequently try to obtain the exhaustive list

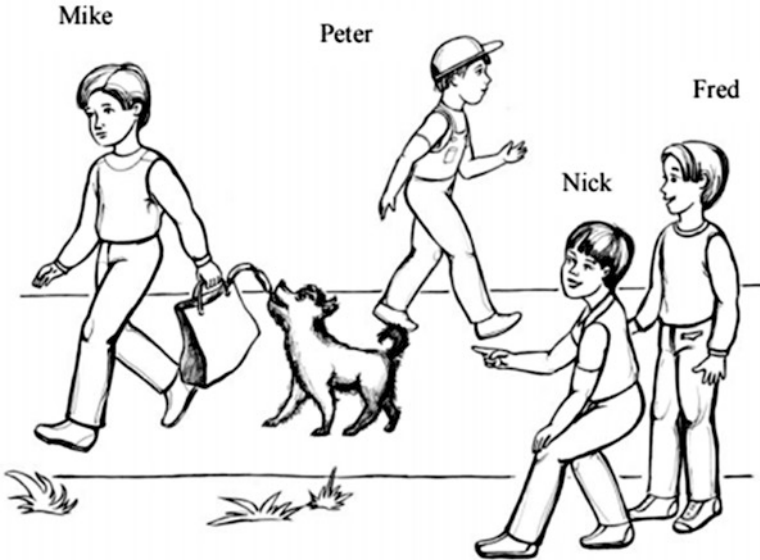
<p>WOL search system allows a literature fan to extend the knowledge base with the new favorite story or novel and to specify the major ways of accessing it (asking about it). This toolkit processes the <b>combination</b> of the answer (an abstract of a story, introducing the heroes and their interactions) and a set of questions or statements (explicitly expressing the mental states these interactions are based on).</p>	
<p>When does a person pretend about her intention to know something?</p>	<p><b>The Carriage of holly gifts</b> by P. Merimee An old-aged king wants to learn from his secretary if the young girl he loves is faithful to him. The secretary is anxious to please the king...</p>
<p><i>Add to Knowledge</i></p>	<p><i>Compile Knowledge base</i></p>
<p><b>Domain extension code:</b> <i>pretend(person, other_person, want(person, know(person,Smth))) :- do201.</i> <i>do201:-output(\$The Carriage of holly gifts... \$).</i> <b>Domain is compiled. Ask a question to the updated domain</b></p>	
<p><input type="text"/></p>	<p><b>Ask</b> Now you can ask the questions for the domain extension as well as for the base domain, varying the phrasings.</p>

Fig. 9.17 An autistic child learns the mental interaction with the characters (participants of the scene), using the suggested system

of literary works, memorize the querying results and enjoy sharing WOL plots with the others.

The demonstration encourages the users (players, students) to demonstrate their knowledge of classical literature, from medieval to modern, asking questions about the mental states of the characters and compare the system’s results with their own imagination. The system stimulates the trainees to extract the mental entities, which can be formalized, from the totality of features of literature characters. After an answer is obtained, it takes some efforts to verify its relevance to the question. It takes a little variation in the mental expression to switch from one literary work to another. More advanced users are offered the option of adding a new literary work. For mental a intervention (particularly, CwA), certain visualization aids are useful in addition to the literary work search system (Fig. 9.17).

Examples of questions the children may ask the system, while watching the scene, are shown in Fig. 9.18. Involving more and more complex mental states helps



**Fig. 9.18** Example queries

- |  |
|--|
| <ol style="list-style-type: none"> <li>1. Does Mike see that the dog is eating the sausages?</li> <li>2. Does Peter see what is happening with Mike and the dog?</li> <li>3. Does Nick know what is happening with Mike and the dog?</li> <li>4. How does Nick express his emotions?</li> <li>5. Does Fred know whether Peter knows what is happening with the sausages?</li> <li>6. Does Nick want to keep the dog from eating the sausages?</li> <li>7. What would Fred do if he wants to let Peter know what is happening?</li> </ol> |
|--|

**Fig. 9.19** A scene that serves as a playground for asking questions about mental states

(Fig. 9.19) the playing children to develop creativity and imagination, as well as the communication skills of understanding mental states of others.

## 9.6 Discussion and Conclusions

Recent studies (e.g., Dawson et al. 2007) have reported that autistic people perform in the normal range on the Raven Progressive Matrices test, a formal reasoning test that requires the integration of relations as well as the ability to deduce behavioral rules and to form high-level abstractions. Morsanyi and Holyoak (2010) compared autistic and control children, matched on age, IQ, and verbal and non-verbal working memory, using both the Raven test and pictorial tests of analogical reasoning. They found that autistic children’s reasoning capabilities are similar to those of



controls on reasoning with relations tests. The authors concludes that the basic ability to reason systematically with relations in the physical world, for both abstract and thematic entities, is intact in autism.

Gokcen et al. (2009) investigated the potential values of executive function and social cognition deficits in autism. While the theory of mind is generally accepted as a whole, a number of researchers suggested that it can be separated into two components (mental state reasoning and decoding). Both aspects of the theory of mind and verbal working memory abilities were investigated with the focus on mental reasoning for parents of children with autism, who had verbal working memory deficits as well as a low performance on a mental state reasoning task. The parents had difficulties in reasoning about others' emotions. In contrast to findings in the control group, low performance of mental state reasoning ability was not associated with a working memory deficit in control parents. Social cognition and working memory impairments may represent potential genetic risks associated with autism.

In the physical world, children with autism perform relatively well so it should not be a limitation for their team formation capabilities. Autistic participants outperformed non-autistic participants on abstract spatial tests (Stevenson and Gernsbacher 2013). Non-autistic participants did not outperform autistic participants on any of the three domains (spatial, numerical, and verbal) or at either of the two reasoning levels (concrete and abstract), suggesting similarity in abilities between autistic and non-autistic individuals, with abstract spatial reasoning as an autistic strength.

For an abstract reasoning system, experiencing difficulties in forming teams does not necessarily mean that deficiencies are in the domain of reasoning about the mental world. It could be a general incapability to adjust to a given environment (Galitsky and Peterson 2005), general problems in non-monotonic reasoning (Galitsky and Goldberg 2003; Galitsky 2007), autistic planning (Galitsky and Jarrold 2011) and autistic active learning (Galitsky and Shpitsberg 2014). However, it turned out that the root cause of autistic difficulties in team formation are due to reasoning in the mental domain, as demonstrated by its direct correlation with the real world performance.

We explored team formation at the following levels:

1. Abstract reasoning in mental world
2. Team formation in controlled, assessment tasks
3. Team formation in real world

We found a strong correlation between (1) and (2), and a weak, qualitative correlation between (2) and (3). We used the computational tool capable of solving similar problems (reasoning about mental states, Galitsky 2013) to the ones which were given to CwA. In the case of children, we simulated the peculiarities of autistic reasoning on one hand and supported rehabilitation exercises on the other hand. We used the following hybrid teams of agents: autistic + autistic, autistic + control and autistic + software agents.

We found that the main determining feature of autistic team formation is their reasoning capabilities. This observation can be extended to the case of software agents, where behavioral algorithms can be affected by a broad range of circumstances.

For software agents, the bottleneck of reasoning about mental states can be less noticeable, but we expect it to be as almost as strong as for the case of autistic reasoning.

Our study has certain implications for how the autonomy features of abstract agents can be modeled via aspects of human behavior. Obviously, autistic reasoning not only leads to unusual and frequently inappropriate behavior but also causes error in controlling the outside world. Our finding confirms the theory of social interdependence in its simplest form, applied to naïve autistic reasoners: once agents become capable of operating in the mental world, they are able to form teams: no special, additional skills are required. Once children form teams, their mental reasoning capabilities improve, but they do not need to learn anything more besides mental states and actions to learn about forming simple teams. In this respect, our findings back up the traditional individual methodological perspectives (e.g., cognitive architectures). They assume that individuals are more stable than labile from the social interactions in which they engage: once individual reasoning skills are adequate, the collective behavior becomes adequate as well.

## References

- Bai Q. and Zhang, M.,2005a. Dynamic Team Forming in Self-Interested Multi-Agent Systems. Sydney, Australia, LNAI Vol 3809, Lecture Notes in Artificial Intelligence, Springer-Verlag Berlin Heidelberg, pp. 674-683.
- Bai Q. and Zhang, M.,2005b. Flexible Agent Team Forming in Open Environments, In Proceedings of the Fifth International Conference on Intelligent Technology, Phuket, Thailand, pp. 402-407.
- Baron-Cohen, S., 1989. The autistic child's theory of mind: a case of specific developmental delay. *Journal of Child Psychology and Psychiatry*, 30, 285-297.
- Bogdashina, Olga. Communication Issues in Autism and Asperger Syndrome: Do We Speak the Same Language? Jessica Kingsley Publishers, 2005.
- Dawson M, Soulières I, Gernsbacher MA, Mottron L. The level and nature of autistic intelligence. *Psychol Sci*. 2007 Aug;18(8):657-62.
- Erikson, Erik H. (1968) Identity, youth, and crisis. New York: W. W. Norton.
- Galitsky B. (2000) Simulating autistic patients as agents with corrupted reasoning about mental states. AAAI FSS-2000 Symposium on Human Simulation, Cape Cod, MA.
- Galitsky B., Shpitsberg I. Evaluating Assistance to Individuals with Autism in Reasoning about Mental World. Artificial Intelligence Applied to Assistive Technologies and Smart Environments: Papers from the 2015 AAAI Workshop.
- Galitsky B., Shpitsberg I. 2006. How one can learn programming while teaching reasoning to children with autism. AAAI Spring Symposia Stanford CA.
- Galitsky, B. 2003. Natural language question answering system: technique of semantic headers. *Advanced Knowledge Intl*. Adelaide Australia.
- Galitsky, B. 2007. Handling representation changes by autistic reasoning. AAAI Fall Symposium - Technical Report FS-07-03, pp. 9-16.
- Galitsky, B. & Goldberg, S. 2003. On the non-classical reasoning of autistic patients. International Conference on Neural and Cognitive Systems Boston University, MA.
- Galitsky, B. 2013. A computational simulation tool for training autistic reasoning about mental attitudes. *Knowledge-Based Systems*, Volume 50, September 2013, Pages 25–43, 2013.

- Galitsky, B. and Peterson, D. 2005. On the Peculiarities of Default Reasoning of Children with Autism. FLAIRS-05.
- Galitsky, B. 2016. Computational Autism. Springer Human-Computer Interaction Series.
- Galitsky, B., de la Rosa JL, Boris Kovalerchuk, B. 2011. Discovering common outcomes of agents' communicative actions in various domains. *Knowledge-Based Syst.* 24(2): 210-22.
- Galitsky, B., Jarrold, W. 2011. Discovering patterns of autistic planning. AAAI Workshop - Technical Report WS-11-16, pp. 2-9.
- Galitsky, B., Shpitsberg, I. 2014. Finding faults in autistic and software active inductive learning. AAAI Spring Symposium - Technical Report.
- Ghera M, Marshall P, Fox N, Zeanah C, Nelson CA, & Smyke AT (2009). The effects of foster care intervention on socially deprived institutionalized children's attention and positive affect: Results from the BEIP study. *Journal of Child Psychology and Psychiatry*, 50: 246-253.
- Gokcen S, Bora E, Eremis S, Kesikci H, Aydin C. Theory of mind and verbal working memory deficits in parents of autistic children. *Psychiatry Res.* 2009 Mar 31;166(1):46-53.
- Lawless, W. F., Llinas, J., Mittu, R., Sofge, D.A., Sibley, C., Coyne, J., and Russell, S. 2013. Robust Intelligence (RI) under uncertainty: Mathematical and conceptual foundations of autonomous hybrid (human-machine-robot) teams, organizations and systems. *Structure and Dynamics*, 6(2).
- Lopes, M., Melo, F. and Montesano, L. 2009. Active Learning for Reward Estimation in Inverse Reinforcement Learning. ECML.
- Morsanyi K, Holyoak KJ. Analogical reasoning ability in autistic and typically developing children. *Dev Sci.* 2010 Jul;13(4):578-87.
- Rao, M., P. Georgeff. 1995. BDI-agents: From Theory to Practice. Proceedings of the First International Conference on Multiagent Systems (ICMAS'95).
- Stevenson, JL, Gernsbacher MA, 2013. Abstract Spatial Reasoning as an Autistic Strength. *PlosOne*. [10.1371/journal.pone.0059329](https://doi.org/10.1371/journal.pone.0059329).
- Yi, L., Junhao Pan, Yuebo Fan, Xiaobing Zou, Xianmai Wang, Kang Lee. 2013. Children with autism spectrum disorder are more trusting than typically developing children. *Journal of Experimental Child Psychology*.
- Yu, K., J. Bi, and V. Tresp. Active learning via transductive experimental design. In Proceedings of the International Conference on Machine Learning (ICML), pages 1081-1087. ACM Press, 2006.
- Zhou, S., Qingcai Chen, and Xiaolong Wang. 2010. Active deep networks for semi-supervised sentiment classification. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters (COLING '10). Association for Computational Linguistics, Stroudsburg, PA, USA, 1515-1523.
- Zhu X. Semi-supervised learning literature survey. Computer Sciences Technical Report 1530, University of Wisconsin-Madison, 2005.

# Chapter 10

## Semantic Vector Spaces for Broadening Consideration of Consequences

Douglas Summers-Stay

Reasoning systems with too simple a model of the world and human intent are unable to consider potential negative side effects of their actions and modify their plans to avoid them (e.g., avoiding potential errors). However, hand-encoding the enormous and subtle body of facts that constitutes common sense into a knowledge base has proved too difficult despite decades of work. Distributed semantic vector spaces learned from large text corpora, on the other hand, can learn representations that capture shades of meaning of common-sense concepts and perform analogical and associational reasoning in ways that knowledge bases are too rigid to perform, by encoding concepts and the relations between them as geometric structures. These have, however, the disadvantage of being unreliable, poorly understood, and biased in their view of the world by the source material. This chapter will discuss how these approaches may be brought together in a way that combines the best properties of each for understanding the world and human intentions in a richer way.

### 10.1 Designing for Safety

Failure Mode and Effects Analysis documents are used for ensuring safety in complex systems such as automotive design. Engineers painstakingly analyze each subsystem for its probability of failure and build in layers of redundancy depending on the seriousness of system failure. Fail-safes (systems that, when they fail, do so in a way that leaves them safer), layers of redundancy, and hazard and risk analysis, are all tools used to reduce the probability of injury or death to a reasonably low level.

---

D. Summers-Stay (✉)  
US RDECOM, Artificial Intelligence Researcher, Army Research Laboratory,  
Adelphi, MD, USA  
e-mail: [douglas.a.summers-stay.civ@mail.mil](mailto:douglas.a.summers-stay.civ@mail.mil)

Typical machinery makes use of a very simple model of the world. A grocery store conveyor belt, for example, has two states and one binary sensor controlling which state it is in. A safety analysis would consider a richer model of the conveyor belt as a collection of moving parts, any one of which could break, and the much larger set of states that could put the system in, as well as potential consequences of such a failure. The complexity of autonomous systems makes such analysis more difficult. As the system becomes more autonomous, the number of potential actions the system can take and the variety of situations it can find itself in grows very quickly. In addition, useful AI systems must learn and change over time: understanding means incorporating newly acquired facts about the world into the already existing body of knowledge. A-priori consideration of every possible situation becomes impossible. It seems the only solution is to automate the safety analysis itself: we must design the system to perform a safety analysis on its own actions.

Doing this would require the autonomous system to have a rich model of the entire environment it will be interacting with—not just a simplified model that allows it to perform its normal tasks, but a model that takes into account the wider environment so that it understands what its tasks are for, what consequences its actions will have, and which consequences are to be avoided.

Creating such a system to reduce human error would be very difficult, difficult to the point that it has never been seriously attempted. Causal reasoning about physical systems can be performed for limited situations by creating detailed physical simulations, such as finite element analysis for stress analysis or nuclear weapons testing, but such methods are far too computationally intensive to be used for making quick decisions about everyday situations. A more promising approach involves qualitative reasoning about physical systems. In 1985, the dramatically named “Naïve Physics Manifesto” (Hayes 1978) laid out a program for enabling AI to answer questions about real world situations, with some initial success: “figuring out that a boiler can blow up, that an oscillator with friction will eventually stop, and how to say that you can pull with a string, but not push with it.” Hayes’ plan involved entering knowledge about the causal relationships of physical systems into a first order logical system (a knowledge base), and deducing answers to such questions. This approach ran into the common problem of expert systems: brittleness and incompleteness (Lenat 1985). Unless a query was designed carefully by a researcher with intimate understanding of how the knowledge base was constructed and what information it contained, some missing assumption would break the chain of reasoning and no answer would be returned.

There has been substantial work (e.g., Dash 2013) on A.I. planning and the creation of subgoals. While this is important and necessary, as long as these subgoals make use of a simple, incomplete model of the world, they will be inherently unsafe outside of toy applications.

## 10.2 Understanding Intent

Reasoning about physical processes that may lead to accidents, while a huge effort in itself, is only one part of the problem. Without understanding exactly the goal to be accomplished, the AI system may plan for a goal in a way that contradicts other implicit goals, in ways that may prove dangerous.

Amodei (2016) pointed out two mechanisms that can lead to accidents when an objective function is specified. “Negative side effects” can occur because of an insufficient model of chains of causal relations, leading to unanticipated negative consequences. “Reward hacking,” however, occurs when the objective function is technically satisfied, but in a way that contradicts unspoken goals.

“The Sorcerer’s Apprentice” is an old story, probably most familiar from Disney’s version, but originating in the second century A.D. The ancient Greeks also told stories of King Midas turning his daughter to gold or Tithonus, who Zeus grants immortality but not eternal youth. There are similar stories about genies from the Arabian nights, as well as fairy tales about wish-granting fishes, stories of golems from Jewish sources, and stories of deals with Old Nick from frontier America. All these fit the same pattern. In each version of the story, the entity granting the petition has the ability to help humans achieve their goals, but although the petitioner’s goal is technically satisfied, it happens in a way that contradicts real, deeper desires. In discussing issues of A.I. safety, Stuart Russell points out that we have a similar situation: “The primary concern is... simply the ability to make high-quality decisions. Here, quality refers to the expected outcome utility of actions taken, where the utility function is, presumably, specified by the human designer....The utility function may not be perfectly aligned with the values of the human race, which are (at best) very difficult to pin down.” (Russell 2014)

Dietterich (2015) wrote, “Suppose we tell a self-driving car to ‘get us to the airport as quickly as possible!’ Would the autonomous driving system put the pedal to the metal and drive at 125 mph, putting pedestrians and other drivers at risk? ... [T]hese examples refer to cases where humans have failed to correctly instruct the AI system on how it should behave. This is not a new problem. An important aspect of any AI system that interacts with people is that it must reason about what people intend rather than carrying out commands literally. An AI system must analyze and understand whether the behavior that a human is requesting is likely to be judged as “normal” or “reasonable” by most people.”

This is a familiar experience to every programmer. Although programming languages allow us to specify exactly what we want the computer to do, we often end up writing buggy programs that don’t do what we actually want. Autonomous systems are designed to act with less direct, more natural instruction. How do we make a system that will carry out what we want when we ask it to do something? It is impossible unless the system has knowledge of what kinds of things we want and what our words mean.

The problem of A.I. safety, then, is inescapably a version of the same problem of automating understanding that lies at the core of natural language understanding, common sense reasoning, mental modeling, creativity, and many other efforts that have been challenges for A.I. research since its inception. This can be looked at in a

positive, way, though. The same research that is required to make A.I. effective at real world tasks will also be advancing the ability to carry out those tasks safely, without undesirable side effects.

### 10.3 Expressing Intent

Natural languages, unlike programming languages, are imprecise and underspecified. In every uttered sentence, there is a large body of assumed background context, shared knowledge that can remain unsaid. Part of this is innate: all humans have certain shared goals even from infancy, such as air, water, food, and safety from physical and emotional harm—Maslow’s “hierarchy of needs.” Part of this is learned over a lifetime, the cultural body of knowledge such as property rights, social conventions, sarcasm, humor, and so forth.

When a command is expressed in natural language, the command cannot contain all of the limitations and context necessary to carry out the command in a way that matches the intent of the human giving the command. Such precision in language is inherently *un*-natural. If not expressed in the command itself, such values must already be included in the background knowledge brought to bear as the A.I. forms a plan to carry out the command.

One well-established attempt at pinning down some part of human values is the legal system. The legal system attempts to encode some human standard of what is acceptable behavior of an agent interacting with society and the world in very precise language, at least as far as human-readable documents go. The written law, however, is insufficient to decide cases. When cases are actually brought to trial, human lawyers are needed. The lawyers’ role is to search through similar cases which have already been decided, in order to find the nearest analogies with cases in precedent they can which result in a ruling favorable to their side. With lawyers performing this role on each side of the case, a human judge or jury decides which they find to be most similar.

Human judges and lawyers are needed because the law is necessarily insufficiently precise to cover every possible case. In this way it is very similar to hand-created knowledge bases. Attempts to encode knowledge in a system capable only of deductive reasoning were invariably very limited in their usefulness, because they lacked this ability to extend reasoning to new cases by analogy (Speer 2008).

The ability to find analogies is essential to understanding physical systems and human intent. Suppose a boy hits a baseball through a window of the house. A mother provides negative reinforcement, saying “don’t do that again.” But what does she mean by “do that”? It could mean:

- “never hit a ball with a bat”
- “don’t play near the house”
- “don’t hit a ball towards this particular window”
- “don’t move your arms”

And so forth. In order to understand what she means by “do that,” the boy may apply the golden rule: his unconscious reasoning is something like, “I would be angry if someone broke one of my possessions, so she must be angry because I broke one of her possessions.” The boy will recognize that throwing a stick inside the house where it might break a lamp is an analogous situation to be avoided in the relevant sense of “causing an object to move unpredictably where it has the chance to damage someone’s fragile property.” But the ability to pull out this particular meaning of “do that” over any of the others depends on a lifetime of experience and an internal set of desires that corresponds, more or less, with the mother’s.

Is building this kind of “common sense” into an AI system really necessary for it to behave safely? It is such a difficult problem that any way around it seems preferable. In Amodei (2016), several methods were proposed for increasing AI safety that don’t explicitly include such a design. For example, they suggested avoiding side effects, or situations which might potentially have side effects. After exploring this idea for a little while, however, they pointed out situations where such an approach would fail without some notion of the user’s goals and the form that consequences would take. There’s no free lunch: (Amodei 2016, p. 6) “Avoiding side effects can be seen as a proxy for the thing we really care about: avoiding negative externalities. If everyone likes a side effect, there’s no need to avoid it. What we’d really like to do is understand all the other agents (including humans) and make sure our actions don’t harm their interests...However we are still a long way away from practical systems that can build a rich enough model to avoid undesired side effects in a general sense.” The only solution to this problem is the hard one: biting the bullet and building a rich enough model to avoid undesired side effects.

There are two main problems with encoding such common sense background knowledge in a way that an autonomous system can make use of. The first problem is an architectural issue: The meanings of concepts are rich and nuanced. What kind of data structure can allow for such diverse phenomena as being reminded by similar ideas, completing analogies, recognizing objects by their attributes, and recognizing a class by a single example of that class, and still support deductive reasoning?

The second problem is this: once we have an architecture capable of storing concepts and reasoning about them in deductive, inductive, and analogical ways, how can we populate it with the vast amount of common-sense knowledge we all share?

## 10.4 Problem 1: An Encoding for Concepts

Douglas Hofstadter has been writing about the nature of concepts and analogies since the 1980s, pointing out a distinction between how symbolic information is stored in precise logical forms in a knowledge base, and how concepts are held in the mind. “The property of being a concept is a property of connectivity, a quality that comes from being embedded in a certain kind of complicated network.” (Hofstadter 1985, p. 528) In an object-oriented programming language or a



knowledge base, we can represent an object such as a fire-extinguisher with a few facts defining its function as needed in the program. To really count as a *concept*, though, requires much more than that. The concept of a fire-extinguisher includes something of its shape and size, the material it's made from, its appearance, the uses it is put to, how to operate it, where one can be found, a rough idea of how much it costs, what it resembles, and many other such properties. Each of those properties, in turn, must be concepts, with the same richness of internal structure. Concepts that define a class have shades of membership. A bucket of sand might be considered a fire-extinguisher under certain ill-specified conditions. A fire-extinguisher that has not been recharged also has a shaded inclusion in the category.

Concepts have connections of varying strength with many other concepts. "Each new concept depends on a number of previously existing concepts. But each of those concepts depended, in its turn, on previous and more primitive concepts... This buildup of concepts over time does not in any way establish a strict and rigid hierarchy. The dependencies are blurry and shaded rather than precise, and there is no strict sense of higher and lower... since dependencies can be reciprocal. New concepts transform the concepts that existed prior to them, and that enabled them to come into being; in this way, newer concepts are incorporated inside their "parents" as well as the reverse." (Hofstadter 2013, p. 54) To act as a concept, then, requires that the information be stored in a way that admits degrees of similarity, and definitions that are reciprocal, rather than built up from axioms like the definitions of mathematical structures.

Our understanding of concepts is evoked by similar concepts, and the way we think about concepts is largely analogical in nature. "The ability to perceive similarities and analogies, he argues, is one of the most fundamental aspects of human cognition. It is crucial for recognition, classification, and learning, and it plays an important role in scientific discovery and creativity." (Vosniadou 1989, p. 1). Whatever representation of concepts we come up with, it must be able to support reasoning by analogy, and such analogies must be flexible enough to admit ambiguity and imperfect matches.

In early A.I. research, concepts were represented using knowledge bases: as nodes in a relational graph in a database with the capacity for deductive reasoning. The graph expressed first order relations as connections between nodes. This stored symbolic information, but failed to capture the subsymbolic information that is inherent in human concepts. A key problem in storing information this way is that any mismatch between the arrangement of concepts in the knowledge base and the form of a query will cause the query process to fail completely, returning no results at all. For example, the knowledge base may include the fact that gasoline may catch fire:

**causes (gasoline, fire\_hazard)**

but a query asking

**has\_tendency (gasoline, X?)**

Will return no results unless the knowledge base also has rules defining how **causes** and **fire\_hazard** are connected to **has\_tendency** and **burn\_rapidly**.

This isn't just a problem with insufficient rules in the knowledge base, however. Concepts in the human brain seem to be stored in a way that makes them fundamentally different from entries in a knowledge base. We can be reminded of concepts by resemblance in sounds between words, similar parts, or properties between concepts, a similar environment in which the concepts are encountered, and many other ways. Instead of being a discrete graph where each concept in the graph is assigned or not assigned to a particular relation, there are gradations of inclusion by which a pair of terms fits the relation more or less precisely. Many of the relations we can find in our memory seem to be an implicit result of the way the concept is stored, rather than an explicitly learned link.

(Kanerva 1988, p. 2) wrote “although we normally ignore such links, they are there, and they can tell us something about the mathematical space for memory items. Translated into a requirement for the model, memory items should be arranged in such a way that most items are unrelated to each other but most pairs of items can be linked by just one or two intermediate items. This requirement affects the choice of mathematical space for memory items, also called the semantic space.”

## 10.5 Semantic Vector Spaces

What Kanerva suggested was to encode the concepts as vectors in a high-dimensional vector space. High-dimensional vector spaces have some unintuitive properties that make them ideal for representing concepts. One of the most important of these is that between two arbitrary vectors in this space, we can find a vector very close to both but not unusually close to any unrelated vectors. The vector spaces Kanerva worked with were  $n$ -dimensional binary vector spaces, where each element of the vector is 1 or 0, written as  $\{0,1\}^n$ . “The distance between two points of  $\{0,1\}^n$  represents the similarity of two memory items—an association based on form. It is the number of places in which the two patterns differ, so that the closer the points, the more similar the items. Almost all of the space is nearly indifferent to (or about  $n/2$  bits away from) any given point, whereas two points  $n/4$  bits apart are very close together in the sense that an exceedingly small portion of the space lies within  $n/4$  bits of a point. This is intuitively appealing in that any particular concept in our heads is unrelated to most other concepts, but any two unrelated concepts can be linked by a third that is closely related to both.” (Kanerva 1988, p. 25)

Starting with a few primitives, a high-dimensional vector space can build them up to represent more complex ideas. For example, the vector representing **ice** can be located near the sum of the vector for **cold** and the vector for **water**. This new **ice** vector will be similar to both **cold** and **water** and nothing else, except any other terms we may have also formed that include **cold** and **water** as components, such as **snow**. To count as a concept, the vectors would need to be built from many more components representing every aspect of ice that might be of interest, but with a high-enough dimensional vector, many such components can be included. Details

about how many such components can be included in a single vector can be found in Kanerva (1988) or Hawkins (2007).

Such a semantic vector space is capable of representing not just ideas, but also relations between them. For example, suppose we wanted to represent the fact that snow causes icy roads:

**causes\_road\_condition(snow, icy\_roads)**

We define locations in the vector space representing the concepts **precipitation**, **frozen**, and **road**. To represent **snow**, we take the sum of **precipitation** and **frozen** and to represent **icy\_roads** we sum **frozen** and **road**. The relation **causes\_road\_condition** is then the vector which subtracts out **precipitation** and adds in **road** to a concept: the vector (**road - precipitation**). This same relation vector, when added to **rain** will lead us to the vector for **wet\_roads**.

