


STEAM-H: Science, Technology, Engineering, Agriculture,  
Mathematics & Health

Tomas Veloz  
Andrei Khrennikov  
Bourama Toni  
Ramón D. Castillo *Editors*

# Trends and Challenges in Cognitive Modeling

An Interdisciplinary Approach Towards  
Thinking, Memory, and Decision-Making  
Simulations

 Springer

# **STEAM-H: Science, Technology, Engineering, Agriculture, Mathematics & Health**

## **Series Editor**

Bourama Toni, Department of Mathematics, Howard University, Washington, DC,  
USA

This interdisciplinary series highlights the wealth of recent advances in the pure and applied sciences made by researchers collaborating between fields where mathematics is a core focus. As we continue to make fundamental advances in various scientific disciplines, the most powerful applications will increasingly be revealed by an interdisciplinary approach. This series serves as a catalyst for these researchers to develop novel applications of, and approaches to, the mathematical sciences. As such, we expect this series to become a national and international reference in STEAM-H education and research.

Interdisciplinary by design, the series focuses largely on scientists and mathematicians developing novel methodologies and research techniques that have benefits beyond a single community. This approach seeks to connect researchers from across the globe, united in the common language of the mathematical sciences. Thus, volumes in this series are suitable for both students and researchers in a variety of interdisciplinary fields, such as: mathematics as it applies to engineering; physical chemistry and material sciences; environmental, health, behavioral and life sciences; nanotechnology and robotics; computational and data sciences; signal/image processing and machine learning; finance, economics, operations research, and game theory. The series originated from the weekly yearlong STEAM-H Lecture series at Virginia State University featuring world-class experts in a dynamic forum. Contributions reflected the most recent advances in scientific knowledge and were delivered in a standardized, self-contained and pedagogically-oriented manner to a multidisciplinary audience of faculty and students with the objective of fostering student interest and participation in the STEAM-H disciplines as well as fostering interdisciplinary collaborative research. The series strongly advocates multidisciplinary collaboration with the goal to generate new interdisciplinary holistic approaches, instruments and models, including new knowledge, and to transcend scientific boundaries.

### **Peer reviewing**

All monographs and works selected for contributed volumes within the STEAM-H series undergo peer review. The STEAM-H series follows a single-blind review process. A minimum of two reports are asked for each submitted manuscript. The Volume Editors act in cooperation with the Series Editor for a final decision. The Series Editor agrees with and follows the guidelines published by the Committee on Publication Ethics.

*Titles from this series are indexed by Scopus, Mathematical Reviews, and zbMATH.*

Tomas Veloz • Andrei Khrennikov •  
Bourama Toni • Ramón D. Castillo  
Editors

# Trends and Challenges in Cognitive Modeling

An Interdisciplinary Approach Towards  
Thinking, Memory, and Decision-Making  
Simulations

 Springer

### *Editors*

Tomas Veloz   
Departamento de Matemática  
Universidad Tecnológica Metropolitana  
Santiago, Chile

Interdisciplinary Foundation  
for the Development of Science  
Technology and Arts  
Santiago, Chile

Centre Leo Apostel  
Vrije Universiteit Brussel  
Brussels, Belgium

Bourama Toni  
Department of Mathematics  
Howard University  
Washington, DC, USA

Andrei Khrennikov  
International Center for Mathematical  
Modeling in Physics and Cognitive Sciences  
School of Mathematics, Linnaeus University  
Växjö, Sweden

Ramón D. Castillo  
Research Center on Cognitive Sciences  
Universidad de Talca  
Talca, Chile

ISSN 2520-193X ISSN 2520-1948 (electronic)  
STEAM-H: Science, Technology, Engineering, Agriculture, Mathematics & Health  
ISBN 978-3-031-41861-7 ISBN 978-3-031-41862-4 (eBook)  
<https://doi.org/10.1007/978-3-031-41862-4>

Mathematics Subject Classification: 68Q12, 81P16, 81P42, 81P45, 81P68, 81P99

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

This work is subject to copyright. All rights are solely and exclusively licensed 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, expressed 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.

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

Paper in this product is recyclable.

# Preface

The book presents perspectives and development of novel and fundamental ideas for modern understanding of cognitive science modeling and simulation. It is built upon a series of international workshops featuring world experts and collaboration between the Centre for Research in Cognitive Science at the University of Talca, Chile, the Centre Leo Apostel at the Vrije Universiteit Brussels, Belgium, the International Center for Mathematical Modeling at the Linnaeus University, Sweden, and the Foundation for Interdisciplinary Development of Science, Technology and Arts, Chile.

Researchers from various branches, ranging from quantum theory to complex systems, have proposed explanatory framework for thinking and behavior; the topics covered in the book include relation between quantum entanglement and decision-making, generalized probability models, applied advances in simulations of human memory with quantum computers, and representation of conceptual reasoning in geometrical spaces.

More importantly, the book includes the trends and challenges in cognitive modeling and simulations. It is indeed suitable for the multidisciplinary STEAM-H series, which aims at bringing together leading researchers to present their own work in the perspective to advance their specific fields and in a way to generate a genuine interdisciplinary interaction transcending disciplinary boundaries. Contributions are invited only and reflect the most recent advances delivered in a high-standard, self-contained way.

This volume, as others in the series, strongly advocates multidisciplinary with the goal to generate new interdisciplinary approaches, instruments, and models including new knowledge, transcending scientific and mathematical boundaries to adopt a more holistic approach. As a whole, the book certainly enhances the overall objective of the series, that is, to foster the readership interest and enthusiasm in the STEAM-H disciplines (science, technology, engineering, agriculture, mathematics, and health), stimulate graduate and undergraduate research, and generate collaboration between researchers on a genuine interdisciplinary basis. The shared emphasis of these carefully selected and peer-refereed contributed papers is on

important methods, research directions, and applications of analysis, modeling, and simulations including within and beyond mathematical sciences.

The STEAM-H series is by now well established as a reference of choice for interdisciplinary scientists and mathematicians and a source of inspiration for a broad spectrum of researchers, with a high impact through all its volumes published by the world renown Springer Nature.

Talca, Chile  
Växjö, Sweden  
Washington, DC, USA  
Brussels, Belgium  
May 2023

Ramón D. Castillo  
Andrei Khrennikov  
Bourama Toni  
Tomas Veloz

# Acknowledgments

We would like to express our sincere appreciation to all the contributors and to all the anonymous referees for their professionalism. They all made this volume a reality for the greater benefit of the community of science, technology, engineering, agriculture, mathematics, and health.



# Contents

<b>Introduction: Modern Approaches to the Study of Human Cognition</b> .....	1
Bourama Toni	
<b>Use of Agent-Based Modeling (ABM) in Psychological Research</b> .....	7
Enrique Canessa, Sergio E. Chaigneau, and Nicolás Marchant	
<b><i>Nyāyasūtra</i> Proof Pattern: An Interpretation of Similarity as the Fact of Sharing Two Properties</b> .....	21
Miguel López-Astorga	
<b>Using Pheromone Trail Algorithm to Model Analog Memory</b> .....	33
Trung T. Pham, Ramón D. Castillo, Xiaojing Yuan, and Heidi Kloos	
<b>Review on Social Laser Theory and Its Applications</b> .....	53
Andrei Khrennikov	
<b>Challenges from Probabilistic Learning for Models of Brain and Behavior</b> .....	73
Nicolás Marchant, Enrique Canessa, and Sergio E. Chaigneau	
<b>The Emergence of Cognition and Computation: A Physicalistic Perspective</b> .....	85
Karl Svozil	
<b>Analysing the Conjunction Fallacy as a Fact</b> .....	101
Tomas Veloz and Olha Sobetska	
<b>Yes Ghosts, No Unicorns: Quantum Modeling and Causality in Physics and Beyond</b> .....	113
Kathryn Schaffer and Gabriela Barreto Lemos	
<b>Compositional Vector Semantics in Spiking Neural Networks</b> .....	131
Martha Lewis	