In this way, one-to-one relations between concepts can be defined as displacement vectors between the vectors for those concepts. Concepts can be built up from the simplest attributes we wish to define. In a real system, we would, of course want a more refined concept for **icy\_roads** that included the fact that they are slippery, that they sometimes have a reflective appearance, and so forth. The problem of how to get all of the information that needs to be encoded in a concept will be covered in Sect. 10.6. All we are doing here is showing that the vector space representation is capable of holding such information about concepts and their relations.

Following chains of deductive reasoning would be simple in such a vector space. Suppose the space encodes the facts that

**has\_location (finger , cutting\_board)**

and

**uses (cutting\_board, knife)**

then we can conclude that **could\_be\_affected\_by (finger, knife)** using a rule stating that **has\_location(X,Y) ^ uses (Y, Z)** implies **could\_be\_affected\_by (X,Z)**. (Neelakantan (2015) explores such chained reasoning in vector spaces.) In this case, the vector representing **could\_be\_affected\_by** can be found by simply adding the vectors from **X** to **Y**, and from **Y** to **Z** to find the vector from **X** to **Z**.

It is also possible to perform analogical reasoning in such a vector space. Suppose we are given the following analogy to solve: **bear:hiker::shark:X**. If the concepts **bear**, **hiker**, and **shark** are already in the vector space, they are composed of simpler terms. Suppose these simpler terms happen to be **woods**, **sea**, **predator** and **tourist**. Then substituting in the simpler component terms, we have the simpler analogy **woods + predator : woods + tourist :: sea + predator : X**. The relation between the first two terms can be found by subtracting **predator** from **bear**, (leaving **woods**) and adding **tourist** to the result. We then apply that relation to **shark**, to get **sea + tourist**, which is close to the vector for **snorkeler**. The vector arithmetic simplifies to  $\mathbf{D} = \mathbf{B} + \mathbf{C} - \mathbf{A}$ , when trying to solve  $\mathbf{A}:\mathbf{B}::\mathbf{C}:\mathbf{D}$ . While these are too high-level terms to actually be used as primitive concepts, they serve to demonstrate the arithmetic.

To the extent that the fundamental concept vectors are plentiful enough, this vector space also serves as an associational memory, in the sense that summing up a few related terms is enough to bring to mind a term associated with them. For example, adding **France** + **city** + **fashion** gives a vector close to Paris, since the components of **France**, **city**, and **fashion** all added together are similar to the components of **Paris**, plus some leftover that can be treated as noise.

Thus, a memory encoded as a high dimensional vector space is capable of supporting deductive, analogical, and associational reasoning. As further support for the practicality of such an approach, it is interesting to note that such a memory is fairly biologically plausible. As a toy model, each component of the vector could be considered to be a neuron which is activated to some degree, and reasoning would consist of bringing to mind the various concepts in such a way that the end result of the reasoning is the application of the relevant vector arithmetic on those concepts. Hinton (1984) outlined how distributed representations were more biologically plausible than the local representations used in a knowledge base. Brain-imaging studies have likewise suggested that concepts are represented in the brain as distributed networks of neural activation (Rissman 2012; Blouw 2005). In particular, the analogical-reasoning capability of a semantic vector space can be understood as an example of the relational priming model of analog-making outlined in Leech (2008). There is also evidence that object categories are encoded as a continuous semantic space across the surface of the brain (Huth 2012). The brain's slow operating speed and massive parallelism also hint that whatever operations are being performed must be very short, simple programs operating on large vectors.

All this is of little use, however, unless we can find a way to input all the information about the concepts. Indeed, unless we can find some way of automating the process of populating the vector space, we are not getting much more out from it than we have painstakingly entered by hand. Once we know that two concepts share certain qualities or relations, finding associations and analogies is not difficult. It is recognizing those shared properties in the first place that is the harder problem.

This is where the problem stood for some time after the development of the theory of representing concepts as high-dimensional vectors. At that time, finding a way to automatically populate the vector space was not a practical possibility. The development of such a method would come about from attempts to find semantically similar documents.

## 10.6 Problem 2: Distributional Semantic Vector Spaces

A distributional semantic vector space assigns a vector to each word such that words found in similar contexts have similar vectors. Incredibly, this is enough to create (in an approximate, noisy way) the kind of vector space described in the previous section, including some very subtle and difficult to express attributes of concepts that can be used for associational and analogical reasoning.

The simplest distributional vectors represent the meaning of a document by counting up the frequency of occurrence of each word in that document. The vector is the size of the entire English vocabulary  $v$ , with zeros for most words, which never occur in the document, and the occurrence count for the words which do occur. Such vectors are impractically large, and tend to be very noisy in terms of similarity between documents expressing similar ideas, because one author may prefer certain terms to express an idea, while another author would use a different subset of terms to express the same idea.

If we consider a few words to either side of a given word any time it occurs to be its “document”, we can create a  $v * v$  matrix in this way that encodes each word by its context. Using support vector decomposition (SVD), we can reduce this matrix to a more reasonable size (say,  $300 \times 300$  instead of  $v \times v$ ) and at the same time remove much of the noise, so that vectors encoding similar terms end up with similar vectors. Later techniques such as word2vec optimized certain technical parameters and improved the time and memory performance of deriving such vectors (actually performing SVD on a  $v \times v$  matrix requires an impractical amount of memory) to the point where enormous corpora could be practically handled, but the core idea behind the vectors is the same (Deerwester 1988). Since this process assigns vectors with similar meanings similar vectors, all the nice properties listed above—the ability to find analogies, to recall a concept by its associations, to break a concept down by its attributes, to encode a one-to-one relation as a vector, and to compose these vectors to follow chains of reasoning—are *automatically* properties of the distributional vectors.

Consider the analogy

**seat\_belt : car :: life\_preserver : X.**

**seat\_belt** will be found in contexts discussing roads, and in contexts discussing accidents and safety. **Life\_preserver** will also be found in contexts discussing accidents and safety, but instead of being found in contexts having to do with roads, it will be found much more often with words like **sea** and **ocean**. **Car** will be found in contexts having to do with vehicles, as well as contexts having to do with roads. **Ship** and **boat** will also be found in vehicle contexts, near verbs like **travel** which it shares with **car**, for example, but also in the context of **ocean**. Since a word found in two concepts is found near the average or sum of those two concepts, the following will be approximately true:

**road + safety : road + vehicle :: ocean + safety : ocean + vehicle**

This is the same kind of analogy we saw in the constructed example of sharks and bears above. This conceptual arithmetic would not be exact. Life preservers might also share some of the context of vests, while seat belts might share a context of belts, which isn’t captured in the arithmetic above. But as long as there are no nearer terms, such differences can be treated as noise, for which the vectors are surprisingly robust. At some of these tasks the vectors perform very well: given a four-term analogy problem from the SAT with a multiple choice answer, such vectors can reach (Turney 2006) human performance. In other words, distributional

semantic vectors automatically encode a great deal of common-sense about concepts in their structure.

Their ability to represent concepts as a whole makes them even able to handle some tasks that would normally be considered to require some creativity. For example, we performed the following experiment. All pairs of adjectives and nouns starting with the letter “a”, and pairs of rhyming words were generated based on existing word lists. The vectors representing both words in a pair were averaged, and then these were searched to find the nearest match to the vector for a search term. Here are synonymous alliterative phrases it came up with:

- robot: anthropomorphic automaton
- songbird: arboreal artist
- textbook: authoritative algebra
- birdhouse: architectural aviary
- chemistry: academic alchemy
- tin: antique aluminum
- bronze: archaeological alloy
- neon: amber ambiance
- divide: antagonistic arithmetic

and synonymous rhyming phrases:

- cowboy: colorado desperado
- friar: yeast priest, barbarian seminarian
- pillow: head bed
- trampoline: elastic gymnastic
- rocking chair: knitter sitter
- novel: fiction depiction
- orca: dalmation cetacean
- flower: bloom perfume, frilly lily
- Clinton: she nominee

The ability to generate such paraphrases demonstrates that some aspects of the meaning of the terms have been captured by the vectors. Other aspects of the meaning are missing, however. Learning solely from written texts creates some serious gaps in the knowledge implicitly contained in the set of vectors. Vectors nearby to the vector for **safe** include

- synonyms, near synonyms, hypernyms, sister terms, and hyponyms (**secure, healthy, reliable, comfortable, stable, protected, sane, adequate, prudent, reliable**)
- other forms of the word (**safer, safest, safely, unsafe, safeness, Safe, safety**)
- antonyms (**unsafe, dangerous, hazardous**)

This is not ideal. What we would like for the purposes of reasoning is a vector whose near terms are all synonyms of **safe**, with other related terms a little farther away. Perhaps most important is to draw a strong distinction in meaning between terms and their antonyms. One way to do this is to start with a vector space learned

from a large text corpus, but then represent the concepts we are interested in as a sum or average of the synonymous terms that all mean the same concept. We could represent the concept of **safe**, for example, by a vector that is a weighted sum of the terms we would like nearby, with negative weights on the terms from which we would like it to be farther away. This separates the *concept* vectors from the vectors for the *terms* themselves, though the two will still be very close. The sets of synonyms and antonyms for many English terms are already available from Wordnet, so constructing such vectors is not difficult (Rothe 2015).

While a semantic vector space trained on a large, diverse text corpus is successful on some analogy tasks, it fails badly on others. For example, the following is an easy analogy for humans:

**blueberry : blue jay :: strawberry : cardinal**

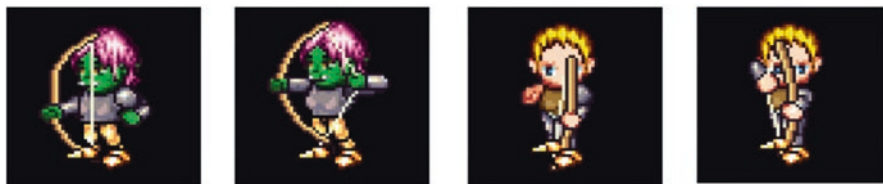
If we try this in word2vec, instead of **cardinal** we get

**blue jays red bellied woodpecker grackle ovenbird downy woodpecker indigo bunting tufted titmouse Carolina wren chickadee nuthatch rose breasted grackle bluejays raccoon spruce grouse robin**

Most of the results are species of birds, but there does not seem to be any tendency for the birds named to be *red*. We can easily find clusters of fruit and songbirds in the vector space. But there is no reason for red things to appear in the same contexts in newspaper articles—no texts that discuss firetrucks, strawberries, and cardinals in the same way. Their color just is not very relevant to the way newspapers talk about those things. (If we trained the system on books for toddlers, that might be different.) Because red things are not clustered together, the analogy fails. The color of objects is not typically one of the facts about them that the vector encodes because it is too elementary a fact to be mentioned in the corpus.

In the last few years, computer vision researchers (Sadeghi 2015; Reed 2015; Upchurch 2016) have built deep-learning neural network architectures whose weights, treated as a high-dimensional vector, organize visual representations in the same way as distributional semantic vector spaces organize the textual world. These systems are able to form “visual analogies” that could potentially solve the analogy above, not in verbal form but with actual pictures of birds and fruit. Such systems are not trained on context at all (each image is learned in isolation) and yet they are able to build up a similar representation. This is because context serves mainly as a way of discovering similarity. If we can learn similarity directly, through the representations formed in a deep neural network, similar things (birds, fruits, items of particular colors or shapes) will be represented by similar vectors and form clusters, enabling the analogical arithmetic above to go through (see Fig. 10.1).

Any method of creating high-dimensional vectors in order to classify data will have these properties of analogy-forming, relation-representation, and associational structure if it is successful in putting vectors nearby in the explicit and implicit categories we are interested in.



**Fig. 10.1** Example of deep visual analogy from Reed (2015). Based on the first three images and a learned image manifold, their deep learning system is able to infer the fourth image by analogy with the other three

## 10.7 What Needs to be Done

The facts presented above give reasonable confidence that researching ways of building better semantic vector spaces and reasoning over them could prove successful. The area has seen a few high-profile advocates in the last few years. Hinton, now at Google, has been a proponent of encoding the meaning of entire sentences in a single vector in what are called “skip-thought” vectors (Ba 2016). Hawkins’s (2007) memory-prediction framework also represents concepts as high-dimensional vectors, though his work concentrates on streams of information rather than single documents or images. Making such a system able to reason effectively, though, is still a challenging problem that will take many researchers some time to complete. Such a program would involve many different areas:

### 10.7.1 Learning More Complex Relations

The method for following chains of deductive reasoning described in Sect. 10.5 is really only effective on one-to-one relations, where each concept in the input is paired with exactly one concept in the output. These include a wide variety of relations, everything from **scientist\_studies\_field (astronomer, astronomy)** to **animal\_makes\_noise(cow, moo)**. But most relations are not so simple. Consider the relation **store\_carries\_product**. Every store carries many products, and every product is carried by many stores. Associational reasoning is still helpful here—if a store carries peanut butter, there’s a good chance that it also carries jelly—but it is not as simple as defining one vector defining the relation between Walmart and its set of products, subtracting out Walmart and adding Costco, and expecting that the set of products found will be very accurate. Using larger vectors, it is possible to define a vector representing a set of semantically similar things, especially if we also give information about where to make the cutoff between things in the set and things not in the set. However, some relations are simply going to be too complex to



handle this way, and will need to be encoded by something more complex than a single vector, no matter how we arrange the rest of the concepts in the vector space. Such relations may need to be learned by a neural network, ideally in such a way that similar relations will be able to share information to develop into similar representations in the neural network.

There has been some exploration of how various relations are encoded into a distributional semantic vector space. Rei (2014) for example, explores how hyponyms swarm around a term. The possibility has also been explored of reshaping a vector space according to verified facts in Faruqui (2014). One problem with doing this is that other relations, not explicitly included in such reshaping, may be distorted and so no longer have the analogical properties they had in the original vector space. Wang (2014) also explores putting a knowledge graph into a vector space.

### ***10.7.2 Distributional Semantics***

Research into distributional semantic vector spaces has exploded in the last few years. Some interesting areas include choosing dependency based-word embeddings (Levy 2014), which modify the context window based on the sentence parse tree. GloVe and word2vec are among the most popular distributional semantic vector spaces at the moment due to their capacity for training on large corpora.

### ***10.7.3 Semantics from Images, Video, and Other Data Streams***

Image vectors are derived from the weights of neural networks trained on images. Image vectors capable of capturing more than just class membership, but also encoding color, texture, pose, shape, and other information are still in their infancy. Some interesting efforts include Sadeghi (2015), Reed (2015), and Upchurch (2016). Similar vectors could potentially be derived from 3D sensors, whether depth-based or tactile, to provide another dimension of context.

### ***10.7.4 Combining Two Vector Spaces to Better Capture the Knowledge Learned from Each***

Simply concatenating the vectors from two spaces is one way to combine them, but there must be ways of knowing which kinds of facts have been captured better by one space than another, and giving more weight to that space. This seems like a task best handled by a neural network.

### ***10.7.5 Encoding the Meaning of Natural Language Phrases and Sentences as Vectors***

Skip-thought vectors (Kiros 2015) are one attempt at encoding sentences and phrases directly as vectors. AnalogySpace, built from the large knowledge base ConceptNet (Speer 2008), is another such example. If we keep to the idea of representing only word-like concepts as vectors so as to preserve the relational properties, a different approach would seem to be required. A sentence can be considered as a series of asserted relations between the terms in the sentence. If we can successfully parse the sentence into these relations, and modify our representations of the terms and relations appropriately, the system will have incorporated the knowledge in the sentence. Such semantic parsing is not completely reliable yet, but it has been improving. It would be best if such a semantic parsing system could be learned together with the semantic vector space so that more subtle conceptual relations could potentially be captured.

### ***10.7.6 Modifying a Semantic Vector Space as New Information Is Learned Without Destroying Already Existing Structure***

Faruqui (2014) showed that the vectors representing terms in a distributional semantic vector space could be modified to better capture certain known relations, but that doing so without regard to existing structure reduced the ability of the system to form analogies of other relations which were not explicitly optimized for. Unless these unplanned-for relations can be preserved somehow, modifying the vector space to incorporate known facts will always run the risk of destroying them.

### ***10.7.7 Performing Reasoning Within Vector Spaces***

When a chain of reasoning consists solely of one-to-one relations that are accurately captured by displacement vectors, deductive reasoning to follow the chain is a single-step process. This gives it a huge advantage over reasoning within a knowledge base, where a tree of possible relations must be explored to find a path between the terms in the relation to be proved. But this only applies to that limited set of one-to-one relations. For reasoning steps that involve more complex combinations of and/or operations on relations (known as Horn rules) or that involve higher-order relations, other techniques must be developed. This has not yet been extensively explored, although Neelakantan (2015) and Lin (2015) have made a beginning at it.

### ***10.7.8 Ways of Discovering and Representing Knowledge About Physical Consequences***

Most of the concepts and relations that need to be represented in order to reason about consequences of actions will not come from either textual sources or from still images, but only through the experiences of an agent interacting with the world in a safe play environment. These experiences will need to incorporate spatial and temporal aspects which are difficult to handle in standard reasoning systems and vector-based ones alike. The kind of system outlined here at least has the capacity, though, to represent and reason about all the relevant concepts in a real-world situation which could involve all kinds of unanticipated objects and actors. The same can't be said of approaches using either a hand-built knowledge base or a learned neural-network representation that does not carry out reasoning processes.

## **10.8 Conclusion**

Semantic vector spaces provide a way of capturing conceptual knowledge and reasoning about it efficiently and flexibly. They allow for analogical and associational reasoning in a way that is completely impractical for purely symbolic approaches to knowledge representation. Such conceptual representations are necessary for interacting with the diversity of situations that arise in the real world and doing so safely. There is still a lot of territory to explore in how best to create and make use of such subsymbolic representations, but research from neuroscience, machine learning, linguistic and knowledge representation communities all seem to be converging on similar methods, though what a finished system will look like is still murky. Any short-cuts to A.I. safety that are unable to reason about physical processes and human goals in order to reduce accidents and human error will behave in unexpected ways when introduced to new environments.

## **References**

- Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.
- Ba, J. L., Kiros, J. R., and Hinton, G. E. (2016). Layer normalization. arXiv preprint arXiv:1607.06450.
- Blouw, P and Eliasmith, C. (2005) A neurally plausible encoding of word order information into a semantic vector space. *35th Annual conference of the cognitive science society* Vol. 1910.
- Dash, D., Voortman, M., and De Jongh, M. (2013, August). Sequences of Mechanisms for Causal Reasoning in Artificial Intelligence. In IJCAI
- Deerwester, S., et al, Improving Information Retrieval with Latent Semantic Indexing, Proceedings of the 51st Annual Meeting of the American Society for Information Science 25, 1988, pp. 36–40.

- Dietterich, T. G., and Horvitz, E. J. (2015). Rise of concerns about AI: reflections and directions. *Communications of the ACM*, 58(10), 38-40.
- Faruqui, M., Dodge, J., Jauhar, S. K., Dyer, C., Hovy, E., and Smith, N. A. (2014). Retrofitting word vectors to semantic lexicons. *arXiv preprint arXiv:1411.4166*.
- Hawkins, J., and Blakeslee, S. (2007). *On intelligence*. Macmillan.
- Hayes, P. J. (1978). The naive physics manifesto. Institut pour les études sémantiques et cognitives/ Université de Genève.
- Hinton, G. E. (1984). Distributed representations.
- Hofstadter, D. (1985). *Metamagical themas: Questing for the essence of mind and pattern*. Basic books.
- Hofstadter, D., and Sander, E. *Surfaces and Essences*. Basic Books, 2013.
- Huth, A. G., Nishimoto, S., Vu, A. T., and Gallant, J. L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6), 1210-1224.
- Kanerva, P. (1988). Sparse distributed memory. MIT press.
- Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Skip-thought vectors. In *Advances in neural information processing systems* (pp. 3294-3302).
- Leech, R., Mareschal, D., and Cooper, R. P. (2008). Analogy as relational priming: A developmental and computational perspective on the origins of a complex cognitive skill. *Behavioral and Brain Sciences*, 31(04), 357-378.
- Lenat, D. B., Prakash, M., & Shepherd, M. (1985). CYC: Using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks. *AI magazine*, 6(4), 65.
- Levy, O., and Goldberg, Y. (2014). Dependency-Based Word Embeddings. In *ACL(2)* (pp. 302-308).
- Lin, Y., Liu, Z., Luan, H., Sun, M., Rao, S., and Liu, S. (2015). Modeling relation paths for representation learning of knowledge bases. *arXiv preprint arXiv:1506.00379*.
- Neelakantan, A., Roth, B., and Mc-Callum, A. (2015, March). Compositional vector space models for knowledge base inference. In *2015 AAAI Spring Symposium Series*.
- Reed, S. E., Zhang, Y., Zhang, Y., and Lee, H. (2015). Deep visual analogy-making. In *Advances in Neural Information Processing Systems* (pp. 1252-1260).
- Rei, M., and Briscoe, T. (2014, June). Looking for Hyponyms in Vector Space. In *CoNLL* (pp. 68-77).
- Rissman, J., and Wagner, A. D. (2012). Distributed representations in memory: insights from functional brain imaging. *Annual review of psychology*, 63, 101.
- Rothe, S., and Schütze, H. (2015). Autoextend: Extending word embeddings to embeddings for synsets and lexemes. *arXiv preprint arXiv:1507.01127*.
- Russell, S. (2014, November 14). Of Myths And Moonshine. Retrieved from [edge.org/conversation/jaron\\_lanier-the-myth-of-ai](http://edge.org/conversation/jaron_lanier-the-myth-of-ai)
- Sadeghi, F., Zitnick, C. L., and Farhadi, A. (2015). Visalogy: Answering visual analogy questions. In *Advances in Neural Information Processing Systems* (pp. 1882-1890).
- Speer, R., Havasi, C., and Lieberman, H. (2008, July). AnalogySpace: Reducing the Dimensionality of Common Sense Knowledge. In *AAAI* (Vol. 8, pp. 548-553).
- Turney, P. D. (2006). Similarity of semantic relations. *Computational Linguistics*, 32(3), 379-416.
- Upchurch, P., Snively, N., and Bala, K. (2016). From A to Z: Supervised Transfer of Style and Content Using Deep Neural Network Generators. *arXiv preprint arXiv:1603.02003*.
- Vosniadou, S., and Ortony, A. (1989). *Similarity and analogical reasoning*. Cambridge University Press.
- Wang, Z., Zhang, J., Feng, J., and Chen, Z. (2014, October). Knowledge Graph and Text Jointly Embedding. In *EMNLP* (pp. 1591-1601).

# Chapter 11

## On the Road to Autonomy: Evaluating and Optimizing Hybrid Team Dynamics

Chris Berka and Maja Stikic

### 11.1 Introduction

As the reliance upon teams and teamwork continues to grow, so does the need for reliable, unobtrusive, and real-time measures of team performance across a wide range of environments and domains, including industry, education, military, medical and sports settings. Artificial Intelligence (AI) systems could be utilized to learn team dynamics and even optimize team performance by implementing mitigation strategies when patterns are identified that could lead to failures in team performance. Potentially dangerous errors associated with less than optimal team interactions, for example in surgical or military teams, could be avoided with appropriate early interventions. The scientific investigation of team interactions is beginning to provide key insights into the collective performance of teams both in co-located and virtual environments. AI systems will likely prove useful in modeling team interactions and serving as integral components of future human-computer hybrid teams.

Traditional methods for assessing team performance include survey-based reporting and scoring based on expert observations. Surveys are typically based on self-reports that are inherently subjective, often qualitative, and are generally used offline—during team downtime, or after teams have been disbanded which introduces potential bias. Scoring based on expert observations is more reliable, but biases and discrepancies between scorers can still occur. Overall, performance based evaluation provides measures of team outcome, but they do not elucidate *why* the team performed optimally or poorly and may not be fully capable of capturing a more detailed insight into dynamics of team process.

To address these shortcomings, neuroscience methods show early promise towards a deeper understanding of the connections between individuals at a physiological

---

C. Berka (✉) • M. Stikic

Advanced Brain Monitoring Inc., 2237 Faraday Ave, Suite 100, Carlsbad, CA 92008, USA

e-mail: [chris@b-alert.com](mailto:chris@b-alert.com); [maja@b-alert.com](mailto:maja@b-alert.com)

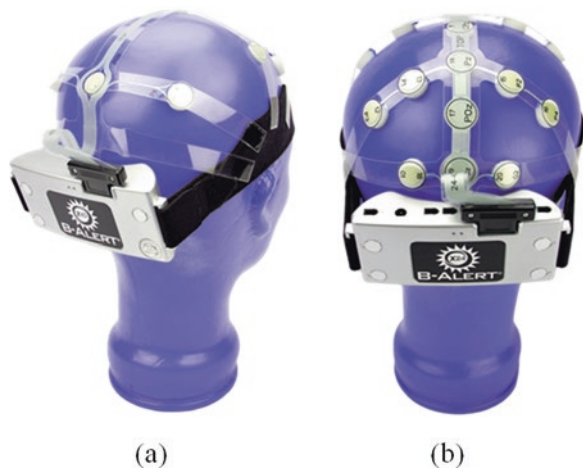
level, thereby enhancing analysis of team processes, and allowing for further insight into team performance. Historically, the use of physiological measures was limited by the obtrusive nature of instrumentation, but this has changed with the advent of miniaturized sensors and embedded platforms capable of supporting complex signal processing techniques online in real-world environments.

Neuroimaging techniques such as electroencephalography (EEG) could provide unobtrusive and objective insight into team members' brain activity patterns. Recent developments in EEG technology have enabled breakthroughs in EEG-based neurophysiologic metrics of team processes. Compared to other neuroimaging techniques, EEG-based assessment is inexpensive, and it provides continuous measures with high temporal resolution. Moreover, current EEG technologies can capture data during a team process without interruption, allowing for team members to be examined simultaneously in real-time. Lastly, EEG-based measures of team performance are quantifiable and the results could be compared against standard databases allowing for measuring test-retest reliability. Recent research suggests that neuroscience-based measures provide ecologically-valid assessment of team variables, such as engagement and leadership (e.g., Waldman et al. 2011b, 2013). In addition, EEG technology shows promise toward comparing brain activity patterns across teams to see which configuration(s) are associated with more cohesion and potential better team performance. These patterns have been organized into a framework, called Team NeuroDynamics (Stevens et al. 2009a), which provides real-time and objective insight into team cognition, with the potential for future applications in team optimization and adaptive training platforms.

Due to evolving task demands, there is a growing interest in studying specific aspects of collaborative teamwork such as engagement (Rich et al. 2010), workload (Funke et al. 2012), or leadership (Carson et al. 2007). In particular, performing a task as a team requires team members to mutually coordinate their actions, and this is the distinguishing factor between the performance of a team versus the same actions performed independently. A team does not equal the sum of its parts, and this introduces difficulties when trying to aggregate individual-level variables to derive team-level measures. For this reason, researchers have started to investigate time series of team members' neurophysiological metrics, with the goal of associating the dynamics of team members' cognitive and emotional states with the team process (e.g., Stevens et al. 2009b). The initial efforts were focused mostly on synchronized electrical activity between regions of the brain within an individual (Hannah et al. 2013; Waldman et al. 2011a). However, new approaches for identifying synchrony between physiological signals of individuals in dyads have started to emerge (e.g., McAssey et al. 2013). Furthermore, Astolfi et al. (2011) introduced an approach based on functional connectivity analysis of the team members.

In the rest of this chapter, we will consider the challenges of using EEG technology to examine team process. We will present the Advanced Brain Monitoring's (ABM, [www.advancedbrainmonitoring.com](http://www.advancedbrainmonitoring.com)) EEG teaming platform, and summarize a number of studies that have been conducted with this emerging technology. Next, we will consider the implications of using this methodology in team studies, and introduce open research questions that might be examined in the future with the proposed EEG approach.

**Fig. 11.1** EEG headsets:  
(a) X10, (b) X24



## 11.2 Teaming Platform

Before embarking on an in-depth discussion of the latest advancements in the analysis of team dynamics based on the EEG data, we will first introduce the B-Alert EEG hardware and software platform, as well as its ability to enable real-world applications in team settings by analyzing the neural patterns of human interactions synchronized across team members. The system is fully portable and easy to apply, so it allows for natural behavior during data acquisition. Furthermore, it is accurate, reliable, and cost-effective.

The platform supports two wireless B-Alert sensor headsets (X10 and X24) featured in Fig. 11.1. The X10 acquires EEG data from 9 scalp locations (Fz, F3, F4, Cz, C3, C4, POz, P3, and P4), while X24 densely covers 20 brain regions of interest (Fp1, Fp2, Fz, F3, F4, F7, F8, T3, T4, T5, T6, Cz, C3, C4, Pz, POz, P3, P4, O1, and O2). In addition, both systems can also acquire an electrocardiogram (ECG) with electrodes placed on the left and right clavicles. The signals are sampled at 256 Hz, filtered, and transferred wirelessly in real-time via Bluetooth link to a nearby computer or tablet, where the data is stored. The platform enables automatic detection and removal of a number of artifacts in the EEG data, such as spikes, amplifier saturations, or excursions. Furthermore, eye blinks and excessive muscle activity introduced by naturally occurring actions like head, jaw, or eye movement are identified and decontaminated by a proprietary algorithm based on a wavelet transformation (Berka et al. 2007). The B-Alert Live software supports a variety of EEG signal processing calculations, such as Power Spectral Densities (PSDs) and Event-Related Potentials (ERPs), as well as empirically-derived EEG-based metrics for quantification of engagement and workload (Berka et al. 2007; Johnson et al. 2011) that have been validated in a number of user studies (e.g., Berka et al. 2004; Westbrook et al. 2004; Behneman et al. 2012; Stevens et al. 2012).