**Optimality, Prototypes, and Bilingualism** ..... 147  
Igor Douven and Galina V. Paramei

**The Dimensionality of Color Perception**..... 165  
Javier Fdez, Oneris Rico, and Olaf Witkowski

**Index**..... 181

# Contributors

**Enrique Canessa** Faculty of Engineering and Science, Universidad Adolfo Ibanez, Vina del Mar, Chile

**Ramón D. Castillo** Research Center on Cognitive Sciences, University of Talca, Talca, Chile

**Sergio E. Chaigneau** Center for Social and Cognitive Neuroscience, School of Psychology, Universidad Adolfo Ibanez, Santiago, Chile

**Igor Douven** IHPST/CNRS Panthéon-Sorbonne University, Paris, France

**Javier Fdez** Cross Labs, Cross Compass Ltd, Kyoto, Japan

**Andrei Khrennikov** International Center for Mathematical Modeling in Physics and Cognitive Sciences, Linnaeus University, Växjö, Sweden

**Heidi Kloos** University of Cincinnati, Cincinnati, OH, USA

**Gabriela Barreto Lemos** Instituto de Fisica, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil

**Martha Lewis** Department of Engineering Mathematics, University of Bristol, Bristol, UK

**Miguel López-Astorga** Institute of Humanistic Studies, Research Center on Cognitive Sciences, University of Talca, Talca, Chile

**Nicolás Marchant** Center for Social and Cognitive Neuroscience, School of Psychology, Universidad Adolfo Ibanez, Santiago, Chile

**Galina V. Paramei** Department of Psychology, Liverpool Hope University, Liverpool, UK

**Trung T. Pham** United States Air Force Academy, Colorado Springs, CO, USA  
University of Talca, Talca, Chile

**Oneris Rico** Cross Labs, Cross Compass Ltd, Kyoto, Japan

**Kathryn Schaffer** School of the Art Institute of Chicago, Chicago, IL, USA

**Olha Sobetska** Faculty of Social Sciences and Philosophy, Institute of Sociology, University of Leipzig, Leipzig, Germany

**Karl Svozil** Institute for Theoretical Physics, Vienna University of Technology, Vienna, Austria

**Bourama Toni** Department of Mathematics, Howard University, Washington, DC, USA

**Tomas Veloz** Departamento de Matemática, Universidad Tecnológica Metropolitana, Santiago, Chile

Interdisciplinary Foundation for the Development of Science, Technology and Arts, Santiago, Chile

Centre Leo Apostel, Vrije Universiteit Brussel, Brussels, Belgium

**Olaf Witkowski** Cross Labs, Cross Compass Ltd, Kyoto, Japan

**Xiaojing Yuan** University of Houston, Houston, TX, USA

# Introduction: Modern Approaches to the Study of Human Cognition



**Bourama Toni**

**Abstract** This introductory chapter presents some background and context for some of the techniques and approaches to cognitive modeling and simulations. It includes the summarizing ideas drawn for the contributed chapters to reveal the topical link between them while presenting hints as to the direction of future research. Indeed cognitive science has been endeavoring to harness the rigor and efficiency of mathematical modeling, in particular, in times of machine learning, artificial intelligence, and quantum computing and information science.

**Keywords** Human cognition · Qualitative modeling-agent-based modeling · Nyayasutra inference · Pheromone trail algorithm · Social Laser · Probabilistic learning · Physicalistic perspective · Conjunction fallacy · Quantum modeling · Compositional vector semantics · Optimality · Color perception

## 1 Background

Human cognition broadly includes perception, attention, memory, learning, reasoning, decision-making, critical thinking, and problem-solving. Its modeling and simulation, that is, the so-called *cognitive modeling and simulation*, necessitates powerful tools including the design and development of mathematical and computational models to decipher the underlying cognitive mechanisms allowing human thinking, behavior, and performance in all aspects of human life. Ultimately cognitive modeling and simulation aim at enhancing our understanding of how the brain processes data to retrieve information leading to human decision-making and related behavior.

*Cognitive modeling and simulation* are therefore major undertaking in psychology, neuroscience, computer science, artificial intelligence, education, and

---

B. Toni (✉)

Department of Mathematics, Howard University, Washington, DC, USA

e-mail: [bourama.toni@howard.edu](mailto:bourama.toni@howard.edu)

human-computer interaction. In these various fields, hypotheses and theories are formulated through mathematical or computational models and then tested and evaluated through simulations to replicate cognitive processes in the form of computer programs or virtual environments. Algorithms are developed and implemented on systems to perform human-like reasoning and decision-making. One of these well-known simulation techniques is the so-called agent-based modeling (ABM), an offshoot of game theory, to assess individual agents and interactions (strategic or otherwise) within a system (simple or complex). As in the broader game theory, ABM has served in the study of the spread of infectious diseases, the emergence of social norms and cooperation, financial markets dynamics, and evolutionary dynamics of biological systems. Social norms are self-enforcing patterns of behaviors, often sustained by multiple mechanisms, to include the desire to coordinate and follow the lead of others, the fear of being sanctioned, etc. Stochastic evolutionary game theory has proven to be the best cognitive model and simulator to study the resulting dynamics of social norms.

## 2 Current Approaches

*Cognitive modeling and simulation* are indeed very powerful tools, and sometimes they are the only tools to understand and predict human data and information processing and decision-making in interacting with complex systems. As a modern scientific endeavor, the theory of *cognitive modeling and simulation* has its challenges; one of these challenges is how to effectively harness the rigor and efficiency of mathematics to achieve the same results as the physical sciences. That is, the lack of an efficient analytical framework. For instance, psychology has been trying to be an exact science, but so far to no avail.

Note that precise measurement and predictions in the physical sciences occur after centuries of experimentation, leading, among other outcomes, to the creation of differential equations in mathematics.

Oftentimes in social and behavioral sciences, modelers resort to off-the-shelves mathematical methods, e.g., using real analysis instead of p-adic analysis, for lack of adequate training in advanced mathematics. Much like a drunk looking around the street lamppost for keys, he lost elsewhere only because “the light is better here.” Work by Khrennikov (Khrennikov 1998, 2000) on the dynamics of mental spaces is a good reference of the complexity of cognitive modeling and simulations. The author discusses the adequate mathematical model of mental space.

*Cognitive modeling and simulation* found an application in artificial intelligence (AI), which is basically the decoupling of intelligence from consciousness. And nonconscious intelligence is developing at a tremendous speed. We are witnessing the emergence of surprisingly powerful AI tools, such as ChatGPT, raising serious questions about the so-called *Mathematical Creativity*. Mathematics has been and is still a human construct. But for how long? AI may soon produce a very different sort of mathematics. A long-held belief has been that human cognition, i.e., human

mind and intelligence, has unique creativity capabilities unknown to animals and machines including computers. Von Neuman (1958) in his remarkable book *The Computer and the Brain* discussed the idea that computer/AI thinking, when this occurs, must be of a very different nature than human's thinking. This was also considered by Alan Turing in *Computing machinery and intelligence* (Turing 1950).

Computers and their capabilities have been assisting mathematical creativity for some time, performing, e.g., heavy logical and numerical tasks oftentimes beyond human capabilities: for example, the proof of the *four-color theorem* by Appel and Haken (1989). The Gosper's algorithm implemented on a symbolic manipulation program and using the *Wilf-Zeilberger pairs* (WZ) has produced new identities involving hypergeometric functions. A well-established trend is the so-called *computer-verified formal proof*. Human is known to have limited memory and to be prone to errors. The famous mathematicians Hadamard and Poincaré have seen doing mathematics as a combinatorial task leading to an elegant theorem by gluing pieces together. If so, then we will soon witness an *artificial mathematical creativity* by computer/AI.

The acquisition of language in the evolution of human cognition seems to have led to the ability to do mathematics, evolution favoring the ability to speak, the ability to count from 1 to 10 but possibly not the ability to master Galois Theory in higher mathematics. More on mathematical creativity and post-human mathematics could be found in Ruelle's book *The Mathematician's brain* (Ruelle 2007) which also inspired some of the above ideas. Ruelle has also coined the term *Artificial Mathematical Creativity*.