The EEG-engagement metric is associated with processes involving information gathering and sustained attention or alertness to auditory and/or visual stimuli. It measures individual involvement in presented information. The metric relies upon PSD variables from the midline region of the brain (differential channels Fz-POz and Cz-POz) that were selected based upon their ability to discriminate four distinct classes of participants' alertness: High Engagement, Low Engagement, Distraction, and Sleep Onset. The metric is individualized using a benchmark session with ABM's Alertness & Memory Profiler (AMP) platform. The AMP integrates physiological and performance measures in an easy-to-administer platform designed for quantitative assessment of neurocognitive functions including alertness, attention, learning, and memory. The second validated metric, EEG-workload, is essential in processing gathered information, and comparing it to internal mental models. EEG-workload is associated with an increase in working memory load during problem solving, integration of information, and analytical reasoning. The workload measure has been developed on a large dataset of EEG recordings, during which participants performed two mental tasks with varying levels of difficulty. The derived general model utilizes PSD variables from differential EEG channels (C3-C4, Cz-POz, F3-Cz, Fz-C3, and Fz-POz) to generate a continuous measure of workload.

The ECG-based metrics provide insight into stress, arousal, and anxiety (Berntson et al. 1997). The ECG signal consists of a well documented sequence of positive and negative peaks known as the PQRST complex (Berntson et al. 1997; Task Force 1996). For assessing stress, arousal and anxiety levels, the ECG is first filtered to improve the contrast between the QRS complex and the T wave. This allows the detection of peaks to be more robust as double peak detection is minimized. The real-time algorithm implemented in B-Alert Live software (B-Alert LIVE software at [www.advancedbrainmonitoring.com](http://www.advancedbrainmonitoring.com)) calculates the inter-beat R-R interval as the number of seconds between consecutive R-waves. Based on that, heart rate (HR) is estimated as a number of beats per minute, i.e.  $60/R-R$  interval. Furthermore, the algorithm assesses the quality of detected beats by monitoring the standard deviation of the consecutive beats. Furthermore, heart rate variability (HRV) is also computed and parasympathetic versus sympathetic arousal is evaluated through the ratio of low frequency (LF) to high frequency (HF) HRV. Higher LF/HF ratios are believed to be associated with increased sympathetic activation, and stress/anxiety states (Berntson et al. 1997; Task Force 1996).

Lastly, affective metrics include positive/negative affective state classifier and empathy metrics based on mu-suppression (EEG mu rhythms are believed to reflect mirror neuron activation). Affective state classifier (Stikic et al. 2014) is a general model that utilizes X24 EEG data to classify the data into positive/negative affective state. The most discriminative set of EEG variables was determined and the model was trained on the data from 98 participants who watched commercially available videos (clips from "America's Funniest Home Videos" and battle scenes from the war drama "Saving Private Ryan") to induce positive and negative states. Mu-suppression is the log ratio of 8–13 Hz PSD over C3, C4, and Cz EEG channels during the task in question and the baseline eyes closed task. Log ratios of less than





Fig. 11.2 Teaming platform

zero are indicative of suppression, values of zero are indicative of no change, while positive values are indicative of enhancement (Cheng et al. 2008, 2014; Corradini and Antonietti 2013).

The overall goal is to combine the previously described EEG and ECG real-time measures of individual team members into quantitative team metrics. However, synchronized and simultaneous EEG data collection from multiple individuals, such as in teaming experiments, imposes a variety of challenges that often compound proportionally with the number of individuals and teams involved. In most cases, teaming experiments involve well-defined problem solving exercises and the integration of physiological recordings must thereby be achieved in a seamless and unobtrusive manner to ensure unbiased and efficient execution of the tasks. The synchronization of data from multiple subjects is another challenge, as team-based metrics are inherently reliant upon the link between specific events in the teaming task with the physiological signals of the individuals participating in the task. Due to the importance of task-specific context, millisecond level synchronization with the task events, as well as between individuals, is necessary for accurate analysis and assessment of team process episodically.

An additional difficulty associated with many team tasks is physical movement of team members, which complicates the acquisition of clean EEG data. As EEG-based features and metrics rely upon clean EEG data from all individuals, concurrent and intuitive data quality monitoring is essential. Early warnings of any data quality issues allow the task administrator to take necessary steps to salvage the data, before completion of the task. Data aggregation and semi real-time analysis and visualization are another set of requirements that are important for task mitigation and early interventions. The novel EEG teaming platform developed by ABM (see Fig. 11.2) allows for recording, synchronization, and viewing of EEG data from each team member, and it enables the aggregation of combined data across all

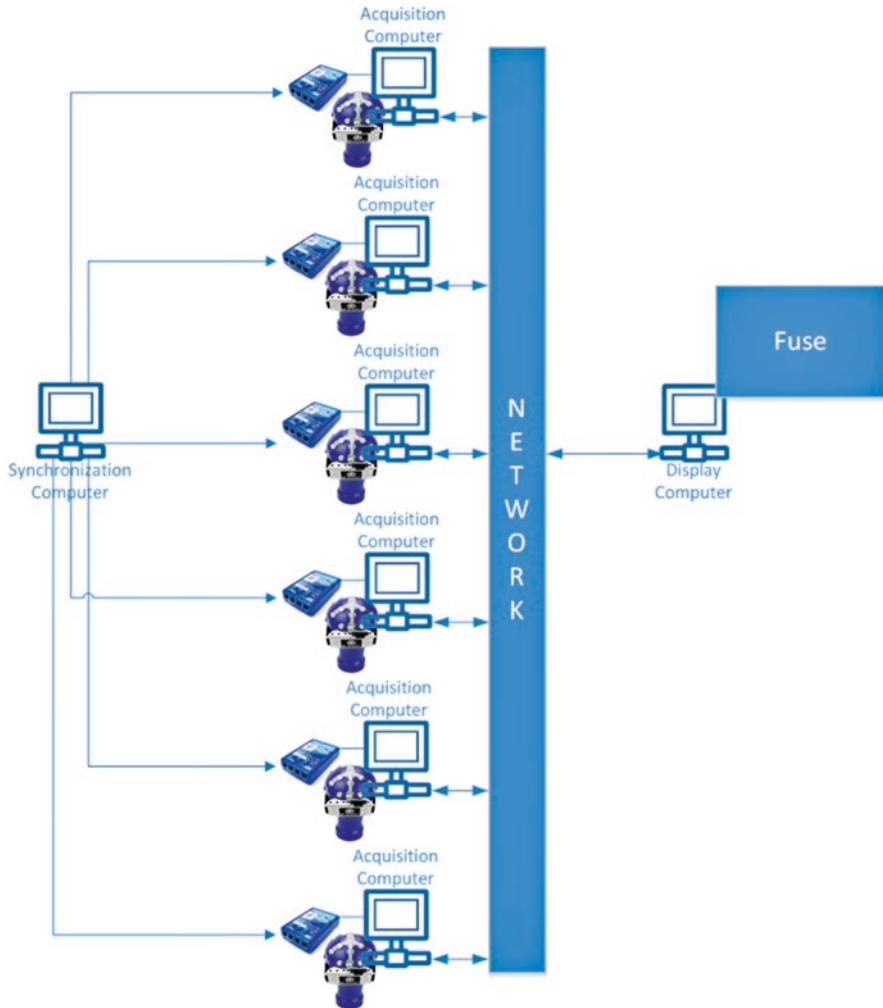
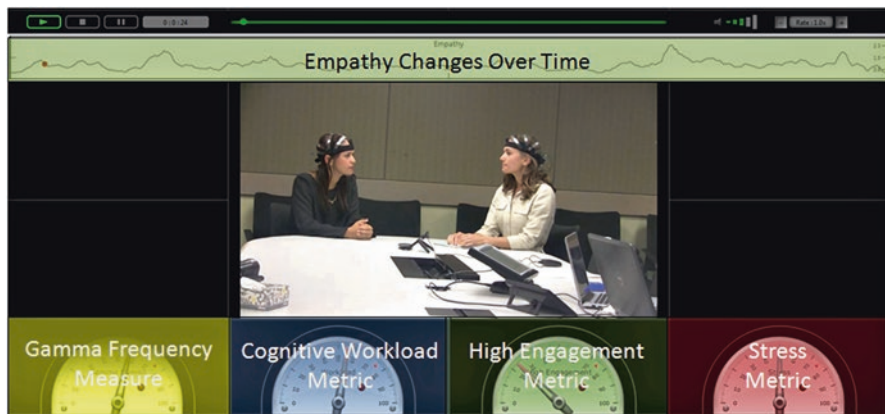


Fig. 11.3 Fuse architecture

team members to provide group metrics. In particular, the teaming platform supports the use of multiple EEG systems, and it addresses the aforementioned issues to a significant extent.

The platform can be operated using either Network Time Protocol (NTP), over Ethernet, or via proprietary serial network synchronization using ABM's External Sync Unit (ESU). The ESU can achieve much higher synchronization accuracy than NTP, however, it limits user separation to the transmission length acceptable in a wired RS232 network. As shown in Fig. 11.3, the teaming platform operates on four communication links: (1) The Acquisition computer interacts with the ESU via a USB port, (2) the EEG Headset transmits data to the ESU via Bluetooth protocol,



**Fig. 11.4** Fuse GUI for study of leadership skills training

(3) the Synchronization computer sends “sync” beacons to the ESU over a serial network at regular intervals, and (4) the Display computer acquires data over the Ethernet. In order to visualize data from multiple headsets, ABM built a software platform called Fuse that runs on the Display computer, and provides unprecedented levels of flexibility and customization; the user can build a custom Graphical User Interface (GUI) with rich multimedia components on-the-fly. Fuse hosts a server, and communicates in a custom and scalable ABM protocol to any number of clients on the network. The installation includes a client Software Developers Kit (SDK), including a wrapper written in C++ that enables easy integration with acquisition software, while abstracting the complexities of the protocol. A rich multimedia library—which includes drag and drop components for video, audio, charts, and gauges such as heat maps and meter gauges—allows users to build context-specific display interfaces. The components can then be associated with a signal source, which may include either raw or processed data. Fuse supports three modes: (1) capture and display of data in real-time, including data rewind; (2) data playback based on a continuous clock; or (3) data playback based on a static clock with the user advancing the clock manually. The latter two modes facilitate offline visualization and data annotation during offline analysis.

The teaming platform has been utilized to collect EEG data from over 100 teams (comprising more than 500 team members) across a range of user studies (e.g., Stevens et al. 2009a, 2013; Waldman et al. 2013; etc.). For example, the teaming platform has proven useful in several studies designed to improve understanding and training of leadership skills such as coaching or time, emotion, and information management. One use case is shown in Fig. 11.4 in which a role-playing task was performed. It involves a customized, structured, and scripted one-on-one conversation between a manager and an actor. For the scenario in question, the supervisor conducted a performance review, while the actor played an outspoken direct report. For the duration of the session, the following measures were recorded: (1) video of both participants through one video feed, (2) audio, (3) two B-Alert X10s with 9

EEG channels each, (4) ECG using the optional port on each B-Alert X10s, and (5) expert ratings. In addition, the B-Alert Live software computed second-by-second EEG/ECG based metrics: (1) EEG-engagement (i.e., cognitive state), (2) EEG-workload, (3) HRV, (4) PSD bandwidths (delta, theta, alpha, beta, and gamma), and (5) empathy. The Fuse GUI enabled easier annotation of data and more detailed review of the selected neurophysiological measures with the video footage of the scenario. After computing all of the neurophysiological measures, the data were reviewed alongside the video footage of the session, and the observational expert scoring was used to discover patterns, and relate the employee's physiology with segments of interest.

### 11.3 Teaming Studies

The next section will review some of the latest EEG studies conducted using the ABM's teaming platform, that cover a wide range of application domains, such as education, gaming, narrative storytelling, industrial/organizational, and medical settings.

### 11.4 Neurophysiologic Synchronies

The EEG-workload metric was explored in a team neurodynamics study (Stevens et al. 2009a) to develop a deeper understanding of how teams collaborate when solving time-sensitive, complex, real-world problems. In this study, participants were solving substance abuse management tasks individually, and then in teams of three. The results indicated that nonrandom patterns of neurophysiologic synchronies could be observed across teams, and within members of a team, when engaged in problem solving. Different patterns were discovered early in the collaboration when the team members were forming initial mental models of the problem at hand, as compared to the patterns found later in the collaboration when the team members were sharing their mental models and converging to a solution. The participants expended more EEG-workload in a teamwork situation than they did when performing the task individually, which may relate to the processing "cost" of collaboration. However, time to solve the problem was greatly reduced in the team setting. Furthermore, the resulting patterns found in the EEG data differed depending on the team's efficiency. The efficiency of a team in completing a task (as measured by time to completion) varied directly with levels of EEG workload of the team members—the more efficient a team, the higher their EEG workload. Overall, efficient teams showed more focused (i.e., less distributed) EEG patterns. This approach shows promise for future utility in monitoring the quality of teamwork and optimizing team performance by adaptively modifying the team "flow" when optimal patterns are not present.

A similar approach was applied by Stevens et al. (2013) to analyze submarine piloting and navigation tasks. EEG-engagement and EEG-workload for each team member were combined into a vector representing an aggregated profile of team neurologically derived engagement and workload. These collective team variables, called neurophysiologic synchronies, were modeled by utilizing a self-organized, artificial neural-net. An entropy-based model of the changes in neurophysiologic synchronies over time revealed a dynamic information structure, characterized by fluctuations in the neurodynamic flow of the team. The experiments of Stevens et al. (2013) showed how teams cognitively organize around changes in the task and how this cognitive organization is altered with experience. Effective and well-trained teams typically demonstrate both stability and flexibility to cooperatively accomplish tasks at hand, while also exhibiting the ability to rapidly respond to evolving task demands (Mathieu et al. 2008). This approach enabled characterization of these reorganizations at the neurophysiological level, providing a useful measure for monitoring the quality of teamwork during complex, real-world tasks.

## 11.5 EEG Predictors of Team Performance

The team of Waldman et al. (2013) used the ABM EEG teaming platform to assess team processes in a management team problem-solving context. The goal of the study was to explore the feasibility of continuous neurophysiological assessment of team performance and different psychological aspects of a team process. The teams consisted of up to five MBA students who discussed and attempted to solve a case problem dealing with child labor and corporate social responsibility (Pless and Maak 2011). At the end of the team ethical decision making, two types of psychological metrics (i.e., engagement and leadership) were assessed by team members, both at the individual and team levels. These metrics showed significant correlations with the team performance scores derived by four trained coders. Two of the coders rated the team's solutions in terms of effective problem solving, decisiveness, and creativity. The other two coders rated the level of moral reasoning displayed in the solutions.

PSD summary metrics over gamma bandwidth during the benchmark AMP tasks were significantly correlated with the leadership scores in the study (Johnson et al. 2013), showing that neurophysiological metrics could be predictors of leadership development potential. The psychological metrics of engagement and leadership were then also assessed based on the EEG data acquired during the teaming discussion session itself (Stikic et al. 2013). Different modeling techniques, such as linear and quadratic discriminant function analysis and linear regression were applied to the processed EEG data. The models were evaluated through auto-validation, but also through cross-validation to test stability of the models in the team-independent training setting. The experimental results suggested that EEG could be effectively used in the team settings to classify individual and team engagement, as well as the leadership qualities shown by team members. Lastly, Waldman et al. (2013) showed that team leaders in a team problem solving context were able to elicit more EEG-engagement

from other team members when they spoke during team discussions. The study suggested that neurological assessment could reveal more about a member's true engagement, beyond that person's overt verbal and nonverbal behavior. In sum, this research showed the potential of neuroscience technology to be applied in real-time to examine issues pertaining to the effects of leaders on teams.

## 11.6 Narrative Storytelling

Another area of interest in team-based research is shared audience experience, which was recently investigated in an EEG-based narrative storytelling study (Correa et al. 2015) which aimed to identify and characterize the neural and physiological correlates of narratives on the audience's cognitive and affective states. The persuasive power of narratives for driving positive, prosocial behaviors was evaluated. The study explored the ability of a narrative to influence audience members to donate to a charity, and whether psychophysiological metrics obtained from the audience during the presentation of a narrative were related to the underlying processes of narrative persuasion and/or any resultant prosocial behaviors. A large number of psychophysiological metrics were considered: EEG-engagement, EEG-workload, PSDs (with the focus on midline theta, prefrontal gamma, and left/right occipital/parietal slow alpha suppression), wavelets, HR and HRV, and positive/negative affective state classifier.

The narrative was built around archetypal themes of fairness and justice, situated in a contemporary and cross-culturally applicable context. Specifically, the story involved themes of injustice against women, illegal immigrants, and people with disabilities in an attempt to elicit strong negative emotional responses, and potentially influence prosocial behavioral decisions. The 11-segment story was developed with three variable segments to enable alternative character descriptions that would potentially increase/decrease empathy and character identification for both the main character and antagonist, with the two distinct versions of narrative resolution that varied in levels of injustice ("least just" versus "most just"). Namely, the final story segment contained the two variable versions of resolution: "least just", in which the antagonist was not punished for the crime he committed, and "most just", in which the antagonist could not escape justice, but even the "most just" story version did not result in significant punishment. The participants were grouped into 3-person teams and watched one of the two story versions (Fig. 11.5). Afterwards, participants were asked if they would like to donate to a particular charity out of a list of three foundations related to the narrative, and any charity donations were deducted from their compensation for participation.

Analysis of the charity donations showed that the participants who heard the "most just" story version donated money more often than the participants who heard the "least just" version of the story. A  $2 \times 2$  ANOVA was conducted to determine whether there were statistically significant differences in the psychophysiological metrics between the narrative versions ("most just" and "least just") and donation behaviors ("donated" and "did not donate"). The ANOVA analysis has shown that: (1) subjects



**Fig. 11.5** Narrative storytelling experiment

who did not donate had greater HRV LF:HF ratios during the narrative than those who donated, which reflects an increase in stress response that might have affected their decision-making process; (2) while audiences from both story versions experienced negative affective state, the “most just” story version overall induced a lower level of negative affect than the “least just” version; (3) the participants who donated their money had an overall higher level of negative affect than the participants who did not donate; and (4) those who donated and viewed the least just version experienced a significantly more negative affect than those who donated and watched the most just version, while affective states did not significantly differ across story versions for those who did not donate.

Lastly, in an attempt to characterize donation behavior, it was examined if: (1) one could predict whether the participants would donate to the charity by developing a discriminant function analysis classifier based on the averaged levels of negative affect over the story segments; and (2) one could predict the amount of money donated by performing step-wise linear regression. The developed classifier exhibited accuracy of up to 72% in predicting the donation behavior, while the regression was able to explain up to 88% of the variance. While these pilot results show promise, further cross-validation on a larger sample size is needed to evaluate the generalization capabilities of the trained models. On a higher level, this study was able to demonstrate the importance of psychophysiological measures to understand narrative persuasion, evaluate how psychophysiological measures during key portions of a narrative were related to post-narrative behaviors, and explain variances in such behaviors based on intrinsic characteristics of the audience members, reactions to the narrative, and psychophysiology during the narrative.

## 11.7 Tutoring Dyads

In a recent teaming study focused on tutoring, dyads’ psychophysiology and cognitive processes were investigated with the goal of effectively training individual’s decision making and problem solving abilities, as well as increasing skill

acquisition speed. For that purpose, Stone et al. (2014) explored interactions between tutor and tutee to assess the learning curve by examining synchronous psychophysiological metrics of the dyad's HR, EEG-engagement, and EEG-workload during a spatial reasoning video game. They tested the hypothesis that increased tutor/tutee neural synchrony would correlate with improvements in performance (i.e., increased learning). Initial results indicated small but statistically significant correlations between the analyzed synchrony metrics and performance. First, the individual HR, EEG-engagement, and EEG-workload were analyzed. This analysis showed that HR and EEG-engagement were significantly elevated for the tutees playing the game, as compared to the levels of the tutor. In contrast, there was no statistically significant differences in overall EEG-workload of tutor versus tutee. After computing individual measures, correlation analysis of the tutor's and tutee's psychophysiological metrics was performed, and the resulting relationships indicate moderate levels of psychophysiological synchrony. Third, step-wise regression was explored to determine whether the analyzed metrics could explain the tutee's performance. The psychophysiological metrics for the tutor, tutee, and the correlations between the two were regressed onto the game performance. While the individual psychophysiological metrics were only able to explain a minority of the variance in performance, the correlations were consistently responsible for the majority of the variance explained. These preliminary results imply that synchrony on a psychophysiological level between tutor and tutee impacts tutee performance. The study demonstrated that EEG-based metrics of tutor/tutee dyad neural synchrony may serve as reliable and objective measures of learning effectiveness and pedagogical efficiency, with direct correlations to training and task performance. These findings lead to the hypothesis that neural coherence across dyads may provide more insight into the tutoring learning process, particularly in a task requiring attention and working memory.

In a related study, Stone et al. (2015) explored EEG coherence of the tutor-tutee dyads. Coherence is a measure of the amount of association between two signals in the frequency domain, and it is calculated as a ratio of the cross-spectrum and the auto-spectra of the analyzed signals. Stone et al. (2015) investigated both intra-individual and inter-individual coherence across increasing game difficulty levels. Intra-individual coherence is calculated across two EEG channels within an individual to assess functional brain connectivity, while inter-individual coherence is analyzed between tutor and tutee to estimate gross changes in synchrony. The tutees were grouped into low-skilled and high-skilled players, and only fronto-parietal coherence was analyzed, as it is most closely related to working memory. Intra-individual coherence analysis showed that both low-skilled and high-skilled players had elevated fronto-parietal coherence while playing the game. In comparison to the high-skilled players, however, the low-skilled players exhibited weaker coherence. This suggests that skill levels may be associated with underlying brain connectivity. Initially, inter-individual coherence was higher for the low-skilled than for the high-skill players due to their need to rely more on input from the tutor. As the game progressed, this trend became less prominent. Overall, the coherence-based approach shows promise in delineating between highly skilled and lower skilled tutees.



## 11.8 Quality of Surgical Operations

In an initial study, Guru et al. (2014) investigated the utility of EEG-based cognitive assessment in a medical training environment, during robot-assisted surgery, and compared it against the traditional tool-based metrics and subjective ratings. Surgeons with varying operative experience (i.e., beginners, competent/proficient, and experts) performed basic, intermediate, and advanced skill tasks. The tool-based metrics showed statistically significant differences between extreme groups (i.e., beginners and experts) and/or tasks (i.e., basic and advanced), however, the cognitive EEG-engagement metric was also able to discern between experts and competent/proficient surgeons when performing the advanced-level tasks. Thus, the study results suggested that EEG-based cognitive assessment may aid in defining levels of expertise in performing complex surgical tasks once competence is achieved.

The goal of the next proof-of-concept surgical study, performed by the Naval Surface Warfare Center (Panama City, FL) was to quantify the ability of medical personnel to perform critical surgical procedures onboard, when faced with high sea states that cause increased deck accelerations. The selected surgical procedures for this study included: stabilizing a fractured pelvis, treating a displaced femur fracture, treating an open wound of the abdominal wall, and the treatment of an amputated leg. The involved medical personnel included a surgeon, nurse, surgical technician, and anesthesiologist. The tests were conducted with medical personnel simulating medical treatments in high sea states within a realistic training environment. Participant's physiological indicators of EEG-workload were analyzed during performed surgical procedures under both motion and no-motion conditions. Surgeons had consistently higher workload than technicians, while, technicians' workload increased in the motion condition compared to the control no-motion condition. The overall performance of the teams, however, was not negatively impacted in the simulator's motion condition. Based on the data collected, there was no compelling evidence to suggest that the procedures considered should be avoided during high sea state conditions. However, this study did not take into account operational conditions on ships, such as fatigue and combat stress, which may further impact or interact with workload and overall team performance. Given the small sample size and large number of confounding factors, further explorations of these preliminary and promising results are still necessary.

## 11.9 Discussion

The technology and research reviewed above have the potential to advance the current understanding of team performance, and to identify why some teams perform better than the others. Neurologically-based methods could be used in the future to enhance individual and team performance by leveraging measures of cognition,

emotions, and team processes in different ways, e.g., (1) to objectively assess team interactions within and across teams, (2) to predict team performance in advance of training and adjust the training accordingly, (3) to characterize and train effective team leaders, or (4) to use metric-based feedback to adapt and optimize team training.

For example, neurological metrics could be used to identify team weaknesses, or track team improvements over time. Namely, effective and well-trained teams typically demonstrate both stability and flexibility to cooperatively accomplish tasks. Additionally, effective teams are also able to rapidly respond to evolving task demands. The approach based on neurophysiologic synchronies allows for characterization of these team reorganizations, and thereby may provide a useful tool for monitoring the quality of team work during complex, real-world tasks, and a guideline for adaptively modifying the workflow of team members when the optimal neural patterns are not present, i.e., the teams should be organized in a way to maximize neural synchronies, while the tasks and workload could be adjusted between team members in real-time. For example, situations where a member of the team has lower engagement and/or workload, while the other members are fully engaged and working hard, may indicate a less effective team member. On the other hand, the proposed approach could also detect the situations when certain team members are overloaded, so potential clashes and other issues could be anticipated proactively.

The study performed by Waldman et al. (2013) made an initial step toward quantitative leadership assessment by developing neurological profiles of persons with leadership potential. That approach could be used for selection of natural leaders, but future leadership training could also highly benefit from such profiles by utilizing neuro-feedback to bring the brain into the targeted neurophysiological state of interest. This new line of research and practice holds potential to also enrich information systems with additional data on cognitive state changes of the user that could be employed in a range of operationally relevant applications. Examples include training people to give better presentations with closed-loop feedback on controlling their stress levels, testing the effectiveness of different marketing advertisements, gauging users' reactions to different movie storyline endings, or even to assess responses to political speeches.

The tutoring dyads study demonstrated that neural synchrony may serve as a reliable and objective measure of learning effectiveness, with direct implications to training and task performance. These findings could be used to drive the development of the new intelligent tutoring approaches and their objective assessment. Performance could be improved with real-time or post feedback to encourage faster synchrony in pairs. Furthermore, the proposed approach also shows potential utility towards identifying: (1) those most likely to benefit from tutoring, (2) when a learner has maximized the benefits of the current tutoring session, and/or (3) at what time training modalities should be adapted to optimize cognitive processing.

The aforementioned narrative storytelling study successfully analyzed the donation behavior of an audience. To build upon this early work, another potential use of the combined audience's neural metrics is to incorporate them into closed loop

audience feedback for adaptive narrative structure that could drive real-time changes of the story ending or its characters, depending upon the audience's neural responses. This would result in an individualized narrative tailored by neural signatures associated with cognitive processes such as attention and empathy allowing for optimization of the story to potentially encourage prosocial behavior, such as charity donations.

As indicated by Schaueneman et al. (1984), the lack of surgical knowledge, inadequate skill level, and anxiety associated with learning a new surgical procedure results in higher cognitive load with significant mental demands. In comparison, expert surgeons are desensitized to stress, have adequate knowledge, and are comfortable with surgical planning for execution of the next surgical step, as they have overcome cognitive demands, mastered psychomotor skills, and have already converted fine surgical skills into automated processes. This has also been observed in a study by Guru et al. (2014), in which the expert surgeons exhibited lower EEG-workload and EEG-engagement levels than their less experienced counterparts. Neurological assessment holds great potential to be used as an adjunct to traditional methods for skill assessment in medical training and the operating room. EEG-based feedback could be applied to enhance surgical training, increase the safety, and accelerate the transition from novices to experts.

## 11.10 Future Research Directions

Although neuroscience methods have recently received increasing attention in team research, most existing research in this field is characterized by relatively small sample sizes and controlled laboratory conditions. Before these methods can be applied to real-world applications, larger field studies are needed. For that purpose, one needs to enable synchronized EEG recordings of as many team members as possible for extended periods of time in operational settings where noise in the EEG signal is expected. ABM is working to streamline the EEG teaming platform by further improving acquisition systems, timing accuracy, and artifact decontamination algorithms. Annotation capabilities could be enhanced by automatic triggers when certain patterns are found in the EEG/ECG data that could be of particular interest.

AI systems offer great potential for developing models of EEG across team members through the integration and analysis of patterns over time within these very large data sets. AI models could be useful in rapid assessment of a team to determine compatibility and to make suggestions as to the strength and weaknesses of a particular team.

The majority of previous team studies focused on the traditional co-located teams. However, cyber teams are frequently only virtually linked requiring trust and collaboration without face-to-face interaction. Thus, future research is required to explore whether the discovered neural patterns are applicable in such settings as well. The envisioned future for virtual teams is likely to include AI systems embedded within

the teaming collective where AI could enhance the capabilities of the human team by creating a human-computer hybrid team. This approach has proven successful in medical diagnostics where AI systems can quickly match symptoms, genetics and biomarkers with very large existing medical databases to assist physicians in diagnosis and treatment recommendations. Our work with surgical teams using robotic instruments has already shown team EEG patterns that are conducive to positive outcomes. These results could become part of an AI model for improving training and outcomes in robotic surgeries.

Multimodal inputs beyond EEG and ECG should also be supported to augment the team assessment with additional information from complementary sensor modalities. By fusing multiple sensor modalities, one could compensate for the shortcomings of each separate measure, which could potentially improve accuracy of the algorithms. Fusion of sensor data is another area that could be enhanced with AI applications.