Human cognition went through revolutions transforming an insignificant African ape into the ruler of the world (creation of gods, corporation, cities, empires, writing and money, splitting of atom, reaching to the moon, genetic engineering, nanotechnology, brain-computer interfaces, etc.). Cognitive modeling aims at designing mathematical and computational models of the cognitive mechanisms underlying human behavior and then running these models to simulate human performance on numerous tasks. Many of these mechanisms have been perfected by the biological evolution of the human brain.

However, most of the research today involving cognitive modeling and simulation, i.e., about human mind and brain, is happening in the so-called WEIRD societies (western, educated, industrialized, rich, and demographic); that is to say it is fundamentally biased and flawed. The premise of this research endeavor is that, since Darwin's "On the origin of Species" organisms are seen as just biochemical algorithms similar to the electronic/silicone algorithms, sophisticated since Alan Turing. Indeed, exactly the same mathematical laws and principles apply to both, with the potential that the electronic algorithm will somehow outperform the biochemical one. Human cognitive activity amounts to processing data into information, information into knowledge, and knowledge into decision-making (wise or otherwise). In other words, human beings are reduced to data processing systems that need to be modeled and simulated for a deeper understanding. Cognitive science has been morphing into computational algorithmic science.

It should be emphasized here that cognitive modeling must be of a *qualitative* nature, i.e., not quantitative as for the physical sciences. As claimed by Levins (1974), scientific modeling can maximize at most two of the three virtues: *generality*, *realism*, and *precision*.

Sacrifice generality for precise quantitative predictions about specific systems and maximize realism by representing as many system details as possible.

Sacrifice realism to make unrealistic assumptions so systems can be described with general mathematically tractable equations producing precise quantitative predictions.

Sacrifice precision to abandon quantitative accuracy for qualitative relations between variables for maximum generality and realism.

Indeed in cognitive systems, the relevant information is of qualitative nature. It is usually impossible, infeasible, or impractical to determine the quantitative value or the precise functional form of many of the interactions between system parts, whereas it is often possible to determine the qualitative properties of these interactions. For example, in complex systems, what can be only ascertained is that there is or there is not interaction between the variables, which could be represented by yes or no, 0 or 1 (Boolean models). For instance, in psychology, an accurate mathematical function is not available to represent exactly human behavior, e.g., in imprecise belief states or social preferences.

In most complex systems including cognitive systems, relevant information resides in the rules of construct of the system, and not in the absolute quantitative values; what is being analyzed/investigated (data, phenomena, behavior, etc.) is essentially qualitative. Several qualitative modeling methods are available in the literature; we have proposed one based on the so-called Dynamical Roles of Jacobian Feedback Loops, in Toni (2014). See also Justus (2006) and Puccia and Richard (1985). In other words, qualitative analysis should be the main tool to understand complexity, key in the evolution of systems, versus the usual quantitative idealization of most mathematical models.

### 3 Conclusion

The volume features a variety of approaches and applications in the science of cognitive modeling and simulation. Drawing from the respective chapters' abstract, we highlight the outcomes of the study undertaken in each one of the chapters.

Enrique Canessa, Sergio Chaigneau, and Nicolas Marchant address in Chapter “[Use of Agent-Based Modeling \(ABM\) in Psychological Research](#)”, stressing, in particular, the infrequent use in psychology of a tool the authors deem the main research tool in social sciences. The authors present some general drawbacks and some specific to psychology; they include the benefits/advantages of using such a tool to outweigh its shortcoming.

The chapter titled Miguel Lopez-Astorga discusses “[Nyāyasūtra Proof Pattern: An Interpretation of Similarity as the Fact of Sharing Two Properties](#)”. The author proposes a relation to first-order calculus and explores the link between



the Nyayasutra, an inference by Schayer to first-order predicate logic, to Carnap's reduction sentences, showing that, despite their differences, the Indian inference and Carnap's reduction sentences have a similar potential for the analysis of scientific definitions.

The Chapter "[Using Pheromone Trail Algorithm to Model Analog Memory](#)" by Trung Pham, Ramon Castillo, Xiaojing Yan, and Heidi Kloos proposes such a use to model how the human mind registers data into memory in the brain. They also extend the model with algorithms to register and recall data embedded in an overlaid manner to represent the analog memory of a theoretical quantum computer. Numerical simulations are provided to illustrate the concept and to demonstrate the workability of the algorithms.

Andrei Khrennikov presents, in Chapter "[Review on Social Laser Theory and Its Applications](#)", an overview completed with new developments and applications, to include the detailed study of the dynamic interactions of the so-called *infons* with social atoms and the processes of absorption and emission.

The Chapter "[Challenges from Probabilistic Learning for Models of Brain and Behavior](#)" contains the contribution by Nicolas Marchant, Enrique Canessa, and Sergio Chaigneau. The authors discuss the historical background of probabilistic learning, its theoretical foundations, and its applications in various fields such as psychology, neuroscience, and artificial intelligence. They also review key findings from experimental studies on probabilistic learning, to include the role of feedback, attention, memory, and decision-making processes.

Karl Svozil discusses in Chapter "[The Emergence of Cognition and Computation: A Physicalistic Perspective](#)" to support the idea that cognition is an emergent property driven by dissipation. Cognitive agents are better equipped to acquire physical resources and means, giving them an advantage in survival and reproduction.

In Chapter "[Analysing the Conjunction Fallacy as a Fact](#)", Tomas Veloz and Olha Sobetska analyze *conjunction fallacy* (Tversky and Kahneman) range of factual possibilities. Reviewing samples of experiments between 1983 and 2016, the authors show that the majority of the related research has focused on a narrow part of the a priori factual possibilities, implying that explanations of the *conjunction fallacy* are fundamentally biased by the short scope of possibilities explored.

"[Yes Ghosts, No Unicorns: Quantum Modeling and Causality in Physics and Beyond](#)" is the imaginative title of the chapter by Kathryn Schaffer and Gabriela Barreto Lemos. Drawing from examples in physics, the authors urge caution in cross-disciplinary modeling comparisons and illustrate the kind of explanatory causal reasoning underlying Bell tests. They also argue that Bell inequalities are not portable: their bounds need to be re-derived and interpreted appropriately for each use.

In Chapter "[Compositional Vector Semantics in Spiking Neural Networks](#)", Martha Lewis proposes a way for compositional distributional semantics to be implemented within a spiking neural network architecture, with the potential to address problems in concept binding, and gives a small implementation. The author also describes a means of training word representation using labeled images.

Igor Douven and Galina Paramei report in Chapter “[Optimality, Prototypes and Bilingualism](#)” a study comparing Italian monolingual, English monolingual, and Italian-English bilingual speakers with regard to focal color choices in the BLUE region of color space suggesting that cultural and linguistic factors play a role in the categorical structuring of color space.

In Chapter “[The Dimensionality of Color Perception](#)”, Javier Fdez, Oneris Rico, and Olaf Witkowski study the trade-off between finding an embedding for color perception with the minimal number of dimensions while maximizing the discriminations between colors. They experiment with 13 subjects reporting the similarity between 20 colors randomly generated using the Munsell color system: their result is that the optimum number of dimensions is 3 when using a cosine similarity measure, indicating a resemblance to the way the perception of colors is cognitively encoded from mere physical properties of color maps.

**Acknowledgments** We express our sincere gratitude to all the contributors for their state-of-art research on such an important topic and to all the anonymous referees for their professionalism.