Lastly, although neuro-feedback is a very promising approach for improving team performance, team members should still have the control over the team process, i.e. the future AI systems need to manage complex relationships within teams and interfere only when necessary.

## References

- Astolfi L, Toppi J, Cincotta F, Mattia D, Salinari S, Fallani FDV, Wilke C, Yuan H, He B (2011) Methods for the EEG hyperscanning. Simultaneous recordings from multiple subjects during social interaction. In: Proceedings of the 8th International Symposium on Noninvasive Functional Source Imaging of the Brain and Heart and the 8th International Conference on Bioelectromagnetism, pp. 5-8
- Behneman A et al. (2012) Neurotechnology to accelerate learning: during marksmanship training. *IEEE Pulse* 3(1):60-63
- Berka C, Levendowski DJ, Cvetinovic MM, Petrovic MM, Davis G, Lumicao MN, Zivkovic VT, Popovic MV, Olmstead R (2004) Real-time analysis of EEG indexes of alertness, cognition, and memory acquired with a wireless EEG headset. *International Journal of Human-Computer Interaction* 17:151-170
- Berka C et al (2007) EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviation, Space, and Environmental Medicine* 78:B231-B244
- Berntson GG, Bigger JT, Eckberg DL, Grossman P, Kaufmann PG, Malik M et al (1997) Heart rate variability: origins, methods, and interpretive caveats. *Psychophysiology* 34(6):623-648
- Carson JB, Tesluk PE, Marrone JA (2007) Shared leadership in teams: an investigation of antecedent conditions and performance. *Academy of Management* 50:1217-1234
- Cheng C, Lee PL, Yang CY, Lin CP, Hung D, Decety J (2008) Gender differences in the mu rhythm of the human mirror-neuron system. *PLoS ONE* 3(5):e2113
- Cheng Y, Chen C, Decety J (2014) An EEG/ERP investigation of the development of empathy in early and middle childhood. *Developmental Cognitive Neuroscience* 10:160-169
- Corradini A, Antonietti A (2013) Mirror neurons and their function in cognitively understood empathy. *Consciousness and cognition* 22(3):1152-1161
- Correa KA, Stone BT, Stikic M, Johnson RR, Berka C (2015) Characterizing donation behavior from psychophysiological indices of narrative experience. *Frontiers in Neuroscience* 9:301.

- Funke GJ, Knott BA, Salas E, Pavlas D, Strang AJ (2012) Conceptualization and measurement of team workload: a critical need. *Human Factors* 54:36-51
- Guru KA, Esfahani ET, Raza SJ, Bhat R, Wang K, Hammond Y, Wilding G, Peabody JO, Chowriappa AJ (2014) Cognitive skill assessment during robot-assisted surgery: Separating the wheat from the chaff. *BJU International* 115(1):166-174
- Hannah ST, Balthazard PA, Waldman DA, Jennings PL, Thatcher RW (2013) The psychological and neurological bases of leader self-complexity and effects on adaptive decision-making. *Journal of Applied Psychology* 98:393-411
- Johnson RR, Popovic DP, Olmstead RE, Stikic M, Levendowski DJ, Berka C. (2011) Drowsiness/alertness algorithm development and validation using synchronized EEG and cognitive performance to individualize a generalized model. *Biological Psychology* 87(2):241-250
- Johnson RR, Berka C, Waldman D, Balthazard P, Pless N, Maak T (2013) Neurophysiological predictors of team performance. In: *Proceedings of the 7th International Conference of Augmented Cognition. Foundations of Augmented Cognition, Lecture Notes in Computer Science*, 8027, pp 153-161
- Mathieu J, Maynard MT, Rapp T, Gilson L (2008) Team effectiveness 1997-2007: A review of recent advancements and a glimpse into the future. *Journal of Management* 34:410-476
- McAssey MP, Helm J, Fushing H, Sbarra DA, Ferrer E (2013) Methodological advances for detecting physiological synchrony during dyadic interactions. *Methodology European Journal of Research Methods for the Behavioral and Social Sciences* 9(2):41-53
- Pless N, Maak T (2011) Levi Strauss & Co: Addressing child labour in Bangladesh. *Readings and Cases in International Human Resource Management and Organizational Behavior* 446-459.
- Rich BL, Lepine JA, Crawford ER (2010) Job engagement: antecedents and effects of job performance. *Academy of Management*, 53:617-635
- Schaunenman AL, Pickleman J, Hesslein R, Freeark RJ (1984) Neuropsychologic predictors of operative skill among general surgery residents. *Surgery* 96(2):288-295
- Stevens RH, Galloway T, Berka C, Sprang M (2009a) Can neurophysiologic synchronies provide a platform for adapting team performance? *Foundations in Augmented Cognition*, 5638: 658-667
- Stevens RH, Galloway T, Berka C, Sprang M (2009b) Neurophysiologic collaboration patterns during team problem solving. In: *Proceedings of the International Symposium of Human Factors and Ergonomics Society Annual Meeting* 53(12): 804-804
- Stevens RH, Galloway TL, Wang P, Berka C (2012) Cognitive neurophysiologic synchronies: What can they contribute to the study of teamwork? *Human Factors* 54:489-502
- Stevens RH, Galloway TL, Wang P, Berka C, Tan V, Wohlgemuth T, Lamb J, Buckels R. (2013) Modeling the neurodynamic complexity of submarine navigation teams. *Computational and Mathematical Organization Theory* 19(3):346-369
- Stikic M, Berka C, Waldman D, Balthazard P, Pless N, Maak, T. (2013) Neurophysiological estimation of team psychological metrics. In: *Proceedings of the 7th International Conference of Augmented Cognition, Foundations of Augmented Cognition, Lecture Notes in Computer Science* 8027: 209-218
- Stikic M, Johnson RR, Tan V, Berka C (2014) EEG-based classification of positive and negative affective states. *Brain-Computer Interfaces* 1(2):99-112
- Stone B, Skinner A, Stikic M, Johnson R. (2014) Assessing neural synchrony in tutoring dyads. In: *Human-Computer Interaction International, Lecture Notes in Computer Science*, 8534: 167-178
- Stone B, Correa K, Thor N, Johnson R (2015) EEG coherence within tutoring dyads: A novel approach for pedagogical efficiency. In: *Foundations of Augmented Cognition, Lecture Notes in Computer Science* 9183:697-706
- Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology (1996). Heart rate variability: standards of measurement, physiological interpretation, and clinical use. *Circulation* 93(5):1043-1065

- Waldman DA, Balthazard PA, Peterson SJ (2011a) Leadership and neuroscience: Can we revolutionize the way that inspirational leaders are identified and developed? *The Academy of Management Perspectives* 25(1):60-74
- Waldman DA, Balthazard PA, Peterson SJ (2011b) Social cognitive neuroscience and leadership. *The Leadership Quarterly* 22:1092–1106
- Waldman DA, Wang D, Stikic M, Berka C, Balthazard P, Richardson T, Pless N, Maak T (2013) Emergent Leadership and Team Engagement: An Application of Neuroscience Technology and Methods. In: *Academy of Management Annual Meeting Proceedings*, 73:1-6.
- Westbrook et al (2004) Quantification of alertness, memory and neurophysiological changes in sleep apnea patients following treatment with nCPAP. *Sleep* 27:A22

# Chapter 12

## Cybersecurity and Optimization in Smart “Autonomous” Buildings

Michael Mylrea and Sri Nikhil Gupta Gourisetti

### Abbreviations

AI	Artificial Intelligence
AITA	Artificial Intelligence based Insider Threat Analyzer
B2G	Buildings-to-Grid
BACnet	Building Automation Control network
BAS	Building Automation System
B-C2M2	Building Cybersecurity Capability Maturity Model
BCF	Building Cybersecurity Framework
BEMS	Building Energy Management System
CCA	Critical Cyber Assets
CCTV	Closed-Circuit Television
CFR	Commercial, Federal, Residential buildings
CI	Critical Infrastructure
CMT	Configuration Management Tool
DDoS	Distributed Denial of Service
DER	Distributed Energy Resource
DHS	Department of Homeland Security
DOE	Department of Energy
DoS	Denial of Service

---

M. Mylrea (✉)

Manager, Cybersecurity and Energy Technology, Pacific Northwest National Laboratory, Richland, WA, USA

Executive Cybersecurity Doctoral Program, George Washington University, Washington, DC, USA

e-mail: [michael.mylrea@pnnl.gov](mailto:michael.mylrea@pnnl.gov)

S.N.G. Gourisetti

Research Engineer (Smart-Grid Cybersecurity), Electricity Infrastructure, Pacific Northwest National Laboratory, Richland, WA, USA

Engineering Sciences and Systems Doctoral Program, University of Arkansas at Little Rock, Little Rock, AR, USA

e-mail: [srinikhil.gourisetti@pnnl.gov](mailto:srinikhil.gourisetti@pnnl.gov)

EERE	Office of Energy Efficiency and Renewable Energy
EIA	U.S. Energy Information Administration
EIoT	Energy Internet of Things
FCU	Fan Coil Unit
FPS	Federal Protective Service
GAO	U.S. Government Accountability Office
HIDPS	Host Intrusion Detection and Prevention System
HIDS	Host Intrusion Detection System
HVAC	Heating, Ventilation and Air Conditioning
ICS	Industrial Control System
ICS-CERT	Industrial Control Systems Cyber Emergency Response Team
ICT	Information and Communications Technology
ID	Identification
IDPS	Intrusion Detection and Prevention System
IDS	Intrusion Detection System
IED	Intelligent Electronic Device
IoT	Internet of Things
IPS	Intrusion Prevention System
IT	Information Technology
MAC	Media Access Control
NBAD	Network Behavior Anomaly Detection
NCA	Network Connected Assets
NIDPS	Network Intrusion Detection and Prevention System
NIDS	Network Intrusion Detection System
NIST	National Institute of Standards and Technology
OT	Operations Technology
PIDS	Physical Intrusion Detection System
PLC	Programmable Logic Controller
PNNL	Pacific Northwest National Laboratory
RCM	Risk Characterization Matrix
RFID	Radio Frequency Identification
RTU	Remote Terminal Unit
SCADA	Supervisory Control and Data Acquisition
SCI-RAD	Social Engineering Autonomy for Cyber Intrusion Monitoring and Real-time Anomaly Detecting
SIEM	Security Information and Event Management/ Log Analyzer
SSID	Service Set Identifier

## 12.1 Introduction

Smart buildings and energy technology continue to make innovative advances with machine learning and artificial intelligence, which increasingly take humans out of the loop. Key buildings operations, from energy management to cyber security, business to social behavioral analytics, are increasingly incorporating machine



learning and AI algorithms to optimize productivity, innovation and resilience. As we network and digitize cyber and physical systems, operational technology (OT) and information communication technology (ICT), new opportunities presented to make our buildings, and organizations they support, more autonomous and efficient. However, new cyber threats are also rapidly emerging. Cyber threats to smart buildings exploit smart buildings’ target rich energy-internet-of-things (EIoT) environments because they are often times not designed, configured or operated with security in mind. This chapter highlights how AI software, algorithms and OT that rely on ubiquitous sensors and rapid data collection and exchange can exacerbate this challenge by increasing the amount of data and opening up new attack nodes. At the same time, AI based cybersecurity systems are also needed for smart decision-making and autonomous defense in response to evolving threats such as polymorphic malware and hybrid cyber-physical attacks (Mylrea 2016). AI cybersecurity systems can help improve the state-of-the art by rapidly responding to the dynamic cyber-attack landscape and enhancing the overall cyber situational awareness for building operators even as the threat evolves.

This is especially important, as the cyber threat to smart buildings is complex, non-linear and rapidly evolving. Cyber-attacks have been used to exploit smart building controls and breach corporate networks, cause critical building system failures and enable hackers to pivot from building control system to IT enterprise networks, which are increasingly connected (Mylrea 2016). Attacks targeting building automation system (BAS) and smart energy technology are especially difficult to detect as current intrusion detection systems often times do not monitor OT. This is alarming as OT in buildings, from building automation systems to fire, alarm and access controls networks, are increasingly connected to IT enterprise networks. This trend and challenge will likely increase as smart buildings become increasingly intelligent with AI enabled software and IoT devices. Lack of security as part of vendors and building operators design, deployment and operating criteria certainly exacerbates these challenges. To help pave the way for a more secure adoption of AI enabled technology in buildings this paper explores the cybersecurity opportunities and challenges for Energy Internet of Things (EIoT) environments in smart buildings.

While the terms AI and Machine Learning are often times used leniently and interchangeably in literature, for this paper: data based decision making neural networks that are designed with a feedback loop with the capability to learn over time is a machine learning system. A class of machine learning system that can not only learn from the defined datasets but also can make data based smart decision rather than data based decisions can be classified as AI systems (Marr 2016).

## 12.2 Smart Building Opportunity

AI enabled smart building automation systems help integrate and optimize EIoT environments into smart buildings, presenting many opportunities to network control and automate key aspects of organizations run out of these buildings. Pressure

on building owners and operators to adopt smart building technology is being driven by economic and environmental factors. As a result, owners are quickly moving towards autonomous smart systems that integrate IT and OT with systems that support building functions and business applications. The growing embrace of smart buildings and smart energy technology has led to major increases in process visibility, energy efficiency and conservation, cost savings, interoperability, and the integration of systems. This is especially important as buildings account for nearly 75% of the nation's electricity use, and 65% of the load growth projected by EIA through 2040 (DOE/EIA 2015). Investments in smart building and energy efficiency technology offer a potential gross energy savings worth more than \$1.2 trillion dollars, can help reduce end use energy consumption by 23% of projected demand and abate up to 1.1 Gt of greenhouse gases annually. With a quick return on investment and clear value proposition, the global market is predicted to grow to \$26 billion by 2019 (Towler 2015). As part of this growth, there will be estimated 20.8 billion connected IoT devices in use worldwide by 2020, up from about 6.4 billion connected devices in use worldwide in 2016 (Gartner 2015).

Building automation sensors can range from passive infrared motion detectors, to closed-circuit television (CCTV) motion detection and radio frequency identification (RFID) technologies. By allowing sensors that are usually applied to a single subsystem to be used by other systems, the building can be made more "intelligent". Several examples include: Using RFID tokens to control access to the building or building zones; providing access to the corporate network and retrieving documents on communal printers; and using building security sensors and CCTV motion detection to enhance operation and control of energy management systems. Future smart buildings supported by the combination of AI, big data, and self-learning sensors may enable more robust cyber and physical security and more productive organizations (Mylrea 2015).

### 12.3 Smart Building Challenges

The grand challenge is that as you make buildings smart by networking building automation systems and controls and intelligent through AI it creates an EIoT environment where big data sets are being collected and exchanged it increases complexity and expands the attack surface, creating a target rich environment (Mylrea 2016). This is especially important as in commercial and federal buildings that serve critical economic and national security functions. A cyber attack on buildings can impact both continuity of operations, physical safety, delivery of services and lead to exploitation and theft of sensitive information. Exacerbating the challenge, AI enabled smart buildings require an increasing number of networked sensors and control systems that integrate critical IT and OT assets in buildings; often times with few security controls in place.

As a result, points of vulnerability often time increase. Securing buildings is essential to secure critical infrastructure (CI) sectors. Six of the 16 critical infrastructure

sectors are buildings, which contain assets, systems, and networks, whether physical or virtual, that are considered so vital to the United States that their incapacitation or destruction would have a debilitating effect on security, national economic security, national public health or safety, or any combination thereof (DHS 2016). Despite the importance of buildings, there is a lack of essential building-specific cybersecurity standards, policies, procedures and risk management frameworks (BCF 2016). The U.S. Government Accountability Office (GAO) highlighted these gaps in a report that stated that the federal government was not “addressing cyber risk to building and access control systems particularly at the nearly 9,000 federal facilities protected by the Federal Protective Service (FPS) as of October 2014.” GAO also noted that the government “lacks a strategy that: (1) defines the problem, (2) identifies the roles and responsibilities, (3) analyzes the resources needed, and (4) identifies a methodology for assessing this cyber risk” (GAO 2014).

Cyber challenges to buildings are complex and constantly evolving as the technology evolves. New operational and risk management processes, security practices and paradigms are essential to overcoming these challenges. Traditionally, hardware and software dedicated to monitoring, detecting, and actuating physical processes, also known as operational technology (OT) had been isolated from IT systems. However, the growing convergence of IT and OT in smart buildings has increased overall operational complexity, introduced unanticipated risks, and created new challenges regarding organizational roles, responsibilities, and risk management. The convergence of cyber and physical systems and their underlying complexity also exacerbates the problem of accurately distinguishing the cause of building system anomalies. It is difficult to prioritize the appropriate response if an operator can’t identify and detect the difference between a software or hardware failure, human error, cyber or physical attacks or any combination of failures. AI and machine learning can help improve the state-of-the-art in identifying baselines of IT and OT networks and systems as well as detecting and automatically responding to anomalies caused by cyber and physical events.

This is especially important as cyber-attacks are increasing and include a wide range of actors, such as: terrorists, cyber-criminals, industrial spies, disgruntled insiders and hacktivists (See Table 12.1). During fiscal year 2015, Department of Homeland Security’s (DHS’s) Industrial Control Systems Cyber Emergency Response Team (ICS-CERT) responded to 295 cyber incidents, which represented a 20% increase over the number of incidents in 2014 (ICS-CERT 2015). The energy sector had the second highest number of incidents, while the number of incidents reported by the critical manufacturing sector nearly doubled. A detailed analysis of these attacks revealed that several vulnerabilities could have potentially used by hackers for exploiting processes, controls, and building-connected smart energy technology. DHS reported that hackers are continuously probing critical infrastructure ICS networks in the United States in search of new vulnerabilities to be exploited. Part of the challenge is that control systems found in buildings include legacy systems designed decades ago without any cybersecurity controls.

Moreover, these systems are often not configured securely and remain vulnerable to cyber attacks. Conventional security measures such as access control, authentication,

**Table 12.1** Objectives and motivation for hacking smart building energy technology and controls systems

Denial of service
Theft of intellectual property
Compromising a company's building systems as part of a blackmail scheme
Negatively affecting the public image of a company and thereby taking advantage of a predictable drop in its share price
Directly aiding a competitor who can benefit from the victimized company's loss of production
Compromising a building system to expose the other networks comingled with the control system
Terrorists, "hacktivists", or even disgruntled employees might want to disrupt operations, endanger personnel, damage property, or damage reputations
The threat of state-sponsored agents stealing information and other strategic info from building system data
Disrupting critical energy services through sending malware back into other critical energy infrastructure.

and encryption are necessary, but these measures can be easily circumvented by cyber-attacks that are both sophisticated (zero-day exploits<sup>1</sup>) and polymorphic malware or simple phishing attacks. Hence, if an adversary successfully launches a cyber-attack, more comprehensive security design and risk management solutions are needed to identify and protect potential cyber threats, as well as to detect, respond, and recover (NIST 2014) from cyber incidents (Hagerman 2016). In addition to external adversaries, disgruntled insiders pose a significant threat to smart building control systems. Insiders have physical access inside buildings and often times privileged access to networks and knowledge of the systems, enabling them to launch cyber and cyber-physical attacks that can be difficult to detect. Data from an IBM report indicates that in 2015, 60% of all cyber-attacks were carried out by insiders and 44.5% of them were designed with malicious intent (Kim 2016).

In 2013, Target Corporation was hit by a massive data breach that exploited hacked credentials from a heating, ventilation and air conditioning (HVAC) contractor. The cyber attack exposed the credit card and other sensitive information of almost 40 million customers who purchased from its stores during the first few weeks of the holiday season. As a result, Target agreed to pay as much as \$67 million to resolve claims by financial institutions. The attack led to sales losses, stock prices to fall and caused enormous damage to the retail chain's reputation. Eric Chiu, co-founder and president of Hytrust, a cloud security automation company, told Security Week "In this new 'IoT' world, heating is connected to the same corporate networks that run other systems such as point-of-sale applications and customer databases.

<sup>1</sup>As defined by Wikipedia at [https://en.wikipedia.org/wiki/Zero-day\\_\(computing\)](https://en.wikipedia.org/wiki/Zero-day_(computing)), a zero-day (also known as zero-hour or 0-day) vulnerability is an undisclosed computer-software vulnerability that hackers can exploit to adversely affect computer programs, data, additional computers or a network. It is known as a "zero-day" because once the flaw becomes known, the software's author has zero days in which to plan and advise any mitigation against its exploitation (for example, by advising workarounds or by issuing patches).

This concentration of systems, networks, and data creates a treasure trove for attackers looking to steal information” (Security Week 2014; Hagerman 2016).

Building automation systems are directly or indirectly connected control systems that lack the essential defenses and present a rich target to malicious cyber-attackers. To highlight these vulnerabilities, IBM’s ethical hacking team performed cyber-tests using simple scanning techniques on a building management company, which operated more than 20 buildings across the United States. Basic security errors and flaws were revealed in the firmware which helped in accessing the building management system in one building. A remote execution flaw was discovered which provided them access to the company’s central server and provided them the ability to control the building automation systems in of all the 20 different buildings that were controlled by the company (Hagerman 2016). With root access to a building control system network, hackers could easily cause damage. For example, damage could be inflicted on a data center in a building by simply shutting off the air-conditioning and turning up the heat. In addition, BASs are increasingly being installed and connected to the IT infrastructure by building owners and operators, so penetrating the BAS could open up access to the IT network and vice versa (Ionesco 2016).

Another major challenge is the rise in the number of networked devices and control systems in buildings (Hardin et al. 2015). According to forecasts by Gartner Inc., since 2015 there will have been a 30% increase in the number of connected devices worldwide, reaching 6.4 million connected devices in 2016 and increasing to around 20.8 billion by 2020. In 2016, 5.5 million new things were connected every day (Gartner 2015). EIoT and IoT devices found in buildings prioritize functionality, user-friendliness, and price; and security is often times an afterthought. This fact was highlighted by HP’s recent study noting that 70% of the most common IoT devices contained vulnerabilities, with an average of 25 vulnerabilities per device (HP 2014). Thus, both the expansion in the attack landscape and inherent vulnerability of the devices means that traditional security measures like secure configuration, whitelisting,<sup>2</sup> patch and inventory management<sup>3</sup> offer limited mitigations to emerging threats found in smart buildings.

Converging IT and OT networks in buildings offers a treasure trove of data and potential points of compromise for hackers. In 2009, a security guard at a Dallas-area hospital introduced malware to the hospital’s computers controlling the ventilation, heating and air-conditioning systems for two floors. This could have

---

<sup>2</sup>As defined by Wikipedia at <https://en.wikipedia.org/wiki/Whitelist>, A **whitelist** is a list or register of entities that are being provided a privilege, service, mobility, access or recognition. Entities on the list will be accepted, approved and/or recognized. Whitelisting is the reverse of **blacklisting**, the practice of identifying entities that are denied, unrecognized, or ostracized.

<sup>3</sup>According to Technopedia at <https://www.techopedia.com/definition/13835/patch-management>, **Patch management** is a strategy for managing patches or upgrades for software applications and technologies. A patch management plan can help a business or organization handle these changes efficiently. Technopedia defined Network Inventory Management at <https://www.techopedia.com/definition/29987/network-inventory-management> as, Network inventory management is the process of keeping records of all the IT or network assets that make up the network.

threatened patient medications, treatments and well-being. While hospitals increased reliance on IoT and building automation could make them more vulnerable to cyber attacks, this attack could have played out in almost any commercial or federal buildings with similar vulnerable configurations and devices. A report about federal facility cybersecurity by GAO noted that security officials interviewed said that “cyber-attacks on systems in federal facilities could compromise security countermeasures, hamper agencies’ ability to carry out their missions, or cause physical harm to the facilities and their occupants” (GAO 2014). Cyber attacks on building automation have caused physical damage. According to a recent report, a blast furnace at an unnamed German steel mill suffered damage due to a cyber-attack. The attackers used spear-phishing attack to gain initial access to the business network, and then pivoted to other parts of the system to access and control the networked physical control systems for plant equipment on its production network. These incidents emphasize the vulnerabilities created by the convergence of IT enterprise and OT networks (Hagerman 2016).

The cyber-attacks on Target, Home Depot, and the German steel mill attracted national media attention. However, numerous attacks go undetected or unreported. A recent industry study investigating the vulnerability of building automation systems found that very basic security flaws which would permit the most basic hackers into a company’s networks were present in 55,000 networked smart buildings systems (Automated Buildings 2014). Many of these building automation and industrial control system devices are woven into the digital fabric of our nation’s critical manufacturing and operations. A study by Symantec highlights that vendors often place more importance on ease and interoperability than on cybersecurity. The following vulnerabilities were commonly found: Lack of encryption, use of default/weak passwords, transfer of sensitive information over open networks, and access points that make it easy for hackers to intercept or manipulate information, to control devices, and to break into corporate networks (Wueest 2015).

A study on Shodan, an easy-to-use search engine that navigates the back channels of the Internet looking for connected devices and interrogating available services collected over 1,000,000 unique IP addresses that appear to belong to Internet-facing control systems’ assets such as remote terminal units (RTUs), programmable logic controllers (PLCs), intelligent electronic devices (IEDs), BASs, and the like. Of these, 13,475 were HVAC, building automation and devices manufactured by popular vendors. The study indicates that these systems allow easy access to the connected networks and provide an indirect avenue for cyber-attacks. Many of these devices had default, weak, or non-existent credential requirements. Most detected systems did not implement firewalls or have adequate encryption for defending against a hacker’s entry into the networks (Radvanovsky 2013). Researchers also found, due to system integrators’ prioritization of functionality and ease-of-use over security controls, 204,416 serial-to-Ethernet devices that bypass traditional firewalls (O’Harrow 2012).

## **12.4 AI Enabled Building Automation Is Blurring the Lines Between Information Technology and Operations Technology**

AI enabled building automation is converging and automating information technology and operations technology that were traditionally isolated, independent and analog. In a recent survey involving Building Operating Management, 84% of respondents said their building automation systems were connected to the Internet. However, building owners and operators are often times not aware of devices that are network-connected and they therefore lack the necessary inventory of assets and segmentation of network. Lack of cyber situational awareness is attributed to the fact that historically BAS, and corporate IT systems were managed by operations teams and IT teams respectively, each with different operational processes, practices, and governance policies. These organizational boundaries, combined with systems integration and interconnection, can introduce significant operational complexity and cyber risk into intelligent buildings. For example, removing a virus or malware from a building management system may be significantly more complex as cyber and physical devices are weaved together and speak different protocols. The problem may be further exacerbated by legacy systems and unsecure third party access to the systems (Pullen 2014).

Increased connectivity and automation can improve the energy efficiency and conservation potential of a building to achieve substantial cost savings. On the other hand, increased connectivity provides a number of vulnerabilities that can be exploited to compromise a building’s connected systems availability, confidentiality, and integrity. A cyber-attack can also amplify impacts from operator errors or system misconfigurations, which may be intentional or unintentional. Building systems can also be breached when well-intentioned but inadequately trained building owners/operators are unable to properly configure sophisticated control systems. Another challenge is that building operators often times don’t know how their building networks are structured or connected insecurely to the Internet.

## **12.5 AI Enabled Autonomous Building Automation to Enhance Security**

AI security enabled smart buildings can help mitigate emerging cyber threats and vulnerabilities to buildings continuity and the availability of critical infrastructure. This chapter examines some of these new AI enable security opportunities as they relate to mitigating cyber-physical vulnerabilities in smart buildings. AI enabled autonomous system can be designed to deliver active defense, improve cybersecurity situational awareness and overall security posture of buildings. Finally, AI enhanced systems can help maintain more dynamic end-to-end cybersecurity defenses in buildings operating in an environment of evolving threats.

### ***12.5.1 AI Enabled Threat Identification and Mitigation***

Integrating smart devices in a building helps enhance the comfort level of the occupants but can expose the building to new cyber threats. An examination of the 64,199 cyber incidents and 2260 cyber breaches reported in the United States indicate an increasing threat to IT and OT devices found in buildings and other critical infrastructures. These attacks also highlight the challenge of identifying the threat in EIoT building environment being attacked by evolving hard to detect threats such as: insider threat, polymorphic malware and zero day exploits, just to name a few.