## References

- K. Appel and W. Haken. *Every Planar Map is four-Colorable*. AMS Providence, 1989
- J. Justus, *Loop Analysis and Qualitative Modeling: limitations and merits*. Biology and Philosophy, 21, 647–666, (2006)
- A. Khrennikov, *Human subconscious as the  $p$ -adic dynamical systems*. J. of Theor. Biology 193, 179–196, 1998
- A, Khrennikov,  *$p$ -adic discrete dynamical systems and collective behaviour of information states in cognitive models*. Discrete Dynamics in Nature and Society 5, 59–69, 2000
- R. Levins, *The qualitative analysis of partially specified systems*. Annals of the New York Academy of Sciences, 231, 123–138 (1974)
- J. von Neuman. *The Computer and the Brain*. Yale U.P., New Haven, 1958.
- C. Puccia and L. Richard. *Qualitative Modeling of complex systems*. Cambridge, MA: Harvard University Press. (1985).
- B. Toni, *Dynamical Roles of Jacobian Feedback Loops and Qualitative Modeling*. New Frontiers of Multidisciplinary Research in STEAM-H. Spring PROMS 90, 205–240 (2014)
- A. Turing. *Computing Machinery and Intelligence*. Mind 59, 433–460 (1950)
- D. Ruelle, *The Mathematician’s Brain*. Princeton U.P., Princeton, 2007

# Use of Agent-Based Modeling (ABM) in Psychological Research



Enrique Canessa, Sergio E. Chaigneau, and Nicolás Marchant

**Abstract** In this chapter, we introduce the general use of agent-based modeling (ABM) in social science studies and in particular in psychological research. Given that ABM is frequently used in many disciplines in social sciences, as the main research tool or in conjunction with other modeling approaches, it is rather surprising its infrequent use in psychology. There are many reasons for that infrequent use of ABM in psychology, some justified, but others stem from not knowing the potential benefits of applying ABM to psychological research. Thus, we begin by giving a brief overview of ABM and the stages one has to go through to develop and analyze such a model. Then, we present and discuss the general drawbacks of ABM and the ones specific to psychology. Through that discussion, the reader should be able to better assess whether those disadvantages are sufficiently strong for precluding the application of ABM to his/her research. Finally, we end up by stating the benefits of ABM and examining how those advantages may outweigh the potential drawbacks, thus making ABM a valuable tool to consider in psychological research.

**Keywords** Agent-based modeling (ABM) · Psychological phenomena · Cognitive dynamic

---

E. Canessa

Faculty of Engineering and Science, Universidad Adolfo Ibáñez, Viña del Mar, Chile  
e-mail: [ecanessa@uai.cl](mailto:ecanessa@uai.cl)

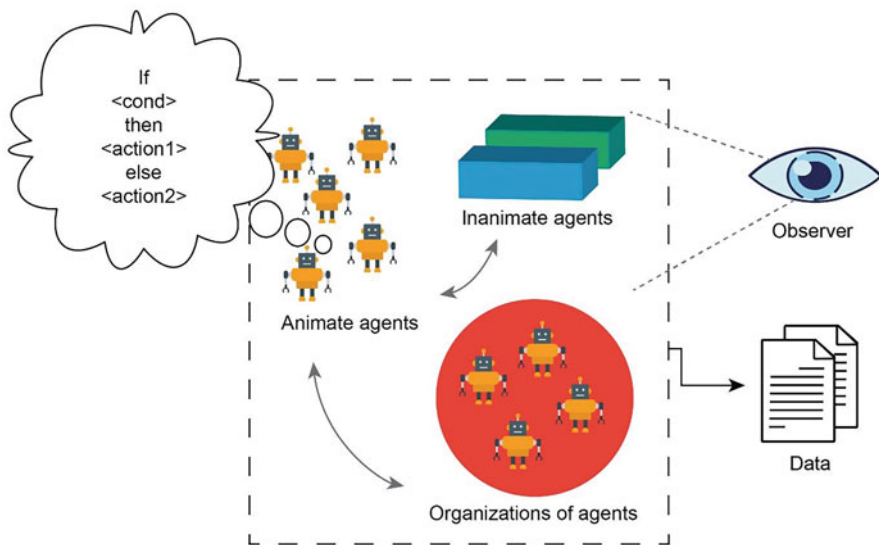
S. E. Chaigneau · N. Marchant (✉)

Center for Social and Cognitive Neuroscience, School of Psychology, Universidad Adolfo Ibáñez, Las Condes, Santiago, Chile  
e-mail: [sergio.chaigneau@uai.cl](mailto:sergio.chaigneau@uai.cl); [nicolas.marchant@edu.uai.cl](mailto:nicolas.marchant@edu.uai.cl)