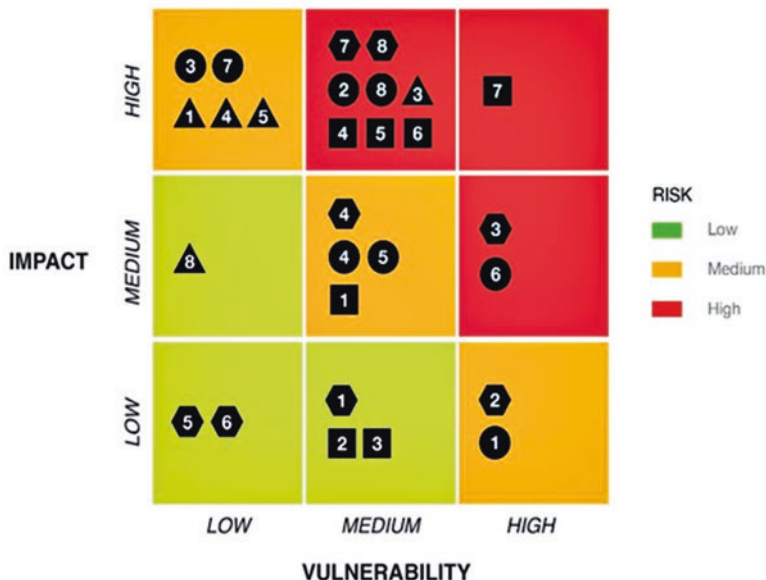
#### **12.5.1.1 Theoretical Concept: AI Based Identification System**

AI enabled threat identification and access control systems in buildings can provide more intelligent self-learning security systems to improve the state of the art in detection of cyber related anomalies to EIoT and building automation systems. To realize this goal, AI systems will draw on both the cyber signatures of the EIoT sensors and communications, logs from the intrusion detection and access control systems as well as the physical properties of the smart devices to enhance the overall security of the building's critical cyber assets. This requires the improvement of AI enabled algorithms aggregating data from physical properties to establish a baseline of what normal IT and OT operations look like as well as automatically respond to any critical deviations caused by an attack. This theoretical concept could also be applied to enable an AI based identification system to autonomously classify assets and develop a risk-characterization matrix (RCM) as shown in Fig. 12.1. Based on the RCM, the AI system could establish roles based access controls, estimate the impact of an attack and help reduce human related vulnerabilities by increasingly taking them out of the loop. In addition, the AI system could help perform more effective threat vulnerability assessments (PNNL 2016) based on the factors such as estimated scale of an attack, safety concerns, financial losses, degradation of services, loss of privacy, and degradation of comfort, etc.

#### **12.5.1.2 AI Based Security Learning System: *Theoretical Concept***

The increase in automation and connectivity has created an EIoT environment in smart buildings that increasingly challenging to protect from emerging cyber threats. Traditional information assurance best practices, such as conducting an inventory and whitelisting critical cyber assets are challenging to implement in an environment of rapid collection, aggregation and exchange of large data sets and a plethora of networked devices. Exacerbating this challenge, users increasingly bring their own devices and leverage cloud-based services, and virtualization and automation further limit centralized monitoring and cybersecurity controls. In response, AI enabled building control systems could draw from the physical





CRITICAL ASSETS BY SECTOR



**Residential**

1. Smart Thermostats
2. Building Lighting System Controller
3. Smart Water Heaters
4. Smart Refrigerator, TV
5. Home Monitoring Cameras
6. Smart Meter, Smart PV Inverter, EV Charger
7. Smart Smoke Detectors & CO Monitors
8. Intrusion Detector (Physical)



**Small/Medium Commercial**

1. Garden Sprinkler
2. Smart Backup Generation
3. RFID
4. CCTV System
5. Refrigeration
6. Cooling Fans (FCU) Controller
7. Building Access Control Management Server
8. Fire Alarm System Controller & Silent Alarms



**Large Commercial**

1. Vending Machines
2. Building HVAC Controller
3. Guest Devices
4. Printers/Scanners
5. Gas Supply Controller
6. Electric System Controller
7. BEMS Server/PLCs



**Federal**

1. Biometric & Key Cards
2. Environment Sensors
3. Body Scanners
4. Printers
5. Chemical/Biological/Radiation Detectors

Fig. 12.1 Illustrative risk characterization matrix—heat map of smart building’s assets. Source: BCF (2016). Note: this illustrative risk characterization matrix varies based on type of building, configuration, etc.

properties of devices to log and learn voltage consumption and frequency norms to identify and improve baseline algorithms of what normal looks like and better respond to all hazards based on potential impact. If physical or cyber anomalies occur, AI enabled building systems will be able to better detect what caused the deviation (cyber-attack, environmental, computational or human error). Depending on the cause, the AI enabled devices could drop malicious commands if the malicious payload was already delivered. AI enabled networks and control systems could also provide increased resilience against cyber-attacks that are hard to detect and deflect, such as zero day exploits, polymorphic malware and even insider attacks. For example, if an intrusion detection system fails to detect a malicious signature; physical change in the AI enabled devices power or voltage consumption caused by the malware could still be detected leveraging the physical properties of the device.

Adaptable and resilient AI enabled protection schemes such as a next generation intrusion detection system could also automate and improve the process of identifying a system's health based on known vulnerabilities to building automation systems as shown in Fig. 12.2.








## ***12.5.2 AI Enabled Cybersecurity Protection***

AI enabled smart buildings create new cybersecurity opportunities and challenges for organizations. There are a number of challenges involved with the integration of an AI system to current IT and OT networks, such as the potential to increase the number of attack vectors and amount of data being collected, stored and exchanged. However, AI enabled building IT and OT networks can also help collect, aggregate and make more intelligent security decisions based on patterns exhibited by building occupants communications and movements as well as signature based patterns of building automation and ICT network logs and traffic. As a result, AI enabled protection schemes may be able to better detect a wide variety of potential cyber and physical threats and increase cyber situational awareness from threats that are difficult to detect, such as an insider threat or an adversary that has already gained root access.

Regardless of the building type, organized cybersecurity training and strict access controls will remain key elements, along with network segregation, dynamic asset protection schemes. The following cybersecurity controls will help secure the increasing number of potentially vulnerable end points found in AI enabled EIoT (Fig. 12.3).

### ***12.5.2.1 The Role of AI in Cybersecurity Protection: Theoretical Concept***

While various cybersecurity controls, such as the list above, can help mitigate risk of cyberattacks, the increasing number of devices and data used in smart AI enabled smart energy technology require new cybersecurity protection policies, procedures and systems. Traditional security controls like whitelisting and inventories are being

-  Lack of inventory and identification of Critical Cyber Assets (CCA)
-  Lack of IT & OT security roles and responsibilities
-  Lack of patch management
-  Lack of separation between IT and OT networks
-  Lack of physical and cyber access control
-  Lack of authentication and encryption of CCA
-  Lack of periodic threat vulnerability assessments, penetration tests & mitigation efforts
-  Lack of cybersecurity training and security audits
-  Lack of redundancy
-  Poor password management policies
-  Default software and network configurations
-  Lack of data and configuration backups
-  Lack of response and recovery plans
-  Lack of secure communication protocols
-  Lack of a risk management strategy

**Fig. 12.2** Typical cyber vulnerabilities found in building automation systems

challenged as the size and speed of the data being exchanged as well as the number of devices increase exponentially. Adding to the challenge, technology is becoming dispersed and decentralized due to policies such as bring your device as well as processes and systems that increasingly leverage the cloud and virtualization. Certainly, this landscape is vulnerable to cyber-attacks and would benefit from more resilient systems enabled by AI and machine learning. Future smart buildings would benefit from AI enabled protection systems that leverage cyber and physical sensors to automate the process of identifying the critical sections of the network that needs to be segregated. An AI enabled protection system would have the capability to better monitor and enforce roles based access rules and determine what assets can see or touch any of the IT and OT networks.

### Protection Schemes

-  Implement network segregation
-  Implement password management policies (strong passwords)
-  Change default SSID. Hide SSID and MAC filtering (especially for residential)
-  Implement firewalls and configuration policies
-  Encrypt all means of data transfer/information communications
-  Determine roles and responsibilities; establish access controls
-  Provide cybersecurity awareness education and training to all building personnel
-  Securely store and exchange control system data to protect against data/privacy breaches
-  Implement plans for asset and network redundancy
-  Implement plans for asset transfers and backups
-  Implement integrity checks for automation software and firmware
-  Implement a backup mechanism for sensitive information
-  Run periodic vulnerability, continuity, and penetration tests
-  Implement a vulnerability management plan
-  Authenticate, approve, and log the remote maintenance of building assets
-  Maintain and protect audit logs such as Firewall logs and network audits

**Fig. 12.3** Cybersecurity controls for buildings automation systems

The AI protection system could also provide more adaptive defenses that continue to learn to respond to the evolving threat, which is increasingly necessary to mitigate threats and attacks to implement cybersecurity control schemes in smart buildings. Having such autonomous AI protection systems could also improve the required protection schemes and the information about how the implemented regulations and schemes can be updated in response to an evolving threat.

### ***12.5.3 AI Enabled Cyber-Physical Intrusion Detection System***

Malicious cyber actors continue to evolve in their methods to compromise critical cyber building assets. A more agile AI based Intrusion Detection System (IDS) will help automatically detect, log and even mitigate potential cyber intrusions to building technology and provide a more active defense. This is especially important as most attacks either go undetected or are detected too late. In fact, the average amount of time before an organization detects a cyber attack is about 220 days (Verizon 2016). Moreover, IDS are often set up and configured for IT enterprise environments and don't register attacks on OT. Next-generation AI based IDS will include self-learning algorithms that improve existing capabilities to continuously monitor and respond to both IT signatures as well as physical properties in OT such as voltage, energy consumption, frequency, line currents, etc. The AI based IDS will leverage these physical properties to establish IT and OT baseline configurations as a reference case. This reference case could be used to autonomously identify deviations from expected behavior by assessing the performance of the assets and monitoring traffic flow between the assets. Upon a successful mitigation, the AI system could update the reference case log and strengthen the security of the assets.

While monitoring the building's assets for any anomalous behavior, the AI based IDS could also look for deviations in fundamental asset parameters such as:

- Frequency deviations (PNNL 2012)
- Voltage and current deviations
- Pressure drops and rises (for example, in air handlers)
- Possible unauthorized access attempts or denials
- Multiple access attempts
- Authentication failures
- Warnings or pop-up windows
- Unknown traffic (to be determined if expected or unexpected)
- Cyber traffic overflow, lack of routine traffic, or buffer overflow. (DDoS attacks result in such behavior.)
- An unusually slow cyber system
- A change in website design or the deletion of entire pages

Some of the AI based IDS could include:

- **Host IDS (HIDS)** to identify cyber-attacks/events based on the behavior of the host-asset.
- **Network IDS (NIDS)**, which identify a cyber-attack/event based on network traffic.
  - Protocol-based controls, which can reside on both HIDS and NIDS in order to monitor communication protocols between control systems and building assets.
- **Physical IDS (PIDS)**, which identify cyber threats to physical systems.

- **Intrusion Prevention Systems (IPS)**, which analyze network traffic flows, automatically drop malicious data packets or block traffic or reset connections upon detecting cyber breaches and before reporting the identified threats. Unlike IDS, IPS are not passive systems. They respond with automated actions to prevent possible intrusions.
- **Network Behavior Anomaly Detection Tools (NBAD)**, which monitor network traffic characteristics such as traffic sources or destinations, traffic volume, protocol use, and others to identify any departure from normal behavior.
- **Security Information and Event Management/Log Analyzers (SIEM)**, which makes security-related data and generated logs available from a single point, making it easier to identify anomalous patterns.
- **Configuration Management Tools (CMT)**, which control the processes for identifying and implementing secure configurations for products and systems, while maintaining their integrity.

Intrusion detection systems and intrusion prevention systems are often used together as an Intrusion Detection and Protection System (IDPS). IDS and IPS are different tools, since their primary functions are, respectively, visibility and control. But when used together they enhance the security framework. The various capabilities of HIDPS and NIDPS are provided in Figs. 12.4 and 12.5 respectively.

Based on a building cybersecurity risk assessment, the AI system could determine which Intrusion Detection and Protection System (IDPS) needs to be installed to mitigate that risk (Martin 2016). As the building and organizations risk profile changes, the AI enabled IDPS could adapt to provide a more holistic and agile active defense.

AI enabled IDPS advances will be supported, in part, by advances in software defined networking (SDN). SDN will help hide networks and firewalls and deploy new devices and network infrastructure by leveraging network flow definitions and automatic dimensioning rules.

### 12.5.3.1 An Integrated AI Based IDPS: *Theoretical Concept*

Using intrusion detection and protection system (IDPS) involves several implementation steps. These may include tuning host and network configurations, notification settings, and determining the strategic locations of networks and hosts. Despite going through such processes, IDPS generates an immense amount of data that could contain false positives, recorded trivialities, and activity information that often leads to missing critical cyber-attack detection information, that is, a loss of valuable information due to obfuscation. Since IDPS and other detection systems use pre-defined baselines, effective use of these systems often requires manual, periodic updates of the baseline. These updates depend on detections, changes in network configurations, changes in expected behaviors of the assets, and more. Based on a condition-set, an AI-based adaptable neural network can automate this

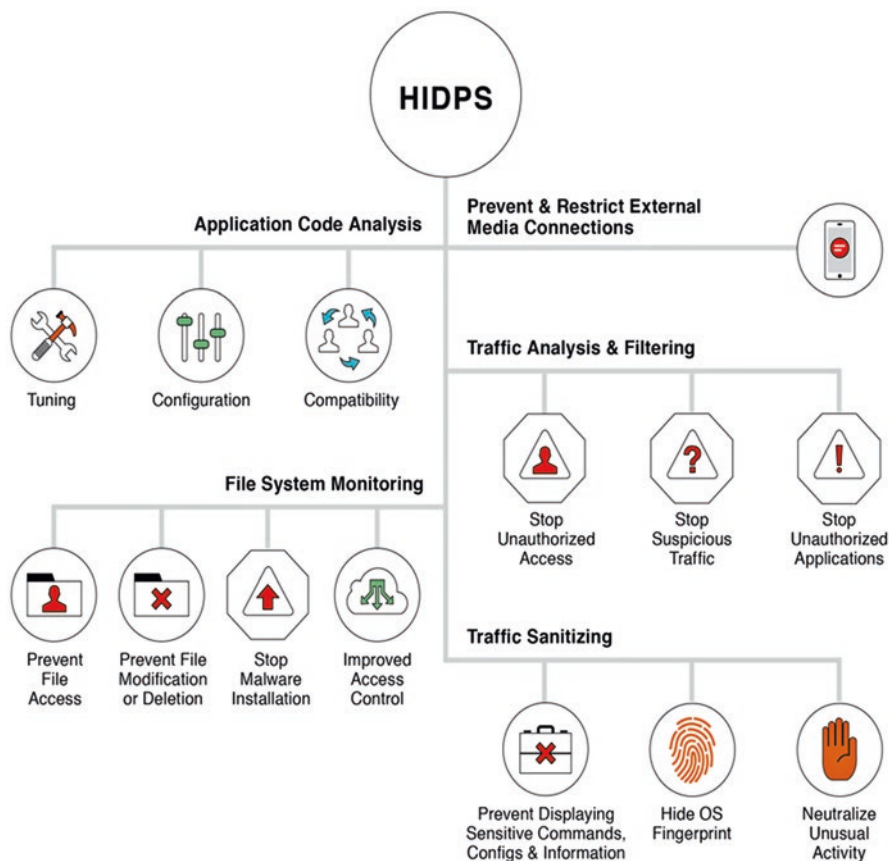


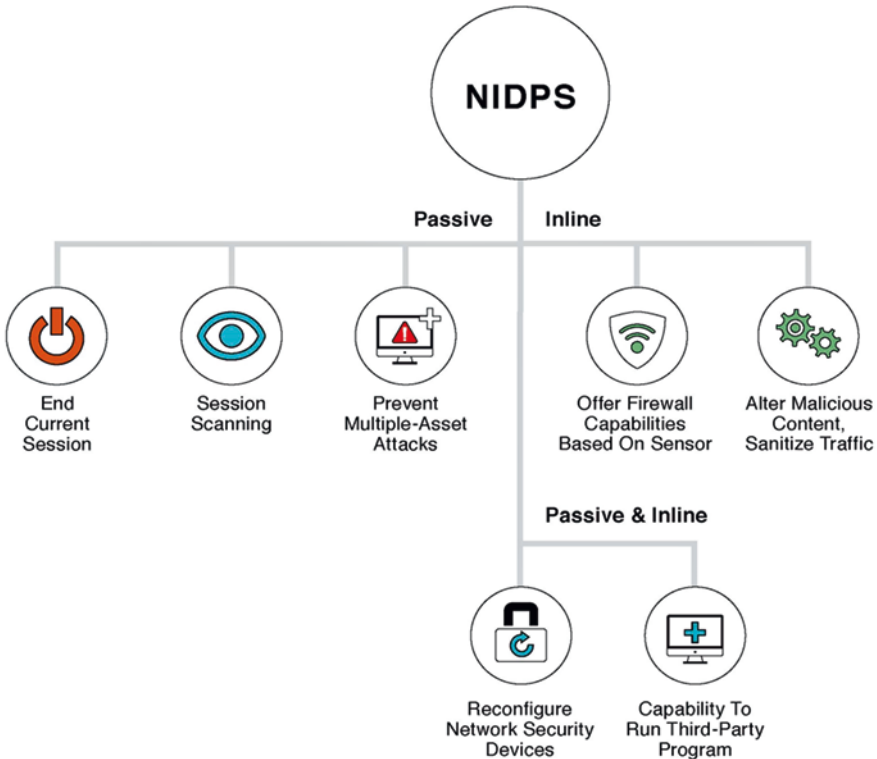
Fig. 12.4 Various functions of host intrusion detection and protection system (HIDPS)

implementation, eliminate false positives, and can create an adaptable baseline to automatically detect and update a system’s baselines.

To make the system more autonomous, this AI network can be designed to:

- Determine strategic locations on the network and identify critical hosts that need to be connected to a detection and prevention system.
- Identify false positives and filter the case logs and network logs, minimizing the detection data that needs to be reviewed upon a cyber-attack or alarm.
- Constantly update the baseline based on the behavior, changes, and message exchanges among assets over time.

Making the IDPS “intelligent” could help mitigate a number of evolving cyber threats and vulnerabilities in smart buildings and other complex cyber-physical environments.



**Fig. 12.5** Various functions of network intrusion detection and protection system (NIDPS)

### ***12.5.4 AI Enabled Cyber Incident Response***

AI enabled incident response platforms could help improve response to a security breach or attack. Current detection systems and operators enforcing every various protection schemes are challenged not only by a complex and evolving cyber-attacks threat, but also due to false positives, poor configuration of security architectures and lack of fidelity to anomalies created by insider threats. Figure 12.6 highlights response planning and implementation imperatives that could potentially be improved by AI advances.

AI based response systems could provide a more dynamic and effective response plans to enable buildings and critical infrastructure to automatically respond to complex cyber events and and:

- Identify the affected network-connected assets in and assess the cyber-event quickly to minimize damage
- Enable automation and control systems to make effective post-cyber event containment decisions



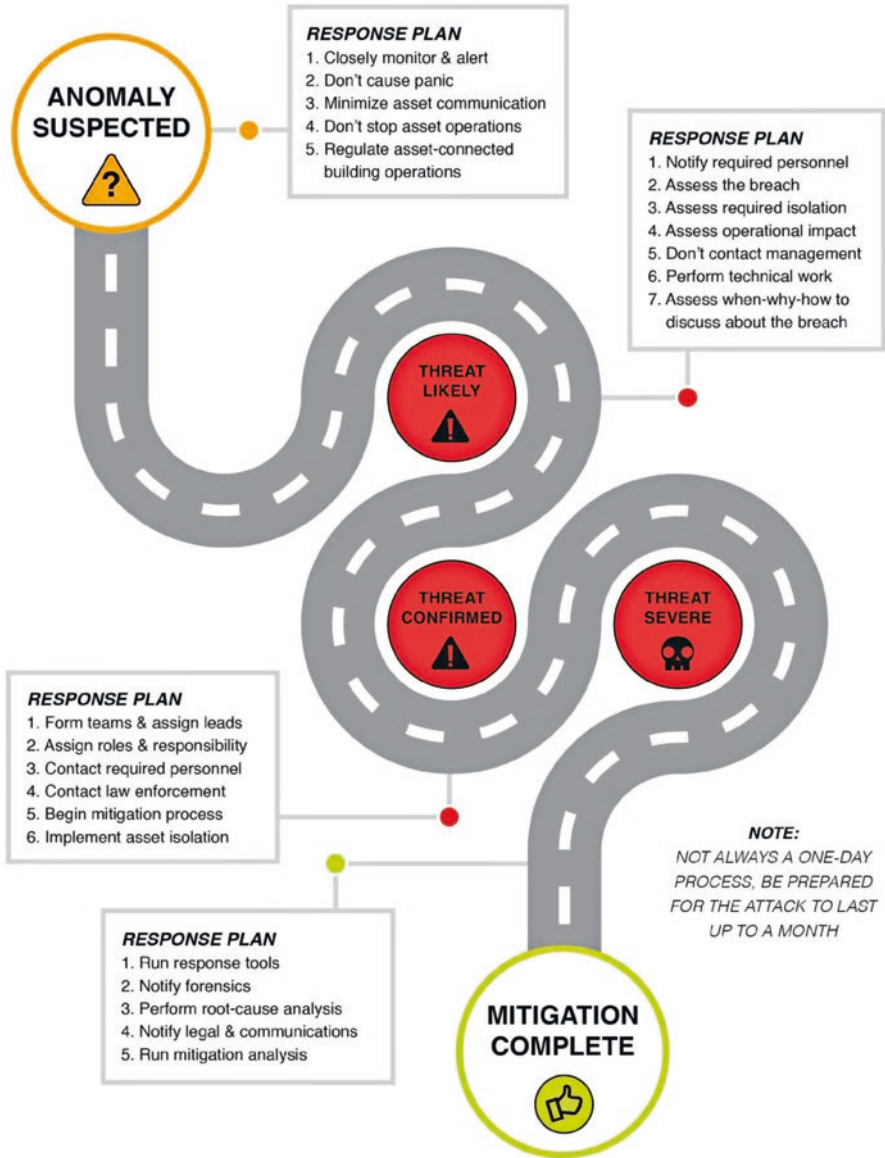


Fig. 12.6 Response planning and implementation from anomaly detection to completion of mitigation processes

- Alert required personnel inside and outside of the building that could be affected by this cyber-event
- Generate and log a set of guidelines to document all of the protocols and regulations about storing/sharing any information about the event.

#### **12.5.4.1 An Autonomous AI Cybersecurity Response System: *Theoretical Concept***

Increasingly the autonomy and resilience of buildings' OT and IT is imperative in response to an evolving cyber threat. An AI-based cyber response system can be connected to the network to help implement a more automated response plan (Search Security 2005). The first step to realizing this goal is for the system to monitor and learn all of the cyber and physical signatures of network-connected assets in the building. Then the system can start to establish baselines of normal operations and causation of anomalies from all hazards. Data aggregated from national threat and malware databases could help make the AI response platform more intelligent to different cyber threats. This would help streamline the complicated processes involved in developing an effective response plan to evolving threats such as polymorphic malware and distributed denial of service attacks. The AI response plan generator should have configuration options to either stay active 24/7, updating the response plan as needed, or to manually run the system once every few months to update the response plan as the building management and operators see fit. It should also allocate and update the "jump kit" a ready-to-go toolkit for emergency responses to cyber incidents.

The AI response plan generator will help automate and improve all of the steps shown in Fig. 12.6. These steps include forming a response team, and automatically coordinating all of the processes once a cyber breach is detected, and until the cyber-attack is mitigated.

### **12.5.5 AI Based Building Recovery System**

A successful cyber response can help provide a more rapid recovery of compromised critical cyber assets in smart buildings. The goal is to restore all the systems to a secure functioning state and remove any points of compromise. The implementation and complexity involved in restoring systems depends upon the extent of isolation executed, the impact and type of the attack, and the inter-dependencies of the assets. Therefore, recovery from a cyber-event requires significant planning that covers technical and non-technical aspects. Examples of technical aspects include the identification of techniques and tools to gather evidence and to determine an event's root causes. Examples of non-technical aspects include identifying roles and responsibilities for key personnel, and procedures for disclosing the incident to stakeholders and the media. These strategies should ideally be captured in a cyber incident recovery plan (CIRP), a document that will serve as a guide to asset owners

for continuous improvement of recovery practices before an incident, and as a play-book for post-incident active recovery.

### ***12.5.6 AI Based Building Recovery System: Theoretical Concept***

An automated AI based recovery system could be triggered upon responding to the cyber event successfully. The AI-recovery system would execute a more dynamic and agile cyber incident recovery plan to initially perform system stability setup functions such as asset prioritization to recover and stabilize, determine and design security interdependency maps, restore and restructure the baselines for IDS, alert recovery personnel when needed for authentication and authorization purposes. After stabilization to pre-event functionality, the AI system could be enabled to execute post-event activities such as evidence gathering to train intelligent cyber incident recovery plan and protection systems, eradication of all traces of malware and loopholes that includes securing the communication channels and re-configuring the firewalls as needed, remediation and reintegration of assets concluding with automated information release to key stakeholders about the current state of the building.

## **12.6 Use Cases**

### ***12.6.1 AI to Mitigate Insider Threat: Cognitive Ubiquitous Sensing and Insider Threat***

Defending organizations against insider threats continues to be a one the most difficult cybersecurity challenges (CERT 2016). The following theoretical scenario highlights how AI enabled building control systems and sensors embedded throughout a smart building can help better detect and respond to insider threats by leveraging cognitive behavioral modeling and cross-check analysis to help eliminate false positives and narrow in on the threat agent (in this case a building employee).

In an AI-based building environment, each employee identification card is embedded with a RFID chip that performs simple functions such as monitoring the employee’s movements and locations and syncing with sensors that have the ability to monitor and track movements, employee interactions as well as body temperature and heat signatures. All of the real-time data collected by these sensors is collected and aggregated by an AI-based insider threat Analyzer (AITA). Smart buildings will need to be equipped with multiple segregated distributed sensor networks that constantly communicate with the AITA. The building’s cognitive sensor network observes all of the operations and employee movements in the building and it measures behavioral

factors such as stress levels. This information will be combined with the information the AITA receives from the IT and OT environment (including the information from an IDPS) to determine any suspicious and anomalous behavior by any individual in the building. The following are the critical functions of an AITA that will advance the state-of-the art insider threat detection and improve cybersecurity posture of building:

- Identifies all network-connected assets, along with their baselines and expected behaviors on the OT side of the building and collects that information.
- Constantly monitors all software-based IDPS to look for anomalous activities such as spam emails, suspicious email conversations, and employee interactions on the web.
- Continually collects information from the cognitive behavioral sensor network in the building (including from wall-mounted and employee ID sensors).
- Constantly verifies and updates protection schemes and detection strategies and taking note of any significant changes.
- Cross-checks the response and recovery plans and ensures these are ready to use under an imminent cyber-attack.

AITA will leverage all of the information above to perform complex data and information analysis to identify possible insider threats. Through collection and aggregation data from multiple sensor networks and nodes, AITA transforms big data into smart data and information into real-time intelligence to improve the state-of-the art in insider threat detection.

### ***12.6.2 AI Enabled Smart Buildings Cybersecurity and Business Optimization***

AI enabled smart buildings can help increase productivity, innovation and cyber-physical security within organizations by collecting and aggregating metadata from both cyber signatures and behavioral data to develop “a causal theory of social structure” (Pentland 2014). AI platforms informed by social physics and cyber meta data may enable smart buildings to better collect, share and optimize the innovation and productivity potential of organizations. For example, if an organization seeks to increase cross-functional interaction between technical R&D teams and management, sensors will help monitor and collect movement and communications patterns to provide automated recommendations for individuals that need to increase their collaboration. For example, an AI enabled communications platform would aggregate metadata from Outlook calendars and email traffic and physical interactions from employees’ RFID badges and recognize when middle staff are not interacting enough based on parameters set by the organization. The AI enabled platform would send an automated calendar invite or reminder to the individuals that need to communicate better. This platform could also recognize patterns that suggest bad cyber

vulnerabilities and respond instantly with the appropriate threat remediation schemes (Mylrea 2015).

These new opportunities and challenges raise a number of questions worthy of future research: Will smart automated buildings and processes take an increasingly important role in leading and shaping organizations? If so, how can they do so without widening the attack surface by integrating enterprise and building automation platforms? Will AI enabled smart building controls take the helm in leading future organizations? Artificial intelligence has already crossed critical thresholds such as self-learning and dynamic conservation. In the same way building controls increasingly collect, monitor, control, and direct the activities of large organizations. As behavioral researchers better understand what makes an organization successful, will they be able to program building controls to lead an organization (Pentland 2014)? What are the limitations of automated systems in complex, distributed and non-linear organizations that increasingly define our globalized economies? All of these questions present an interesting and timely avenue to examine issues at the nexus of cyber security, optimization and autonomous systems (Mylrea 2015).

### ***12.6.3 Uber for Cyber and Energy***

As we network and digitize our energy value chain, we continue to revolutionize how we generate, transmit and distribute energy. Today, our electricity infrastructure is increasingly distributed, using increasing amounts of renewable and clean energy as well two-way communications enhancing visibility and control between grid operators and consumers. This has helped give impetus to new opportunities such as “transactive energy” for managing the generation, consumption and flow of electric power based on market based constructs. The addition of AI and machine learning will help combine disparate data sets, including IT and OT data, building and grid telemetry, cognitive human behavioral data, and organizational data in order to increase the efficiency and sustainability of our energy value chain without compromising its security.

Making these large data sets smart, an AI enabled smart energy platform can make more intelligent decisions to more securely manage the complexities of a digitized energy value chain. Big data provides an opportunity for researchers to perform analysis and studies, but often this leads to incorrect or irrelevant conclusions. The AI network could be designed to filter and use the data aggregation to transform big data into reliable and actionable information, as shown in Fig. 12.7. By doing so, this central AI network adds a layer of intelligence to an IoT based smart building network to transform end-point smart sensors into intelligent sensors. The AI network could be designed to:

- Perform building power system activities such as transactive studies/analysis, energy management, and utilization studies.

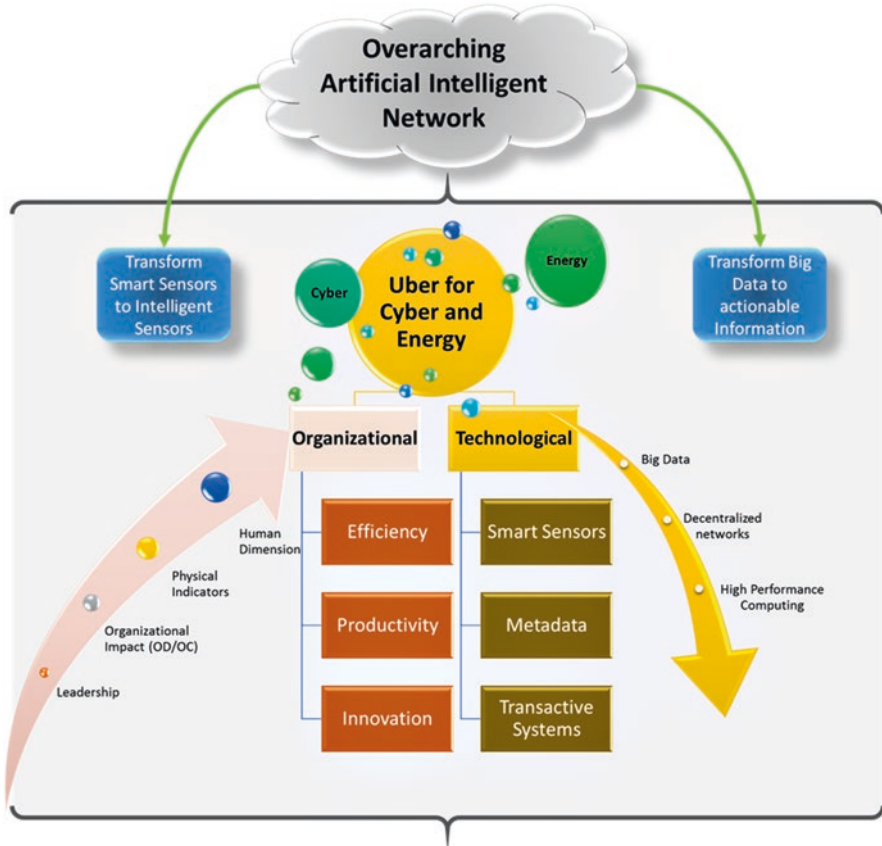


Fig. 12.7 Schematic showing Uber for cyber and energy

- Provide active defense in response to anomalies in both OT power system networks and IT communication networks.
- Implement cybersecurity best practices and mitigation schemes to protect critical cyber assets in the building and recover its systems, as needed, by eliminating the threat.

AI in smart energy technology will play an increasingly important role to buildings to grid modernization efforts. The smart grid will continue to have an integral and increasingly symbiotic relationship with buildings, which plays an increasingly important role for ancillary services, storage, DER deployment. Today, buildings account for nearly 75% of the nation’s electricity use and 65% of its load growth projected by EIA through 2040 (EIA 2016). In this grid 3.0 environment of ubiquitous sensing, distributed generation and networked operational technology, the size and speed of data collection is expected to increase exponentially. The cyber threat landscape also increases significantly, making control issues more important and complex, increasing their need for cybersecurity. The future of the smart grid with,

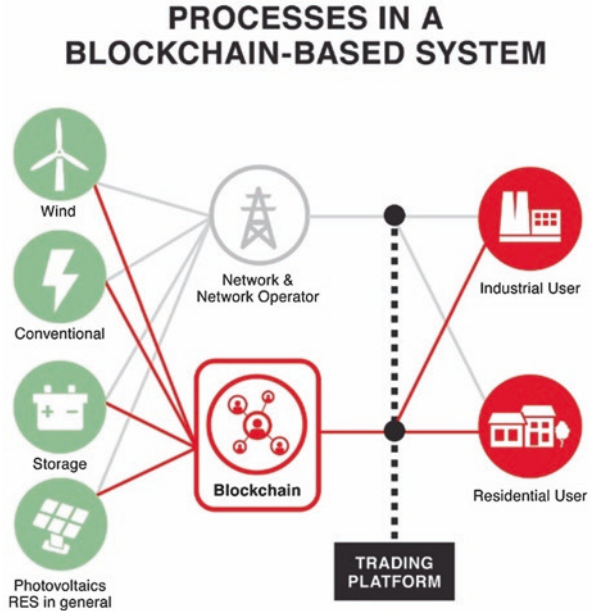
for example, two-way communication, demand response and transactive capabilities (Somasundaram et al. 2014) will require increased cybersecurity fidelity and security controls that can be enabled by AI advances. The owner of future transactive solutions may not be a major utility or even asset owner. A third party with big data analytic capabilities and AI enabled platform could potentially subsidize the installation and deployment of smart sensors and controls to better manage energy use and security in buildings. Both utilities and consumers would benefit from this type of Uber for Energy model enabled by AI advances in this space.

#### ***12.6.4 Blockchain for Power Grid Resilience: Exchanging Distributed Energy at Speed, Scale, Autonomy and Security***

Blockchain is defined as a distributed data base or digital ledger that records transactions of value using a cryptographic signature that is inherently resistant to modification (Tapscott and Tapscott 2016). Combining blockchain based smart contracts with machine learning algorithms presents an opportunity to increase the speed, scale, security and autonomy of transactive energy applications. These improvements present a more resilient path for a decentralized modern grid and integration of internet connected Energy Internet of Things (E-IoT) and grid edge devices (Mylrea and Gouresetti 2017). These grid optimization, automation and resilience improvements are essential operations and design criteria as we modernize our power grid. However, cybersecurity is often an afterthought as vendors and end users prioritize functionality and cost, leaving our power grid, the backbone of our economy, potentially vulnerable to a cyber-attack. This is especially true at the grid’s edge which continues to increase the size and speed of data being collected and exchanged in absence of clear cybersecurity and IoT standards and regulation. Thus, the grid lacks the necessary defenses to prevent disruption and manipulation of DERs, grid edge devices and associated electricity infrastructure. Moreover, as the smart grid increases its connectivity and communications with buildings, cyber vulnerabilities will extend behind the meter into “smart” buildings, which also have a host of cybersecurity vulnerabilities.

Blockchain technology can also be applied to the smart grid to help reduce costs by cutting out third parties and increasing the arbitrage opportunity for individuals to produce and sell energy to each other. Smart contracts facilitate peer-to-peer energy exchanges by enabling energy consumers and procurers to sell to each other, instead of transacting through a multi-tiered system, in which distribution and transmission system operators, power producers, and suppliers transact on various levels. In April 2016, one of the first use cases was demonstrated where energy generated in a decentralized fashion was sold directly between neighbors in New York via a blockchain system, demonstrating that energy producers and energy consumers could execute energy supply contracts without involving a third-party intermediary;

**Fig. 12.8** Role of blockchain in energy sector



effectively increasing speed and reducing costs of the transaction (PWC 2017). In addition to potential cost savings, transaction data might be more secure through decentralized storage and multifactor verification of transactions in the blockchain distributed ledger (PWC 2017). Figures 12.8 and 12.9 highlight how blockchain reduces the need for third parties to process transactions: Electricity is generated → Consumer buys the electricity → blockchain based meters update the blockchain, creating a unique timestamped block for verification in a distributed ledger: (1) At the distribution level, system operators can leverage the blockchain to receive energy transaction data to charge their network costs to consumers; (2) Reduces data requirements and increases speed of clearing transactions for transmission system operators as transactions could be executed and settled on the basis of actual consumption.

Smart contracts execute and record transaction in the blockchain load ledger through blockchain enabled advanced metering infrastructure (AMI). Blockchain based smart contracts can facilitate consumer level exchange of excess generation from DERs, EVs, etc. This could provide additional storage and help substation load balancing from bulk energy systems. Moreover, smart contract data is secured in part through decentralized storage of all transactions of energy flows and business activities. This highlights the disruptive potential for blockchain on energy markets through the introduction of a more autonomous and decentralized transaction model. This peer to peer system may reduce or even replace the need for a meter operator if the meter blockchain is shared with the distribution system operator.

New blockchain opportunities, however, are also accompanied by new challenges. For one, blockchain policies and regulations need to be in place to help



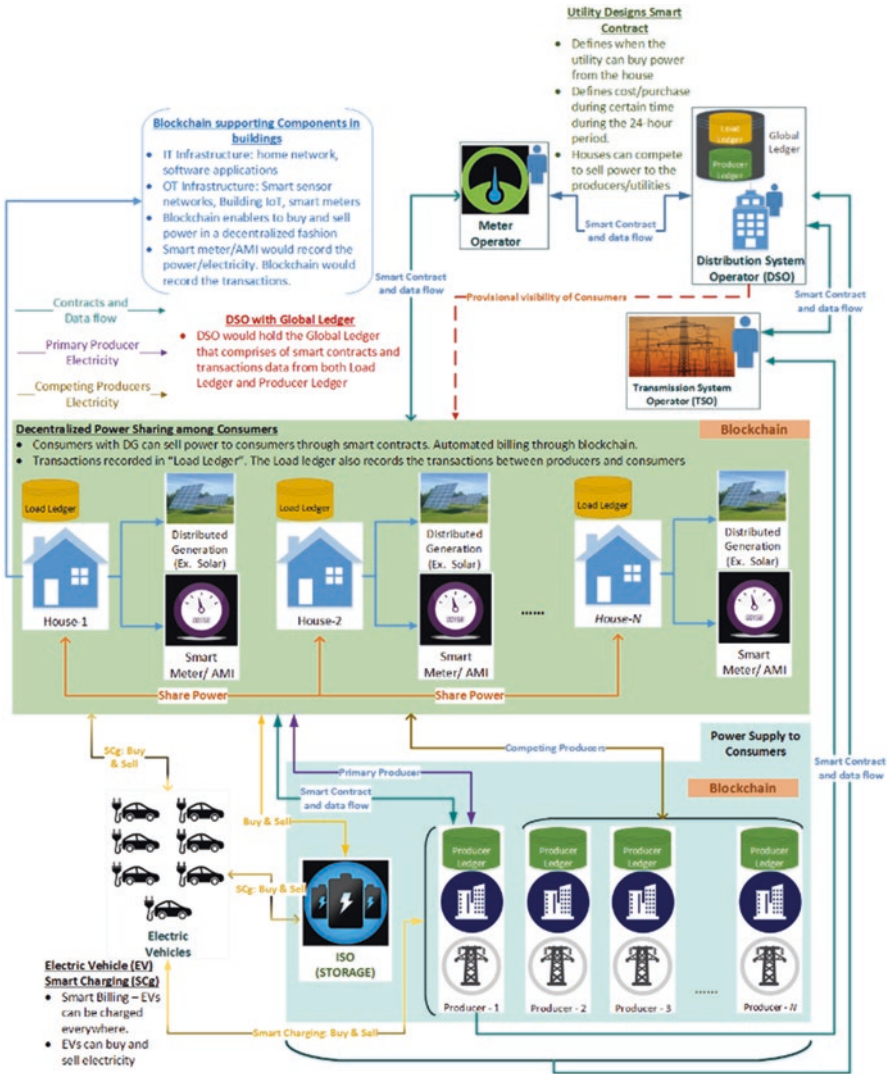


Fig. 12.9 Blockchain application to the electricity infrastructure

determine licensing and other key roles for energy companies. For example, there still needs to be schedule and forecast submitted to the transmission system operator. Another challenge is incorporating individual blockchain consumers into a balancing group and having them comply with market reliability and requirements and submit accurate demand forecasts to the network operator. Managing a balancing group is not a trivial task and could potentially increase costs of managing the blockchain. To avoid costly disruptions, blockchain autonomous data exchanges,

such as demand forecasts from the consumer to the network operator will need to be stress tested for security and reliability before deployed at scale.

### ***12.6.5 Social Engineering Autonomy for Cyber Intrusion Monitoring and Real-Time Anomaly Detecting (SCI-RAD)***

Social engineering continues to be one of the most challenging cyber threats (Allen 2006). One of the major challenges is detecting intruders that attempt to infiltrate by using social engineering (Lord 2016; Lord and Digital Guardian 2016). While several organizations provide training to their employees to be wary of social engineering-based scams and intrusions (Wombat Security 2016; Alexander 2016), there is a major gap in the ability to automatically detect a social engineering attack and provide a more active defense. SCI-RAD is an AI-based software tool concept that would constantly run a “watcher” on the back-end of a phone or web-connected computer. SCI-RAD would be securely connected to a database with thousands of pre-determined phrases that could potentially flag if a social engineering attack is taking place. A dynamic database would be updated along with an option to set control and definition settings manually. When an employee is active on a phone call, SCI-RAD monitors the phone call while looking for suspicious phrases. SCI-RAD does not intrude on privacy, since none of the monitored information is permanently recorded (monitoring can be adjusted, however, based on the policies of an organization). Upon detecting a pattern of suspicious phrases, SCI-RAD would alert the employee with an alert level of green (watchful—possible intruder), orange (critical warning—high probability of an intruder), or red (confirmed detection of an intruder). Recurring alerts by other personnel, such as administrators or a building’s owners can be adjusted based on the needs of an organization and its policies.

## **12.7 Conclusion and Future Research**

As cyber adversaries continue to evolve their tactics, techniques and tools to exploit new targets, smart buildings and building automation systems are increasingly in hackers’ cross hairs. Smart buildings present target rich environments and a cyber-physical attack landscape that includes an expanding array of things (i.e. EIoT, IIoT, IoT, etc.). Even as cyber defenses evolve so do hackers offensive capabilities. For example, hackers are adopting AI enabled cyber-attack tools are being used to target everything from building controls systems to energy management systems in smart buildings. AI enabled cyber payloads are very tough to defend against: polymorphic malware can adapt to defenses, smart AI enabled bots can scan software for new vulnerabilities and learn to exploit them against multiple targets simultaneously,

smart automated scanning tools probe with machine efficiency until they find a vulnerability to exploit, etc. While AI and automation has also improved cyber defenses, the attack landscape continues to expand leaving defenders at a disadvantage. A strong dynamic cyber defense can protect a target against millions of attacks a day, but it only takes one successful attack for an adversary to win. As of the time of this publication, AI appears to be tipping the scale to advantage of well-resourced advanced persistent threats. While improvements in cyber defense gained by AI adoption has helped reduce human error and some of the associated vulnerabilities, in part, by taking humans out of the loop, it also creates new cyber challenges that were addressed throughout this chapter.

For one, automation and AI often times require the exchange of larger data sets that can create new cyber vulnerabilities. This can be seen in AI enabled “smart” energy technology that is being rapidly deployed to make critical infrastructures, from manufacturing to power grids, smart cities to smart buildings, more efficient. Second, the increase in speed at which data is being collected, exchanged and aggregated creates new vulnerabilities. Thus, Moore’s law is playing out to hackers’ advantage in that as we make our infrastructures and technology “smart,” and data processing and storage costs fall, we increase the size and speeds at which process and store data without increasing necessary cyber defenses.

Even as cyber defenses improve, sophisticated threats can exploit cyber defenses (i.e. firewalls, VPNs, cybersecurity personnel, alarms, access control systems etc.) to carry out an attack. Unlike the physical defenses that once protected physical structures (i.e. moats, draw bridges, impenetrable walls), cyber offensive tools can exploit defenses to their advantage. Defending against this increasingly dynamic threat requires organizations to move beyond manual efforts, take humans out of some of the decision loops and into new ones, while continuing to leverage new artificial intelligence tactics, tools and technology to better manage cyber risk. In realization of this goal, this chapter proposed a number of AI cybersecurity concepts and solutions, from a more agile cyber-physical intrusion detection system to smart buildings deployed with ubiquitous sensors and machine learning algorithms to better detect insider threats, blockchain based smart contracts to protect the integrity of more distributed energy transactions to AI enabled business optimization opportunities in smart buildings.

This chapter also explored how AI enabled buildings and smart energy technology can provide more flexible learning systems to accelerate the intelligence, awareness and active defenses of critical cyber assets and systems in smart buildings. AI enabled smart building solutions can help organizations respond to evolving cyber-physical threats and vulnerabilities. AI based cybersecurity systems provide more robust, agile and autonomous defenses to evolving cyber threats. This chapter also highlighted a number of case studies that examined how combining AI (informed by cognitive behavioral sciences and organizational development) with innovative smart-energy technologies can increase both cybersecurity and energy efficiency in smart buildings. While a number of theoretical and applied solutions were given, certainly this area requires additional future research; exploration and application of AI enabled technology at the energy cyber nexus.

This research is timely as critical sectors from energy to manufacturing, health to defense embrace advances in AI and automation. As data sets and exchanges of information increase in size and speed, connecting our cyber and physical systems to the Internet, our smart buildings and infrastructures will increasingly become distributed IoT environments that require more autonomous, resilient and secure cybersecurity solutions. While this study focused largely on how AI could change the smart buildings and cybersecurity systems, future research should also take a closer look at some of the related human factors, such as: how can AI enabled smart buildings optimize energy management, while reducing cyber vulnerabilities?; how could big data sets be displayed to empower cyber defenders of critical infrastructures?; how can AI enabled cyber risk management algorithms facilitate risk quantification and response in a resource strained environment?; what is the best way for AI enabled cybersecurity systems to take humans out of the loop to reduce cyber vulnerabilities (Mylrea 2015).

## References

- Alexander M, SANS (2016) Methods for Understanding and Reducing Social Engineering Attacks. <https://www.sans.org/reading-room/whitepapers/critical/methods-understanding-reducing-social-engineering-attacks-36972>
- Allen M, SANS (2006) Social Engineering: A Means to Violate a Computer System. <https://www.sans.org/reading-room/whitepapers/engineering/social-engineering-means-violate-computer-system-529>
- Automated Buildings, AutomatedBuildings.com (2014) Innovations in Comfort, Efficiency and Safety, Solutions. <http://www.automatedbuildings.com/news/jun14/interviews/140528015505petock.html>
- BCF, Buildings Cybersecurity Framework (2016). Forthcoming publication by the U.S. Department of Energy's Building Technology Office.
- CERT., Cert.org (2016) insider threat. <https://www.cert.org/insider-threat/>
- DOE/EIA (2015) Annual Energy Outlook 2015 with projections to 2040. [https://www.eia.gov/outlooks/aeo/pdf/0383\(2015\).pdf](https://www.eia.gov/outlooks/aeo/pdf/0383(2015).pdf)
- Gartner, Inc. (2015) Gartner Says 6.4 Billion Connected "Things" Will Be in Use in 2016, Up 30 Percent From 2015. <http://www.gartner.com/newsroom/id/3165317>
- Hagerman J (2016) The National Opportunity to Secure Buildings and Facilities from Emerging Cyber Threats. Forthcoming White Paper to be published by U.S. Department of Energy, Buildings Technology Office.
- Hardin DB, Corbin CD, Stephan EG, Widergren SE, Wang W (2015) Buildings Interoperability Landscape (No. PNNL-25124), Pacific Northwest National Laboratory (PNNL), Richland, WA. [http://www.pnnl.gov/main/publications/external/technical\\_reports/PNNL-25124.pdf](http://www.pnnl.gov/main/publications/external/technical_reports/PNNL-25124.pdf)
- HP News (2014) HP Study Reveals 70 Percent of Internet of Things Devices Vulnerable to Attack. <http://www8.hp.com/us/en/hp-news/press-release.html?id=1744676#.V41Wm01f3X6>
- Ionesco P, IBM X-Force (2016) Research Penetration testing a building automation system. Is your smart office creating backdoors for hackers? <https://securityintelligence.com/is-your-smart-office-creating-backdoors-for-cybercriminals/>
- Kim E (2016) The people you trust most could be planning the next big cyber attack on your company. <http://www.businessinsider.com/ibm-report-says-majority-of-cyber-attacks-at-companies-involve-insiders-2016-6>
- Lord N (2016) Social Engineering Attacks: Common Techniques & How to Prevent an Attack. <https://digitalguardian.com/blog/social-engineering-attacks-common-techniques-how-prevent-attack>

- Lord N, Digital Guardian (2016) The History of Data Breaches. <https://digitalguardian.com/blog/history-data-breaches>
- Marr B (2016) What is the Difference Between Artificial Intelligence and Machine Learning? <http://www.forbes.com/sites/bernardmarr/2016/12/06/what-is-the-difference-between-artificial-intelligence-and-machine-learning/#220efb3a687c>
- Martin C (2016) Intrusion Detection and Prevention Systems in the Industrial Automation and Control Systems Environment. <http://docplayer.net/6290577-Intrusion-detection-and-prevention-systems-in-the-industrial-automation-and-control-systems-environment.html>
- Mylrea M (2015) Cyber Security and Optimization in Smart “Autonomous” Buildings. In: 2015 AAAI Spring Symposium Series.
- Mylrea, M (2016) Energy Security 3.0: The Next Generation of Energy Wars and Diplomacy. U.S. Department of State, Ralph Bunch Library Speaker Series Lecture.
- Mylrea, M, Gouresetti, S (2017) Applying Blockchain Based Smart Contracts to Grid Modernization: A Path to Speed, Scale and Security at the Grid’s Edge. IEEE Resilience Week Publication. Forthcoming, September, 2017
- ICS-CERT, NCCIC/ICS-CERT Year in Review (2015). [https://ics-cert.us-cert.gov/sites/default/files/Annual\\_Reports/Year\\_in\\_Review\\_FY2015\\_Final\\_S508C.pdf](https://ics-cert.us-cert.gov/sites/default/files/Annual_Reports/Year_in_Review_FY2015_Final_S508C.pdf)
- NIST (2014) Framework for Improving Critical Infrastructure Cybersecurity, Version 1.0. <http://www.nist.gov/cyberframework/upload/cybersecurity-framework-021214.pdf>
- O’Harrow R, The Washington Post (2012) Cyber Search Engine Shodan Exposes Industrial Control Systems to New Risks. [https://www.washingtonpost.com/investigations/cyber-search-engine-exposes-vulnerabilities/2012/06/03/gJQAIK9KCV\\_story.html](https://www.washingtonpost.com/investigations/cyber-search-engine-exposes-vulnerabilities/2012/06/03/gJQAIK9KCV_story.html)
- Pentland A (2014) Social Physics: How Good Ideas Spread-The Lessons from a New Science – a textbook, Penguin.
- PNNL (2012) Grid Friendly Appliance Controller. <http://availabletechnologies.pnnl.gov/technology.asp?id=61>, [http://availabletechnologies.pnnl.gov/PDF/AT\\_61.pdf](http://availabletechnologies.pnnl.gov/PDF/AT_61.pdf)
- PNNL (2016) Buildings Cybersecurity Compatibility Maturity Model. <https://bc2m2.pnnl.gov/>
- Pullen D (2014). Smart Buildings Research for the Future. Science in Parliament
- Radvanovsky B, Tofino Blog. (2013) Project SHINE: 1,000,000 Internet-Connected SCADA and ICS Systems and Counting. <https://www.tofinosecurity.com/blog/project-shine-1000000-internet-connected-scada-and-ics-systems-and-counting>
- Search Security, Searchsecurity.com (2005) Definition incident response. <http://searchsecurity.techtarget.com/definition/incident-response>
- Security Week (2014) Target HVAC Contractor Says It Was Breached by Hackers. <http://www.securityweek.com/target-hvac-contractor-says-it-was-breached-hackers>
- Somasundaram S, Pratt RG, Katipamula S, Mayhorn ET, Akyol BA, Somani A, Fernandez N, Steckley A, Foster N, Taylor ZT (2014) Transaction-Based Building Controls Framework, Volume 1: Reference Guide. PNNL-23302, Pacific Northwest National Laboratory, Richland, WA. [http://www.pnnl.gov/main/publications/external/technical\\_reports/PNNL-23302.pdf](http://www.pnnl.gov/main/publications/external/technical_reports/PNNL-23302.pdf)
- D. Tapscott, A. Tapscott (2016), The Blockchain Revolution: How the Technology Behind Bitcoin is Changing Money, Business, and the World
- PWC Global Power and Utilities (2017) Blockchain opportunity for energy producers and consumers
- GAO, The U.S. Government Accountability Office (2014) Federal Facility Cybersecurity DHS and GSA Should Address Cyber Risk to Building and Access Control Systems. [www.gao.gov/assets/670/667512.pdf](http://www.gao.gov/assets/670/667512.pdf)
- Towler J (2015) World Building Automation & Control Systems Market expected to be worth just over US\$26 bn by 2019. <https://www.bsria.co.uk/news/article/world-building-automation-control-systems-market-expected-to-be-worth-just-over-us26-bn-by-2019/>
- DHS, U.S. Department of Homeland Security. (2016) Critical Infrastructure Sectors. <https://www.dhs.gov/critical-infrastructure-sectors>
- EIA, U.S. Energy Information Administration (2016) International Energy Outlook 2016. [http://www.eia.gov/outlooks/ieo/pdf/0484\(2016\).pdf](http://www.eia.gov/outlooks/ieo/pdf/0484(2016).pdf), <http://www.eia.gov/outlooks/ieo/>

- Verizon (2016) Data Breach Investigations Report.. <http://www.verizonenterprise.com/verizon-insights-lab/dbir/2016/>
- Wombat Security, Wombat security Social Engineering (2016) Teaching Users to Recognize and Avoid Social Engineering Scams. <https://www.wombatsecurity.com/suggested-programs/social-engineering>
- Wueest C, Symantec (2015) Is IoT in the Smart Home giving away the keys to your kingdom? <http://www.symantec.com/connect/blogs/iot-smart-home-giving-away-keys-your-kingdom>

# Chapter 13

## Evaluations: Autonomy and Artificial Intelligence: A Threat or Savior?

W.F. Lawless and Donald A. Sofge

### 13.1 Introduction

AI has been able to defeat humans playing its most challenging games (e.g., chess; Go; poker; in Lien 2016) and outperformed human decision making even in medicine (e.g., AI diagnoses of pulmonary hypertension are better than those by cardiologists; in Austin 2017). But Bill Gates and other technology thought-leaders have worried that super-intelligent AI threatens humanity (Mamiit 2015). However, AI has not been able to satisfactorily model the human-human interaction (Lawless 2017). Instead, AI theorists are convinced that intuition produces “a realistic model of physics” except when “its predictions can deviate from objective reality” (Battaglia et al. 2013). Bacharach (2006, p. 44), the game theorist, supported intuition’s value; however, Kahneman (2002) countered that intuition fails under uncertainty; Simon (1978, p. 367) added that intuition fails when faced by complexity. From our research (Lawless 2017), intuition is limited when movement occurs, as in the human-human interaction. For movement, humans adopt a crude physics similar to what existed three centuries before Isaac Newton (e.g., McCloskey 1983). Movement, as in an interaction, however, is critical to the discovery of physical reality (e.g., Laudisa and Rovelli 2013), including, we argue, human nature, its measurement and its interactions with machines and robots (Lawless 2017).

Other than intuition, little is accepted about what makes humans human; one idea often repeated is the theory of the human mind as a way for humans to communicate (Bekoff 2011), viz., Shannon’s mutual information or interdependence

---

W.F. Lawless (✉)

Departments of Psychology and Mathematics, Paine College, Augusta, GA 30901, USA  
e-mail: [w.lawless@icloud.com](mailto:w.lawless@icloud.com)

D.A. Sofge

Distributed Autonomous Systems Group, Navy Center for Applied Research in Artificial Intelligence, Naval Research Laboratory, Washington, DC 20375, USA  
e-mail: [donald.sofge@nrl.navy.mil](mailto:donald.sofge@nrl.navy.mil)

(Conant 1976). But Arrow (1951, 1963, p. 9) considered interpersonal preferences to be meaningless; indeed, based on individual preferences, Lucas, Nobel economist and president of the American Economic Association, stated in 2003 that macroeconomics had matured sufficiently to prevent another economic calamity, a misjudgment followed by the Great Recession in 2007 (Lanchester 2017). Similarly, based on the individual, Pfeffer and Fong (2005) concluded that organizational theory could not justify theoretically the existence of the organization. From the self-reports by individuals judging their own reality, individuals have a poor grasp of their own actual behaviors (Zell and Krizan 2014). Yet, compounding the failure to replicate important social science research (Nosek 2015), most of social science is focused on the individual, including in economics (Ahdieh 2009). For example, experimental social psychologists recommend that the effects of interdependence be statistically removed (Kenny et al. 1998).

Traditionally, teams have been organized around a division of labor, negating the benefits from exploiting interdependence with multitasking (MT; in Bartel et al. 2013). Individuals are poor at MT (Otto and Sentana 2015), the function of teams (Lawless 2017); e.g., like the independent roles for the players who MT when playing together as a baseball team. Interdependence governs the dynamics of interaction, teams and society (Cooke and Hilton 2015), but it is a difficult concept to grasp theoretically and mathematically. Based on our research, interdependence creates a bistable state (Lawless 2017) where two or more individuals can hold incommensurable interpretations of a single social reality simultaneously (e.g., Republicans versus Democrats; prosecutors versus defense attorneys; Einstein's interpretation of quantum reality versus Bohr's); forces the convergence of concepts to align into a single interpretation of social reality that is always incomplete (e.g., the first forecast by Tetlock & Gardner's superforecasters converged into the failed prediction that Brexit would be approved by UK voters; see Kennedy 2016); and generates states of uncertainty, the more equal or opposed are those who interact (divorces; business spin-offs; political contests; scientific conflicts; tribal war; e.g., Chagnon 1988). The existence of interdependent states led us to conclude that in the affairs of humans, whether in government, science, courtrooms and on, the best approach to determine social reality under uncertainty is not with Simon's (1989) bounded rationality for an individual human or single robot, but with a team as the fundamental social unit.

With our model of interdependence (Lawless 2017), we have concluded that the skills needed to grasp social reality is unlikely for individuals or single robots acting independently, reducing the likelihood of autonomy and of reaching maximum productivity. Individuals are poor at multitasking (MT; in Wickens 1992); but by forming a team to MT, a team of bistable agents is more effective than the same agents acting independently to perform the same tasks (Lawless 2017). For a team of bistable agents, however, a team's convergence into its biased interpretation of reality reduces its autonomy. Instead, autonomy requires a MT team observing *social reality* to be contradicted by an opposing team. Two teams of MTs best capture social reality (Smallman 2012); underscoring the value of competition in determining reality, but these two teams may still be insufficient for autonomy. Full autonomy



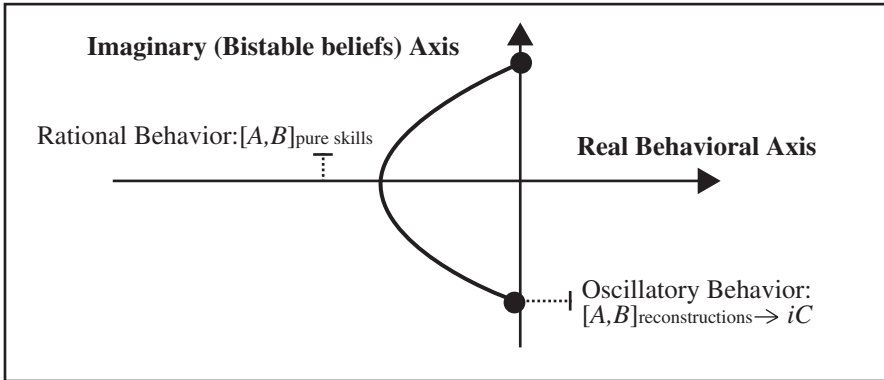
requires three teams: two opposing, well-structured MT teams that drive their constructions of reality in an attempt to attract members of an amorphous team of fluid bistable neutral agents, interdependently entangled with both teams sufficient to determine the team with the best grasp of reality. Thus, given two competitive teams probing for vulnerabilities in each other, adding a third, but loose team that constitutes a spectrum of neutral bistable agents able to invest freely (properties, ideas, works), act freely (joining and rejecting either team), and observe freely makes the greatest contribution to autonomy, to mitigating mistakes, to innovation and to improving social welfare.

The effects from interdependence are indirect but observable (Lawless 2017). For example, contradicting traditional social network theorists (Centola and Macy 2007, p. 716), we predicted that redundancy in teams under the influence of interdependence is minimized to maintain the lines of communication among teammates as necessary to MT, producing a big effect when comparing teams operating under the fewer political constraints existing in a democracy versus a dictatorship. That is what we found when we looked at the top oil firms from around the world; e.g., Sinopec oil company uses about 548 thousand employees to produce about 4.4 million barrels of oil per day whereas Exxon uses about 82 thousand employees to produce about 5.3 million barrels of oil per day. Political effects can also impede businesses: it “takes VW twice as many workers to build a car as it does Toyota” (Jenkins 2017); and in war, it is not unusual for an outnumbered commander to win in battle (e.g., As McMaster won in the Gulf war against much larger forces; in R&O 2017). More to the point, Cummings (2015) found that the more interdisciplinary a scientific team was, the more its scientific performance was *reduced*, indicating team miss-fits. Yet, controlling for interdisciplinarity, Cummings found that the top scientific teams were highly interdependent. These results indicate that there is a positive relationship between a team’s fitness for performance and the minimum number of constituent teammates, without impairing its interdependence.

### ***13.1.1 Mathematical Model of Autonomy: Entropy of Teamwork***

From Ambrose (2001), teams form to solve the problems that an arbitrary collection of individuals performing the same actions are ineffective at solving, like MT in competitive or hostile environments. Firms form to produce a profit (Coase 1937); generalizing, teams or firms succeed when they produce more benefits than costs (Coase 1960).

The mathematics that follow may be counterintuitive to a general understanding because humans tend to think rationally (intuitively). We combine quantum mathematics (matrix algebra of two community operators that convert into Fourier pairs to account for the incompleteness of situational awareness; from Cohen 1995); uncertainty relations (information flow in orthogonal models of teams; Lawless 2017) and



**Fig. 13.1** The horizontal axis displays real behavior while the vertical (imaginary) axis displays the social construction of reality. The two end points for imagined (subjective) beliefs reflect oscillatory dynamics (e.g., in 2005, the Nuclear Regulatory Commission and the Department of Energy’s High-Level radioactive Waste tank closures from about 2007–2011 led to endless debates between the two federal agencies until citizens recommended that the tanks be closed; in Lawless et al. 2014). Rational behavior produces no oscillations; the curve between the imaginary and rational axes reflects dynamics increasingly dampened as the real axis is approached (i.e., when opposing teams are more likely to agree)

biology (the movement of individuals between different teams can be tracked with limit cycles in Lotka-Volterra type equations; from May 1973). The results are metrics that, in the limit, represent Least Entropy Production (LEP; see Nicolis and Prigogine 1989) for team structure, and Maximum Entropy Production (MEP; Martyushev 2013) for team performance.

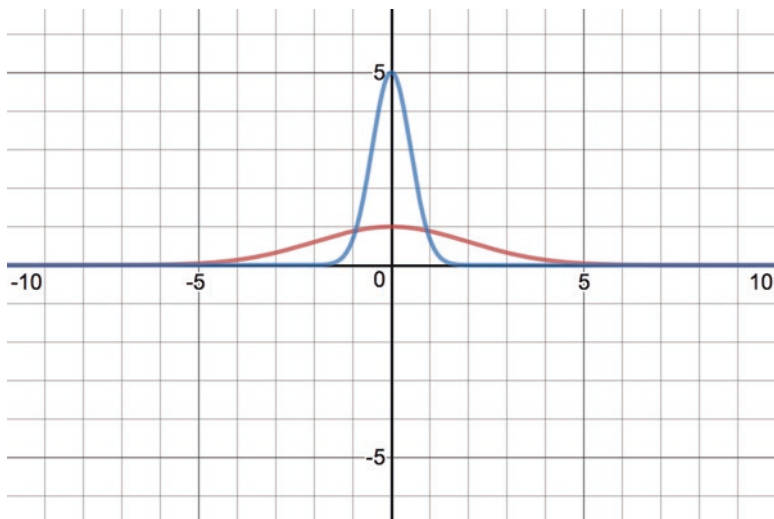
Given that observation and motor activities in the brain are controlled by independent systems (i.e., Rees et al. 1997), we assume that human behavior occurs in physical reality, while observations are reconstructed as interpretations (e.g., beliefs; situational awareness; illusions; or mistakes; e.g., Graziano 2013). We also assume that when two teams agree, no oscillations occur; but that when they disagree, oscillations occur (Fig. 13.1).

Bistability from interdependence occurs between competing claims as well as between actions and observations. When socially constructed reality is challenged by those with opposing interests, social dynamics occur (Lawless 2017).

We model bistability with signal detection theory (SDT) for two operators,  $A$  and  $B$ , to represent competing teams. When two erstwhile competitors agree, their combined social system is stable, no oscillations or limit cycles exist (the goal of an autocracy), and their operators commute:

$$[A,B] = AB - BA = 0 \quad (13.1)$$

But with disagreement between two competitors, operators do not commute (i.e., their eigenvalues are not equal), orthogonality exists between the two viewpoints, causing oscillations:



**Fig. 13.2** The wider Gaussian at the bottom is Fourier transformed to the narrower one. While the Standard deviation for the wider one is 5.0, that for the narrower one has reduced to about 0.33; the two multiplied together roughly constitute a constant value greater than 1/2

$$[A,B] = iC \tag{13.2}$$

where  $C$  measures the “gap” in reality between  $A$  and  $B$ . Based on Adelson’s (2000) work with illusions, we claim that humans cannot improve on signal detection theory (SDT) for sensory perceptions; e.g., humans easily misjudge Adelson’s checker-square illusion even though photometers do not. Cohen (1995, pp. 45–6) converted Eq. (13.2) into (13.3):

$$\sigma_A \sigma_B = \frac{1}{2} \tag{13.3}$$

With Eq. (13.3), Cohen concluded for SDT that a (see Fig. 13.2):

narrow waveform yields a wide spectrum, and a wide waveform yields a narrow spectrum and that both the time waveform and frequency spectrum cannot be made arbitrarily small simultaneously.

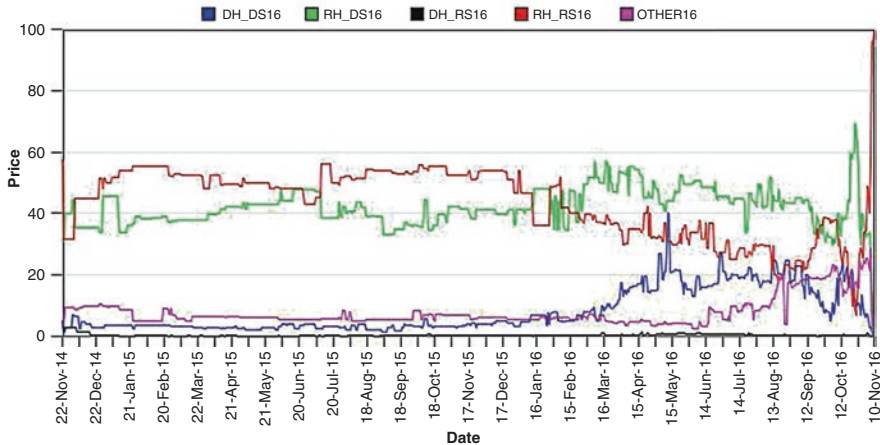
As a bistable example of skills coupled to awareness, Arthur Anderson, the auditor of Enron in 2000, missed Enron’s collapse. KPMG is accused of repeating Arthur Anderson’s mistake (Kowsmann et al. 2014): “KPMG faces criticism for Espírito Santo audit work. Bank’s collapse raises questions whether KPMG should have detected problems earlier.” Equation (13.3) shows that a focus on the wrong aspect of interdependence can account for why intuition fails.

For a team, as MT improves, the tradeoffs internal to each group’s focus on MT interferes with its bistable interpretation of how best to improve its performance, motivating tradeoffs, giving:

$$\sigma_{Skills} \sigma_{Interpretations} \geq \frac{1}{2} \tag{13.4}$$

where  $\sigma_{Skills}$  is the standard deviation of variable  $A$  over time,  $\sigma_{Interpretations}$  is the standard deviation of its Fourier transform, the two forming a Fourier pair that reflects tradeoffs between the physical expression of skills and the social interpretation of skill applications. For example, from Eq. (13.4), as uncertainty in a team’s or firm’s skills decrease (e.g., improved MT skills), uncertainty in the interpretations of skill efficacy increase (i.e., poorer situational awareness), often requiring that a team engage a relatively independent observer as a coach or consultant to help a team improve its performance.

Equation (13.2) captures disagreement in social processes, the result insufficient to determine the outcome of social dynamics. Generally, those individuals committed to their beliefs remain committed over time. If all individuals are committed to one side or the other, conflict ensues (Kirk 2003). We claim that neutrals enter into a state of interdependence with both sides, allowing them to process both sides of an issue. Neutrals moderate conflict and often decide elections (e.g., NYT 2010). If true, competition for neutrals generates limit cycles (e.g., Fig. 13.3).



**Fig. 13.3** Results of the 2014 race to control both Houses of the U.S. Congress. Notice the primary limit cycles (red and green curves) from about mid-November 2013 until about mid-July 2014. We claim that during this time, interdependence governed, making predictions unreliable. After neutrals had made their decisions, in this case, about mid-July 2014 onward, predictions became increasingly credible (chart from [https://iemweb.biz.uiowa.edu/graphs/graph\\_Congress14.cfm](https://iemweb.biz.uiowa.edu/graphs/graph_Congress14.cfm))

### 13.1.2 Entropy Production

Zipf (1949) concluded that “Frequent behaviors become quicker and easier to perform over time.” Zipf applies to teams, too. But in addition, interdependence represents a reduction in individualism (Kenny et al. 1998) that reduces the degrees of freedom (*dof*) in a social group (viz., team) as the team performs over time. Setting Boltzman’s constant  $k$  to 1 gives:

$$\log(dof_{\text{teammates}}) - \log(dof_{\Sigma \text{Individuals}}) \quad (13.5)$$

Balch (2000) used information to measure the entropy of multi-agent teams. Balch calculated that when three slaves form a unit,  $\log 3/3 = 0$ , and three independent individuals give an entropy of  $3 \times 1/3 \log 1/3 = 1.584$ . But Balch overlooked the interdependence involved in MT.

In contrast, using graph theory (Smith 2014), when a team of independent individuals interdependently completes a circuit for a team to MT, like the different roles played by the independent members of a baseball team (similarly for a team of autonomous multi-UAVs), LEP becomes the entropy produced by a team’s structure. Assuming that a set of tasks performed by the least number of individuals forms a complete circuit to MT, then the individuals become a team producing less entropy than the sum of the same individuals constituting the team.

We revise Eq. (13.3) to give us the standard deviation of LEP (structure) times the standard deviation of MEP (performance). As  $\sigma_{LEP} > 0$ , in the limit we find that  $\lim(\sigma_{MEP}) = \infty$ . As entropy produced by a team’s structure goes to zero, MEP for teamwork reaches a maximum. In other words, at MEP, the best teams are able to perform a maximum search of their environment for solutions to the difficult problems that they were designed to solve. For a perfect team acting interdependently to MT as a single unit, in the limit, its *dof* reduce to one, giving:

$$\log(\text{team's } dof) = \log(1) = 0. \quad (13.6)$$

Equation (13.6) accounts for the loss of information from perfect teams, modeled by subadditivity:

$$\rho_{AB} \leq \rho_B + \rho_B \quad (13.7)$$

Equation (13.7) explains why little information is emitted by perfect teams; why aggregated preferences, by providing no indication of a perfect team’s performance, are meaningless; but also why governments censor information from poorly performing teams, which we model next.

Reversing the limits, as  $\sigma_{LEP} > \infty$ ,  $\sigma_{MEP} > 0$ ; i.e., teamwork can become dysfunctional, possibly due to suppression, the zealous enforcement of consensus rules or authoritarianism. This result accounts for the Department of Energy’s use of ordinary cardboard boxes as its primary disposal container of solid radioactive wastes

until the whistle was blown on it in 1983 (Lawless et al. 2014); it accounts for the environmental problems in China today (Wong 2015); and it accounts for the inability of youth in gang-controlled areas to flourish in school (Howell 2006). As a simple test, using patent applications over the last 13 years (the data from USTPO 2013), we looked at Israel's applications filed in the US with whether or not it was experiencing an Intifada (-1), peace (0) or hostilities (+1), finding a significant correlation ( $r = 0.53$ ,  $p < 0.05$ , two-tailed test), suggesting that internal conflict like an Intifada reduces MEP.

Competition exposes vulnerabilities and strengths in teams, firms and organizations. Once exposed, mergers and spinoffs attempt to transform a poorly performing unit as teams seek to obtain sufficient energy to survive (e.g., GM selling Opel, in Ewing 2017). Team fit is bolstered by mergers in consolidating markets (e.g., the Bayer-Monsanto Deal; in Bunge 2016), and spinoffs in failing markets (e.g., Maersk; in Chopping 2016), implying that team fit obeys the second law of thermodynamics. For these mergers, overlapping employees are let go, reversing the effect for spinoffs; this model simulates water when it freezes, releasing heat, reversing when ice melts.

### 13.1.3 *Emotion*

In the limit, we assume that a perfect team is in a ground state; i.e., when no task conflict or role conflict exists, as it performs its tasks, a perfect team resides in its lowest emotional state.

On the other hand, if, for example, in a divorce, both partners are placed into an emotionally elevated or "excited state," MEP (productivity) reduces to zero. Working backwards from this result means that a team's LEP goes to a maximum as internal conflict forces a team's structure to splinter. We have associated team fragmentation with its loss of control (for example, by the New York Stock Exchange; in Hope 2014). In business, the result is like the 300 stores Sears is spinning off (Kapner and Dulaney 2014); in city government, it is a bankruptcy (e.g., Detroit; in EB 2014); or in Palestine, it is an internal battle for control between Hamas and Fatah (e.g., Casey 2014).

When interdependence is suppressed, individualism reigns, increasing joint entropy:

$$H(x,y) \geq H(x), H(y) \tag{13.8}$$

Comparing Eqs. (13.7) and (13.8), interdependence becomes a resource by generating less entropy. These two equations help us to see that organizational boundaries establish and maintain communication channels (e.g., exposed by bankruptcy; see RadioShack in Fitzgerald and Knutson 2017); reduce external interference to members performing tasks within communication channels; increase the likelihood that interdependence thrives; and help structures to maintain LEP.

### 13.1.4 Evaluations

We evaluate our Chap. 13, part 1 first, followed, on their own terms, by those chapters that dealt more with autonomy (Chaps. 2–5–7–8–10–11). Lastly, after reviewing and evaluating Chap. 13, part 2, we evaluate the rest of the chapters, those dealing primarily with errors (viz., human errors, AI errors, team errors).

Chapter 13, part 1: The mathematical models of interdependence we have presented in the first part of Chap. 13 approximate the interactions within and among teams found in the real world. Interdependence improves the performance of teams but at the cost of reducing information from them as performance increases. From what we have learned, the best situation for autonomy requires competitive checks and balances that limit autonomy to the tasks at hand but while also helping society to better grasp (social) reality.

One of our remaining tasks is to study Arrow's (1951/1963) conclusion that interpersonal preferences are meaningless; from a different perspective, with his impossibility theorem, Arrow proved that the preferences of three or more individuals could not be aggregated mathematically (accomplished in a democracy with a majority vote; in a socialist society by forcibly seeking consensus; or in a dictatorship by making unilateral decisions). Both of Arrow's ideas are supported by Eq. (13.7), which indicate a loss of information as a team's performance improves. But from Shannon, contradicting Arrow, team performance should improve as the mutual information goes to zero (Conant 1976, p. 248); however, Cummings (2015) found the opposite, that interdependence reaches a maximum with a top team's best performance, supporting Eq. (13.7). Interviewed about his 1974 book on the limits of organization, Arrow (1998) stated: "the purpose of organizations ... [requires] many individuals for their effectiveness ... which requires maximization information." We agree if Arrow meant maximum mutual information; from our research, however, maximum MT reflects a skills knowledge that generates minimum information (Conant 1976, p. 244). At this time, predicated on future research, we conclude tentatively that the claim of robots displacing "knowledge labor" (Aberman 2017) is overwrought.

Chapter 2 builds upon many of the criticisms that abound in the intelligence community about the lack of effective automation support for intelligence analysis, and offers a unique functional design for what is argued to be a new and hybrid approach to automated support with formal argumentation methods to combine hard sensor data and subjective data. The foundations of the suggested approach would have been better developed if experiments and associated prototype results had been available to the authors, but that there were none is a shortfall of not only this chapter but also, by implication, the existing research programs. Nonetheless, to the extent possible, the authors "made efforts to garner real-world viewpoints" for their design. They found that many experts and pundits in the intelligence field suggested that evidence-based argumentation provides a solid basis for analyses, motivating the need for composite sensor and textual data in modern analysis

environments. The approach presented by the authors in this chapter to combine both composite sensor and (subjective) textual data is fully sensible and plausible.

Autonomy may 1 day soon provide numerous social benefits. But for autonomous systems to provide these benefits, the authors in Chap. 5 argue that the verification of autonomous systems will become necessary to reduce the errors made by these systems. However, as the authors discovered in their review, today, several challenges for verification exist along with gaps in the very research programs needed to address these challenges across the discipline, which they explore, along with existing tools and the tools needed but not yet devised. In addition to tools, the authors address the numerous unexplored research challenges, including the metrics for verification that must be addressed for verification to become a satisfactory program.

A paradigm shift is occurring from multiple supervisors of a single UAV to a single supervisor of multiple UAVs (unmanned aerial vehicles). The authors of Chap. 7 provide a comprehensive overview of the state of supervisory control research to identify the gaps in the current research in order to assess the performance of the humans who will be using these new, highly automated systems. They have proposed a detailed approach for measuring human performance and the trust in autonomous systems that address these limitations across their discipline along with a proposal for their own test-bed to be used as a tool for human-automation experimentation and research. From their perspective, the goal of the test-bed that they have developed is to determine the limits of systems and to rigorously evaluate single human users of multi-UAVs by assessing each operator's states and the performance of the systems.

The ubiquity of autonomous systems entering society is on a near-term horizon, arriving much sooner than even just recently expected. With her practical, step-by-step approach that includes definitions, common examples of automation, autonomy and robot systems, to govern the new systems already here and those expected to arrive soon, the author of Chap. 8 provides an argument for the shared control of autonomous systems as a careful step forward at this time. For full autonomy, she also argues that there is a need for robots to learn an ethics of behavior that protects humans and the current and future well-being of society. This means that several research and legal challenges lie ahead before full autonomy can be let loose upon the world; e.g., researchers and legal experts need to think through what has to be done about a damaged UAV that is still flying and also still autonomous.

In Chap. 10, the author makes the claim that for a powerful AI to behave in a safe manner, it needs to have a rich understanding of the meaning of its instructions and the potential consequences of its actions across a wide range of topics, applications and behaviors. This chapter provides a discussion of the author's preferred architecture for enabling such a level of understanding for a system designed first to be safe. His proposed system consists of a knowledge graph embedded in a semantic vector space trained on large bodies of data. He outlines how deductive, analogical, and associational reasoning could take place in such a system; he also outlines a program of research towards building a practical version. His high-level approach attempts to get around some of the difficulties that presently plague



these knowledge bases. Reasoning systems based neither on common sense alone nor only on sophisticated vector spaces have worked satisfactorily; however, the author sees promise in a system that combines the best features of both to face the challenges ahead as “physical processes and human goals” interact to behave in unexpected ways in new environments.

In Chap. 11, the authors provide a baseline for the neuroscience of humans participating in teams, expanding the tools to test existing concepts and new ideas. The technology for neuroscience is evolving rapidly, affording new opportunities, new concepts and opening new horizons to the study of individuals and of teams. From their perspective, in the near term, the uniqueness of their approach offers the means to ground future claims made with new theories about how teams function. But the applications the authors review are already important as autonomous technology is introduced into surgical teams, child care, and training leaders.

## 13.2 Introduction. Safety and Human Error

Foundational problems remain in the continuing development of AI for team autonomy, especially with objective measures able to optimize team function, performance and composition. But we want to know whether once this problem has been solved, will we scientists be able to invert the solutions for AI systems in order to mitigate human error.

AI approaches often attempt to address autonomy by modeling aspects of human decision-making or behavior. As in Sect. 13.1, behavioral theory is either based on modeling the individual, such as through cognitive architectures or, more rarely, through group dynamics and interdependence theory. Approaches focusing on the individual assume that individuals are more stable than the social interactions in which they engage. Interdependence theory assumes the opposite, that a state of mutual dependence among participants in an interaction affects the individual and group beliefs and behaviors of participants. The latter is conceptually more complex, but both approaches must satisfy the demand for predictable outcomes as autonomous teams grow in importance and number.

From an intuitive perspective, interdependence may be confusing. As a simple example of interdependence, foraging prey overgraze forests free of predators (Carroll 2016); as another example, we have long known that behavior changes when humans believe they are being observed (Roethlisberger and Dickson 1939).

Despite its theoretical complexity, including the inherent uncertainty and nonlinearity from interdependence, we argue that complex autonomous systems must consider multi-agent interactions to develop predictable, effective and efficient hybrid teams (Lawless 2017). Important examples include cases of supervised autonomy, where a human oversees several interdependent autonomous systems (see Sibley et al. Chap. 7); where an autonomous agent is working with a team of humans, such as in a network cyber defense (see Mylrea and Gourisetti Chap. 12); or where the agent is intended to replace effective, but traditionally worker-intensive team tasks, such as

warehousing and shipping. Autonomous agents that seek to fill these roles, but do not consider the interplay between the participating entities, will likely disappoint.

*Overview of the Problem of Human Error and AI's Role in Its Possible Mitigation* AI has the potential to mitigate human error by reducing car accidents; airplane accidents; and other mistakes made either mindfully or inadvertently by individuals or teams of humans. One worry about this bright future is that jobs may be lost as claimed by Mims (2015), but discounted by Tingley (2017).

Supporting Mims, commercial airline pilots disagree with being replaced by AI (e.g., Smith 2015).

... since the crash of Germanwings Flight 9525, there has been no lack of commentary about the failures and future of piloted aviation ... isn't the best way to prevent another Andreas Lubitz to replace him with a computer, or a remote pilot, somewhere on the ground? ... but a plane no more flies itself than an operating room performs a hip replacement by itself. The proliferation of drone aircraft also makes it easy to imagine a world of remotely controlled passenger planes. ... [but] remember that drones have wholly different missions from those of commercial aircraft, with a lot less at stake if one crashes.

An even greater, existential threat posed by AI is to the existence of humanity, raised separately by the eminent physicist Stephen Hawking, the entrepreneur Elon Musk and the computer billionaire Bill Gates. Garland (2015), the director of the film "Ex Machina", counters them:

... reason might be precisely the area where artificial intelligence excels. I can imagine a world where machine intelligence runs hospitals and health services, allocating resources more quickly and competently than any human counterpart ... the investigation into strong artificial intelligence might also lead to understanding human consciousness, the most interesting aspect of what we are. This in turn could lead to machines that have our capacity for reason and sentience ... [and] a different future ... one day, A.I. will be what survives of us.

With an existential but rigorous view of the possible applications of AI to mitigate human error, when anomalies in human operations occur, or when teams have gone awry, should AI ever intercede in the affairs of humans?

In our applications, we explore both the human's role in the cause of accidents and the possible role of AI in mitigating human error; in reducing problems with teams, like suicide (e.g., the German co-pilot, Libutz, who killed 150 aboard his Germanwings commercial aircraft; from Kulish and Eddymarch 2015); and in reducing the mistakes by military commanders (e.g., the sinking of the Japanese tour-boat by the USS Greenville; from Nathman et al. 2001).

### **13.2.1 Human Error**

Across a wide range of occupations and industries, human error is the primary cause of accidents. For example, pilot error is the leading cause in general aviation (Fowler 2014):

The National Safety Board found that in 2011, 94% of fatal accidents occurred in general aviation

In general aviation, the Federal Aviation Administration (FAA 2014) attributed the accidents that occur primarily to stalls and controlled flight into terrain, that is, to avoidable human error.

Exacerbating the sources of human error, safety is the one area an organization often skimps as it tries to save money. From Gilbert (2015),

“History teaches us that when cuts are made in any industry, the first things to go are safety and training—always, every industry,” says Michael Bromwich, who oversaw the regulatory overhaul after the [Deepwater Horizon] disaster before stepping down as head of the Bureau of Safety and Environmental Enforcement in 2011 in 2011.

In industry, minimizing human error with training is a major commitment (Sanders 2017).

The diminution by an organization in its valuation of safety coupled with human error led to the explosion in 2010 that doomed the Deepwater Horizon in the Gulf of Mexico (Gilbert 2015).

The mistakes that led to the disaster began months before the Deepwater Horizon rig exploded, investigators found, but poor decisions by BP and its contractors sealed the rig’s fate—and their own. On the evening of April 20, 2010, crew members misinterpreted a crucial test and realized too late that a dangerous gas bubble was rising through the well. The gas blasted out of the well and ignited on the rig floor, setting it ablaze.

Human error also emerges as a problem in the management of civilian air traffic control (ATM). ATM’s top five safety risks nearly always involve ‘Human Error’ (ICAO 2014).

Human error was the cause attributed to the recent sinking of a research ship (Normile 2015):

The sinking of Taiwan’s Ocean Researcher V last fall resulted from human error, the head of the country’s Maritime and Port Bureau told local press this week. The 10 October accident claimed the lives of two researchers and rendered the dedicated marine research ship a total loss ... Wen-chung Chi, director-general of the Maritime and Port Bureau, said that a review of the ship’s voyage data recorder and other evidence indicated that the crew should have been alerted that the ship had drifted off course.

What about democracies versus autocracies? From WHO (2015), the US had a crash rate of 10.3 per 100,000 population; a rate of 1.24 crash deaths per 10,000 vehicles; and a rate of 1.10 motor vehicle crash deaths per 100 million vehicle miles traveled. In comparison, China’s crash rate was 18.8 per 100,000 population; and Russia’s was 18.9.

### ***13.2.2 The Role of AI in Reducing Human Error***

The causes of human error could be attributed to endogenous or exogenous factors (we discuss the latter under anomalies and cyber threats). A primary endogenous factor in human causes of accidents is either a lack of situational awareness, or a

convergence into an incomplete state of awareness associated with emotions expressed during decision-making (e.g., the Iranian Airbus Flight 655 erroneously downed by the USS Vincennes in 1988; e.g., Lawless et al. 2013). Many other factors have been subsumed under human error including system complexity; poor problem diagnoses; poor planning, communication and execution; and poor organizational functioning.

What role can AI play in reducing human error? First, Johnson (1973) proposed that the most effective way to reduce human error is to make safety integral to management and operations. Johnson's "MORT" accident tree-analyses attempts to identify the operational control issues that may have caused an accident, and the organizational barriers that resist uncovering the deficiencies that contributed to an accident. MORT has been prized by the Department of Energy as the ultimate tool for identifying possible hazards to the safe operation of nuclear reactors.

Second, checks and balances on cognitive convergence processes permit the alternative interpretations of situational awareness that may prevent human error. Madison (1906) wrote that it is to free speech and a free press, despite all their abuses, that "the world is indebted for all the triumphs which have been gained by reason and humanity over error and oppression." But in closed organizations, like the military in the field, in the cockpit or on a ship's command bridge, the lack of free speech poses a threat that is associated with an increase in errors. Based on Smallman's (2012) plan to reduce accidents in the U.S. Navy's submarine fleet with technology that illustrates in real time the range of opinions existing among a ship's staff for a proposed action, AI can offer alternative perspectives that oppose the very convergence processes that permit errors to thrive (e.g., "groupthink"; Janis 1982).

### 13.2.3 Roles with AI

*The Role of AI in the Discovery and Rectification of Anomalies* An anomaly is a deviation from normal operations (Marble et al. 2015). The factors we wish to consider are those associated with cyberattacks; e.g., the subtle takeover of a drone by a foreign agent. But, we expect, AI will be able to be used proactively for "novel and intuitive techniques that isolate and predict anomalous situations or state trajectories within complex autonomous systems in terms of mission context to allow efficient management of aberrant behavior" (Taylor et al. 2015).

*The Role of AI in Determining Team Metrics and Shaping Team Thermodynamics* Teams enhance the performance of the individual (Cooke and Hilton 2015), the added time to coordinate and communicate among a team's members being a cost and also a potential source of error. But, with the metrics we provided in Sect. 13.1, we propose that under the right conditions, well-performing teams can decrease the likelihood of human error, while poorly performing teams almost always increase the risk of human error.

With Shannon's static, stable agents, the information theory community appears to be more focused on the mutual information shared at the input and output of two variables, forming a channel. Instead, for individual agents forming a team, we devised a measure of team efficiency based on the information transmitted by a team. Along with Cummings, we concluded that (Moskowitz et al. 2015, p. 2), in general, a "high degree of team player interdependence is desirable (mapping to good performance)."

Building on our concept of efficiency, but with bistable agents to replicate humans as observers and actors, we noted that Shannon's team entropy is at a minimum for a team slaved together. Contradicting Shannon except when observation and action are perfectly aligned, or when multiple interpretations converge in perfect agreement, in our model, like the independent brain systems that observe and act interdependently as a team, we set the baseline entropy for a perfect team's structure at least entropy production (LEP); when the structure consumes LEP, that allows a team to maximize its entropy production (MEP) in the pursuit of its mission to achieve maximum performance.

### ***13.2.4 Forecasts with AI and Interdependence***

We propose that limit cycles are created by the interdependence between relatively balanced or equal sides of a debate as they entangle observers (e.g., juries in courtrooms; independents in politics; the un-decideds in science; in Lawless 2017); we identify the source of these limit cycles as those decisions driven by Nash equilibria (e.g., Republicans versus Democrats), arising from a people free to move and capital free to invest as information is developed by competition from the two sides of a debate seeking vulnerabilities in its opponent. When one of those two sides are suppressed, as happens under autocracies, errors increase as checks on decisions are overridden (e.g., Cuba; China; Venezuela), leading to social collapse (e.g., the decaying towns in Russia; from Antonova 2017). In conclusion, while limit cycles are modeled by the simple bistable equations used to model predator-prey relationships, the loss of meaning in social systems is profound, leading to endless debates, but better decisions that vastly improve social welfare (e.g., poorer decisions were made by the Department of Energy when the public had no idea that the consequences of DOE's waste management practices at its Savannah River Site plant in South Carolina lead to disastrous radioactive releases into the air; surface waters and ground waters; and across and under the land; however, nowadays, every decision made at SRS is fought in public between the affected States, National Academy of Scientists, DOE managers and scientists, and local citizens, leading to decisions that have vastly improved the environment at and around SRS; in Lawless et al. 2014).

Finally, we began this chapter by asking the question given interdependence, what can be forecasted? Our answer is that while we can assign probabilities for interdependent outcomes, as is already commonly done, structuring the interaction to increase competition is more predictive and with much better outcomes. The

more competition that exists among teams operating freely, the better will be social welfare and the less chance of human error endangering public safety.

### 13.2.5 Evaluations

In Chap. 13, Sect. 13.3.2, we discussed the use of robots in hybrid teams to reduce human error. From our perspective, we envision drawing a boundary about a human operator (e.g., airline pilot, train engineer, car driver), and a robot assistant (e.g., the airplane, train or car, respectively). We foresee the robot assistant (or human assistant) taking control whenever joint entropy production for a high-performing team exceeds the contributions for the team members (shifting from Eqs. (13.7) to (13.8)).

$$\dot{S}_{AB} \rightarrow H(\mathbf{x}, \mathbf{y}) \quad (13.9)$$

We complete the evaluations of the remaining Chapters next.

Chapter 3 links directly to one of the book's themes about using AI to address the recovery from human error. In this chapter, the authors depicted a multi-agent swarm of robots working in a disaster control situation, where several AI (machine-learning) algorithms have been implemented, like clustering and travelling salesman, to find the most efficient geographic route to navigate and coordinate the robot swarm. Though there have been different implementations of swarms of robots, here applied with robot swarms to model the recovery of the human victims of a shipwreck, their research focused on the integration of swarms of robots in a system of systems environment, using real-world factors as decisive parameters in the algorithms used in the process for the victims such as their location, as well as the location, number, speed and capacity of rescue boats. While the results and demonstrations are convincing, the system defined by the authors can be implemented in fields other than for a cruise ship disaster situation. Although the application the authors have provided in this chapter was to recover from a specific example of human error (viz., a shipwreck), more broadly, they have provided a general computational method for the use of robot swarms controlled with AI.

In Chap. 4, the authors note that human information interaction (HII) is a novel label for the field of human systems integration that recently only included human computer interaction. Looking widely across the field of information integration, these domains that they have studied would be directly related and applicable to the challenges identified by the authors in this chapter at the same time that there is a significant amount of existing research ongoing in these areas. From their perspective, by studying the levels of automation, the authors conclude that human information from the interaction can be automated to assist and guide interactions with information. But, wherever human interactions occur, mismatches make it more likely for an increase in errors to also occur. The authors have named their approach as AI-augmented HII. The authors do not overreach in their claims that AI will be a

proxy for human interactions with information and they do a reasonable job of identifying the bounds of HII. While the chapter provides a thorough review of the relevant literature, from their perspective, future work should prioritize the specific aspects of HII research that will most impact the errors that AI systems may cause in order to identify the ways that quantitative and computational methods can mitigate errors for both human and autonomous systems with explainable AI.

The authors of Chap. 6 indicated that much of the research that they and others have done, especially on trust repair, has taken place in virtual, simulated experiments. Other work (both by the authors of this chapter as well as by others in the literature) has found differences between human behavior in some virtual and physical experiments, but similar results in others. As such, the next step is to repeat similar experiments in physical environments. But, based on what they have already presented, and from the trust modification work focused on repairing broken trust, in physical interactions with robots it is becoming increasingly clear to the satisfaction of the authors that it is difficult to break a human's trust placed in a robot, even when a robot is obviously wrong in its directions, and even after a robot apologizes. Future work should focus on trust modification techniques that appropriately calibrate (i.e., that lower or raise trust in a robot as necessary) the trust by human users that is given to a robot. In the event that robots will be deployed to assist humans in their recovery from human errors or accidents, it is a cautionary tale to know that humans must learn to be more judicious in the trust they give to robots by not over-trusting robots, which can make a bad situation worse.

The authors of Chap. 9 note that, usually, the automated reasoning (of robots) is often compared to the reasoning of mentally healthy humans. Instead, in their chapter, the authors attempt with their research to show the benefits from a comparison between robots and the impaired thinking of autists. Their approach can potentially be more fruitful in terms of understanding how to implement reasoning and improve it in both robots and autists. However, reasoning of individuals with mental disorders is less systematic by definition, so it could be harder to have a valid comparison framework with autists. Still, this chapter combines experimental and theoretical work on reasoning in both humans and robots, demonstrating that it is possible to consider these forms of reasoning within a unified framework. As an experimental study of how autism impairs teams, this work has a rather limited dataset with a possibly restricted coverage. But this research fills an important gap in research. One day, autists assisted by robots dedicated to their care and education may offer a revolution in care and social welfare to not only autistic children, but to the elderly and the infirm as well.

Chapter 12 provides a timely and unique exploration of autonomy at the nexus of cyber-security and the management of energy systems and building systems. The authors explore how AI is creating new opportunities to make critical sectors more efficient through advances in AI and automation as well as the challenges accompanying systems that are more cyber vulnerable and also those that are more protected. They provide specific use cases that highlight how AI is increasing the size and speed of data exchanges among systems as it is connecting cyber and physical infrastructure systems to the Internet, increasing the need for protection and caution.

As systems become more and more autonomous, the authors also provide unique concepts on how smart buildings and infrastructures can leverage AI to become more resilient and secure in a complex environment of evolving cyber and system threats against risks ranging from low to potentially severe. In general, cyber-threats to autonomy suggest the need for better responses to these new threats along with the threats that could potentially compromise those infrastructures that already exist.

## References

- Aberman, J. (2017, 2/27), "Artificial intelligence will change America. Here's how", *Washington Post*, from [https://www.washingtonpost.com/news/capital-business/wp/2017/02/27/artificial-intelligence-will-change-america-heres-how/?utm\\_term=.9835268e8dfd](https://www.washingtonpost.com/news/capital-business/wp/2017/02/27/artificial-intelligence-will-change-america-heres-how/?utm_term=.9835268e8dfd)
- Adelson, E. H. (2000). Lightness perceptions and lightness illusions. The new cognitive sciences, 2nd Ed. M. Gazzaniga. MIT Press.
- Ahdieh, R.G. (2009), Beyond individualism and economics, retrieved 12/5/09 from [ssrn.com/abstract=1518836](http://ssrn.com/abstract=1518836).
- Ambrose, S.H. (2001), Paleolithic technology and human evolution, *Science*, 291, 1748-3.
- Antonova, M. (2017, 3/4), "This Is the Russia You're So Afraid Of?", *New York Times*, from <https://www.nytimes.com/2017/03/04/opinion/sunday/this-is-the-russia-youre-so-afraid-of.html>
- Arrow, K.J. (1951, 1963), *Social Choice and Individual Values* (2nd edn.). New Haven: Yale University Press.
- Arrow, K.J. (1974) *The Limits of Organization*. NY: Norton [Comments on *The Limits of Organization* by Kenneth J. Arrow (1998, Aug 29), JOI ITO (The Harvard Book Store), from <https://joi.ito.com/weblog/1998/08/29/comments-on-the.html>]
- Austin, S. (2017, 3/6), "How Artificial Intelligence Will Change Everything. Baidu's Andrew Ng and Singularity's Neil Jacobstein say this time, the hype about artificial intelligence is real", *Wall Street Journal*, from <https://www.wsj.com/articles/how-artificial-intelligence-will-change-everything-1488856320>
- Bacharach, M. (2006), *Beyond Individual Choice: Teams and Frames in Game Theory*. Natalie Gold and Robert Sugden (eds.), Princeton: Princeton University Press.
- Balch, T. (2000), Hierarchic Social Entropy: An Information Theoretic Measure of Robot Team Diversity., *Autonomous Robots*.
- Bartel, A., Cardiff-Hicks, B. & Shaw, K. (2013), Compensation Matters: Incentives for Multitasking in a Law Firm, NBER Working Paper No. 19412
- Bekoff, M. (2011, 11/4), What Makes Us Uniquely Human? *Psychology Today*, from <https://www.psychologytoday.com/blog/animal-emotions/201111/what-makes-us-uniquely-human>
- Bunge, J. (2016, 9/14), "Bayer-Monsanto Deal Faces Heavy Regulatory Scrutiny. Combining two of the world's largest farm suppliers will test politicians wary of consolidation in the \$100 billion global market", *Wall Street Journal*, from <http://www.wsj.com/articles/bayer-monsanto-deal-faces-heavy-regulatory-scrutiny-1473855922>
- Battaglia, P., Ullman, T., Tenenbaum, J., Forbus, K., Sanborn, A., Gerstenberg, T. & Lagnado, D. (2013), Computational models of intuitive physics, *eScholarship.org*, pp. 32–33.
- Carroll, S.B. (2016). *The Serengeti rules. The Quest to Discover How Life Works and Why It Matters*. Princeton: Princeton University Press.
- Casey, N. (2014, 11/7), "Gaza Explosions Hit Senior Fatah Members' Homes. Blasts Rekindle Tensions Between Two Main Palestinian Political Factions", *Wall Street Journal*, [http://online.wsj.com/articles/gaza-explosions-hit-senior-fatah-members-homes-1415356232?mod=WSJ\\_hps\\_sections\\_world](http://online.wsj.com/articles/gaza-explosions-hit-senior-fatah-members-homes-1415356232?mod=WSJ_hps_sections_world)
- Centola, D. & Macy, M. (2007), Complex Contagions and the Weakness of Long Ties, *AJS*, 113(3): 702–34.



- Chagnon, N. (1988), "Life Histories, Blood Revenge, and Warfare in a Tribal Population", *Science* 239: 985–92.
- Chagnon, N. A. (2012), *The Yanomamo*. New York, Wordsworth.
- Chopping, D. (2016, 9/22), "Maersk to Split Into Two Separate Units. Move comes months after departure of CEO Nils Andersen", *Wall Street Journal*, from <http://www.wsj.com/articles/maersk-to-split-into-two-units-1474529401>
- Coase, R. (1937). "The nature of the firm." *Economica* 4: 386.
- Coase, R. (1960). "The problem of social costs." *Journal of Law and Economics* 3: 1-44.
- Cohen, L. (1995). *Time-frequency analysis: theory and applications*, Prentice Hall.
- Conant, R. C. (1976). "Laws of information which govern systems." *IEEE Transaction on Systems, Man and Cybernetics* 6: 240-255.
- Cooke, N.J. & Hilton, M.L. (eds.) (2015, 4/24), *Enhancing the Effectiveness of Team Science*, Committee on the Science of Team Science, Board on Behavioral, Cognitive, and Sensory Sciences, Division Behavioral and Social Sciences and Education, National Research Council, DC.
- Cummings, J. (2015, 6/2). *Team Science Successes and Challenges*. National Science Foundation Sponsored Workshop on Fundamentals of Team Science and the Science of Team Science, Bethesda MD.
- EB, Editorial Board (EB) (2014, 6/7), "Detroit's Fight Against Blight", *New York Times*, [http://www.nytimes.com/2014/06/08/opinion/sunday/detroits-fight-against-blight.html?\\_r=0](http://www.nytimes.com/2014/06/08/opinion/sunday/detroits-fight-against-blight.html?_r=0)
- Ewing, J. (2017, 3/3), "G.M. Near Deal to Sell Opel to Peugeot Maker, PSA", *New York Times*, from <https://www.nytimes.com/2017/03/04/business/dealbook/psa-group-opel-general-motors-deal.html>
- FAA (2014, 7/30), *Fact Sheet – General Aviation Safety*, U.S. Federal Aviation Admin., from [http://www.faa.gov/news/fact\\_sheets/news\\_story.cfm?newsId=16774](http://www.faa.gov/news/fact_sheets/news_story.cfm?newsId=16774)
- Fitzgerald, D. & Knutson, R. (2017, 3/6), "Radio Shack to Seek Bankruptcy Protection, Again. Chain already closing 200 stores as it prepares to file as soon as Tuesday", *Wall Street Journal*, from <https://www.wsj.com/articles/radioshack-to-seek-bankruptcy-protection-again-1488843485>
- Fowler, D. (2014, 7/16), "The Dangers of Private Planes", *New York Times*, from <http://www.nytimes.com/2014/07/17/opinion/The-Dangers-of-Private-Planes.html>
- Garland, A. (2015, 4/26), "Robot overlords? Maybe not; Alex Garland of 'Ex Machina' Talks About Artificial Intelligence"; *New York Times*, from [http://www.nytimes.com/2015/04/26/movies/alex-garland-of-ex-machina-talks-about-artificial-intelligence.html?\\_r=0](http://www.nytimes.com/2015/04/26/movies/alex-garland-of-ex-machina-talks-about-artificial-intelligence.html?_r=0)
- Gilbert, D. (2015, 4/19), "Oil Rigs' Biggest Risk: Human Error. Five years after Deepwater Horizon, new rules and training go only so far", *Wall Street Journal*, from <http://www.wsj.com/articles/rig-safety-five-years-after-gulf-spill-1429475868>
- Graziano, M.S.A. (2013), *Consciousness and the Social Brain*, Oxford University Press: Oxford, UK.
- Hope, B. (2014, 9/1), "Can the New York Stock Exchange Be Saved? After Historic Takeover by Intercontinental, New Chiefs Plot Turnaround for Storied NYSE"; *Wall Street Journal*, <http://online.wsj.com/articles/can-the-new-york-stock-exchange-be-saved-1409625002?KEYWORDS=NYSE+euronext>
- Howell, J.C. (2006, August), "The Impact of Gangs on Communities", *NYGC Bulletin No. 2* (National Youth Gang Center), from <http://www.nationalgangcenter.gov/content/documents/national-of-gangs-on-communities.pdf>
- ICAO (2014, 8/8), "Integration of Human Factors in research, operations and acquisition", presented at the International Civil Aviation Organization's Second Meeting of the APANPIRG ATM Sub-Group (ATM/SG/2)", ATM/SG/2– IP04 04-08/08/2014, Hong Kong, China, 04-08.
- Kulish, N. & Eddymarch, M.A. (2015, 3/30), "Germanwings Co-Pilot Was Treated for 'Suicidal Tendencies,' Authorities Say", *New York Times*, from [www.nytimes.com/2015/03/31/world/europe/germanwings-copilot-andreas-lubitz.html?\\_r=0](http://www.nytimes.com/2015/03/31/world/europe/germanwings-copilot-andreas-lubitz.html?_r=0)
- Janis, Irving L. (1982). *Groupthink: Psychological Studies of Policy Decisions and Fiascoes* (2nd ed.). New York: Houghton Mifflin.
- Jenkins, Jr. H.W. (2017, 2/14), "Dieselgate Is a Political Disaster. Typical of Western leadership was a policy of huge cost and zero benefit", *Wall Street Journal*, from <https://www.wsj.com/articles/dieselgate-is-a-political-disaster-1487116586>.

- Johnson, W.G. (1973), MORT: The management oversight and risk tree, SAN 821-2.
- Kahneman, D. (2002, 12/8), "Maps of bounded rationality: A perspective on intuitive judgment and choice", Prize Lecture, from [http://www.nobelprize.org/nobel\\_prizes/economic-sciences/laureates/2002/kahnemann-lecture.pdf](http://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/2002/kahnemann-lecture.pdf)
- Kapner, S. & Dulaney, C. (2014, 11/7), "Searsyork-stock-exchange-be-saved-1409625002?KEYWORDS=NYSE+ould Lose \$630 Million Last Quarter, Mulls Move to Spin Off 300 Stores", *Wall Street Journal*, fromt, [http://online.wsj.com/articles/sears-considering-forming-reit-to-boost-liquidity-1415362718?mod=WSJ\\_hp\\_LEFTWhatsNewsCollection](http://online.wsj.com/articles/sears-considering-forming-reit-to-boost-liquidity-1415362718?mod=WSJ_hp_LEFTWhatsNewsCollection)
- Kennedy, S. (2016, 5/18), Superforecasters See 23% Brexit Chance as Economy Wins Out, Bloomberg, from <https://www.bloomberg.com/news/articles/2016-05-18/superforecasters-see-24-chance-of-brexit-as-economy-wins-out>
- Kenny, D. A., Kashy, D.A., & Bolger, N. (1998). Data analyses in social psychology. Handbook of Social Psychology. D. T. Gilbert, Fiske, S.T. & Lindzey, G. Boston, MA, McGraw-Hill. 4th Ed., Vol. 1: pp. 233-65.
- Kirk, R. (2003). More terrible than death. Massacres, drugs & America's war in Columbia, Pub. Affairs.
- Kowsmann, P., Enrich, D. & Patrick, M. (2014, 8/28), "KPMG Faces Criticism for Espírito Santo Audit Work. Bank's Collapse Raises Questions Whether KPMG Should Have Detected Problems Earlier", *Wall Street Journal*, <https://www.wsj.com/articles/kpmg-faces-criticism-for-espírito-santo-audit-work-1409227480>.
- Lanchester, J. (2017, 2/7), "The Major Blind Spots in Macroeconomics", *New York Times Magazine*, from <https://www.nytimes.com/2017/02/07/magazine/the-major-blind-spots-in-macroeconomics.html?>
- Laudisa, F. and Rovelli, C. (2013, Summer), "Relational Quantum Mechanics", *The Stanford Encyclopedia of Philosophy*, Edward N. Z. (ed.), from <https://plato.stanford.edu/archives/sum2013/entries/qm-relational>.
- Lawless, W. F., Linas, James, Mittu, Ranjeev, Sofge, Don, Sibley, Ciara, Coyne, Joseph, & Russell, Stephen (2013). "Robust Intelligence (RI) under uncertainty: Mathematical and conceptual foundations of autonomous hybrid (human-machine-robot) teams, organizations and systems." *Structure & Dynamics* 6(2), from [www.escholarship.org/uc/item/83b1t1zk](http://www.escholarship.org/uc/item/83b1t1zk).
- Lawless, W.F., Akiyoshi, Mito, Angjellari-Dajcic, Fiorentina & Whitton, John (2014), Public consent for the geologic disposal of highly radioactive wastes and spent nuclear fuel, *International Journal of Environmental Studies*, 71(1): 41-62.
- Lawless, W.F. 2017, (published online December 2016; forthcoming), The entangled nature of interdependence. Bistability, irreproducibility and uncertainty, *Journal of Mathematical Psychology*, doi: [10.1016/j.jmp.2016.11.001](https://doi.org/10.1016/j.jmp.2016.11.001)
- Lien, T. (2016, 3/21), "Artificial intelligence has mastered board games; what's the next test? Board games are said to have been perfect tests for artificial intelligence, but real life can be messier. That has researchers pushing deeper into AI development", *Los Angeles Times*, from <http://www.seattletimes.com/business/technology/artificial-intelligence-has-mastered-board-games-whats-the-next-test/>
- Madison, J. (1906), "Report on the Virginia Resolutions, 1799-1800", in *The writings of James Madison*, G. Hunt (Ed.), New York: Putnam & Sons, 6: 386.
- Mamiit, A. (2015, 1/29), "Bill Gates, Like Stephen Hawking and Elon Musk, Worries About Artificial Intelligence Being a Threat", *Tech Times*, <http://www.techtimes.com/articles/29436/20150129/bill-gates-like-stephen-hawking-and-elon-musk-worries-about-artificial-intelligence-being-a-threat.htm>
- Marble, J., Lawless, W.F., Mittu, R. Coyne, J., Abramson, M. & Sibley, C. (2015), "The Human Factor in Cybersecurity: Robust & Intelligent Defense", in *Cyber Warfare*, Sushil Jajodia, Paulo Shakarian, V.S. Subrahmanian, Vipin Swarup, Cliff Wang (Eds.), Berlin: Springer.
- Martyushev, L.M. (2013), Entropy and entropy production: Old misconceptions and new breakthroughs, *Entropy*, 15: 1152-70.
- May, R. M. (1973/2001), *Stability and complexity in model ecosystems*. Princeton U. Press.
- McCloskey, M. (1983, 4), "Intuitive Physics. Although Newton's laws are well known, tests show many people believe moving objects behave otherwise. The subjects of the tests tend to follow

- a theory held in the three centuries before Newton", *Scientific American*, from <https://www.scientificamerican.com/article/intuitive-physics/>
- Mims, C. (2015, 4/19), "Data Is the New Middle Manager. Startups are keeping head counts low, and even eliminating management positions, by replacing them with a surprising substitute for leaders and decision-makers: data", *Wall Street Journal*, from <http://www.wsj.com/articles/data-is-the-new-middle-manager-1429478017>
- Moskowitz, Ira S., Lawless, W.F., Hyden, Paul, Mittu, Ranjeev & Russell, Stephen (2015), A Network Science Approach to Entropy and Training, Proceedings, AAAI Spring 2015, Stanford University.
- Nathman, J.B., VAdm USN et al., (2001, April 13), "Court of inquiry into the circumstances surrounding the collision between USS Greenville (SSN 772) and Japanese M/V Ehime Maru; JAG-INST 5830.1; <news.findlaw.com/hdocs/docs/greenville/ussgrnv1041301rprrt.pdf>
- NYT, *New York Times* (NYT) (2010, 10/28), Many groups blue in '08 are now red; chart from <http://www.nytimes.com/imagepages/2010/10/28/us/politics/28poll-g.html?ref=politics>
- Nicolis, G., & Prigogine, I. (1989), Exploring complexity, New York: Freeman.
- Normile, D. (2015, 3/27), "Human error led to sinking of Taiwanese research vessel", *Science*, from <http://news.sciencemag.org/asiapacific/2015/03/human-error-led-sinking-taiwanese-research-vessel>
- Nosek, B.A. (2015), Estimating the reproducibility of psychological science, *Science*, 349(6251): 943.
- Otto, B. & Sentana, I.M. (2015, 10/23), "Indonesia Prepares Navy Ships to Evacuate Haze Victims. Country suffering from some of its worst agricultural fires in years", *Wall Street Journal*, from <http://www.wsj.com/articles/indonesia-prepares-navy-ships-to-evacuate-haze-victims-1445602767?alg=y>
- Pfeffer, J. & Fong, C.T. (2005), Building Organization Theory from First Principles: The Self-Enhancement Motive and Understanding Power and Influence, *Organization Science*, 16(4): 372-388.
- Rees, G., Frackowiak, R., & Frith, C. (1997). Two modulatory effects of attention that mediate object categorization in human cortex, *Science*, 275, 835-8.
- R&O (2017, 2/21), "A Military Strategist for Trump's NSC. H.R. McMaster wrote a book about the duty to challenge a President", Review & Outlook, *Wall Street Journal*, from <https://www.wsj.com/articles/a-military-strategist-for-trumps-nsc-1487634622>
- Roethlisberger, F. & Dickson, W. (1939). *Management and the worker*. Cambridge: Cambridge University Press.
- Sanders, D. (2017, 2/19), Personal communication.
- Simon, H. (1978, 12/8), Rational decision making in business organizations, Nobel Lecture, from [http://www.nobelprize.org/nobel\\_prizes/economic-sciences/laureates/1978/simon-lecture.pdf](http://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/1978/simon-lecture.pdf)
- Simon, H.A. (1989, 9/23), Bounded rationality and organizational learning, Technical Report AIP 107, CMU, Pittsburgh, PA.
- Smallman, H. S. (2012). TAG (Team Assessment Grid): A Coordinating Representation for submarine contact management. SBIR Phase II Contract #: N00014-12-C-0389, ONR Command Decision Making 6.1-6.2 Program Review.
- Smith, P. (2015, 4/10), Why Pilots Still Matter, *New York Times*, from [http://www.nytimes.com/2015/04/10/opinion/why-pilots-still-matter.html?ref=international&\\_r=1](http://www.nytimes.com/2015/04/10/opinion/why-pilots-still-matter.html?ref=international&_r=1)
- Taylor, G., Mittu, R., Sibley, C., Coyne, J. & Lawless, W.F. (2015, forthcoming), "Towards Modeling the Behavior of Autonomous Systems and Humans for Trusted Operations", in R. Mittu, D. Sofge, A. Wagner & W.F. Lawless (Eds), *Robust intelligence and trust in autonomous systems*, Berlin: Springer.
- Tingley, K. (2017, 2/23), "Learning to Love Our Robot Co-Workers. The most important frontier for robots is not the work they take from humans but the work they do with humans — which requires learning on both sides", *New York Times*, from [https://www.nytimes.com/2017/02/23/magazine/learning-to-love-our-robot-co-workers.html?\\_r=0](https://www.nytimes.com/2017/02/23/magazine/learning-to-love-our-robot-co-workers.html?_r=0)
- Smith, J. (2014, July), Personal communication.
- USTPO (2013, December), Patents By Country, State, and Year - All Patent Types, US Trade Patent Office, from [http://www.uspto.gov/web/offices/ac/ido/oeip/taf/cst\\_all.htm](http://www.uspto.gov/web/offices/ac/ido/oeip/taf/cst_all.htm)
- WHO (2015), Global Status Report on Road Safety, World Health Organization, from [http://www.who.int/violence\\_injury\\_prevention/road\\_safety\\_status/2015/en/](http://www.who.int/violence_injury_prevention/road_safety_status/2015/en/)

- Wickens, C. D. (1992). *Engineering psychology and human performance* (2nd edition), Merrill.
- Wong, C.H. (2015, 12/22), "Landfill Operator in Shenzhen Was Warned About Safety Risks. Two inspection firms flagged potential problems at China landfill before landslide left dozens missing", *Wall Street Journal*, from <http://www.wsj.com/articles/landfill-operator-in-shenzhen-was-warned-about-safety-risks-1450803052>
- Zell, E. & Krizan, Z. (2014), Do People Have Insight Into Their Abilities? A Metasynthesis? *Perspectives on Psychological Science* 9(2): 111-125.
- Zipf, G.K. (1949), *Human behavior and the principle of least effort*, New York: Addison-Wesley.