## 1 A Brief Explanation of Agent-Based Modeling (ABM)

Agent-based modeling (ABM) is a computational analysis tool that simulates a system at various levels of detail and then uses the model to execute controlled experiments. Once the model is built, and similarly to empirical research, data is obtained from the model and analyzed through different statistical tools (ANOVA, regression equations, time-series analyses, and other methods better suited to analyzing nonlinear data) to draw conclusions about the system under investigation. The distinguishing feature of agent-based models (ABMs) is that they are constructed in a “bottom-up” manner, by defining the model in terms of entities and dynamics at a microlevel, that is, at the level of individual actors and their interactions with each other and with the environment (Canessa and Riolo 2006). An ABM consists of one or more types of agents (i.e., actors) and possibly a non-agent environment (see Fig. 1).

Agents are typically small computer code routines that encapsulate the behavior of individuals, i.e., the animate agents in terms of Fig. 1 (e.g., a model of legislators with some moral reasoning mechanisms used in deciding whether to approve or not a given law) or institutions (e.g., political parties), represented by the organization of agents in Fig. 1. Agents’ central characteristic is that they are capable of executing autonomous actions, by performing the *if ... then ... else ...* rules in Fig. 1, but note that rules can be much more complex, e.g., including reinforcement learning (RL) and/or artificial neural networks (ANN). The environment is the landscape where agents reside and interact among them and with the environment, represented



**Fig. 1** Components of an agent-based model

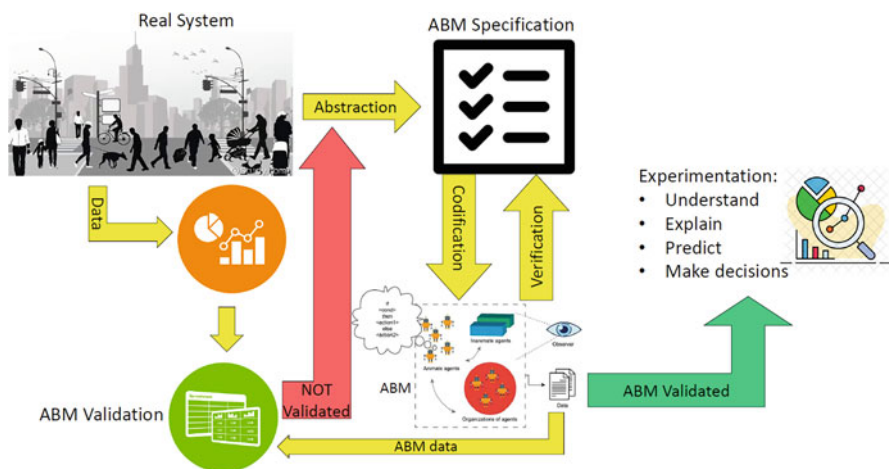
by the dashed-line rectangle in Fig. 1 (e.g., the legislative rules, which legislators need to abide by when processing laws). Note that the environment could have internal states and also evolve by means of internal rules and/or by interacting with animate agents. Additionally, some ABM practitioners find it useful to differentiate between animate agents (already explained) and inanimate agents, which are static and do not have behavioral rules, i.e., they do not change their internal states along time. All these components of an ABM make up an artificial world, which can be seen by a meta-agent called the Observer in Fig. 1. The Observer has access to all the components of the artificial world, including the internal states of the agents and of the environment, which allows it to collect data from this world. These data will then be analyzed to study the dynamics of the ABM. This flexibility makes it possible to study systems at many scales and to integrate parts into a coherent whole. The state of an agent can represent various characteristics, preferences, beliefs, memory of recent events, as well as particular social connections. Agent definitions include specification of their capabilities to carry out particular behaviors, as well as decision-making rules and other mechanisms that agents use to choose their own behaviors. Agents also may have adaptive mechanisms (learning or evolutionary) that allow them to change based on their experience. As an ABM is run, agent behavior is generated as agents make choices that determine which other agents to interact with and what to do in a given interaction. Thus, ABMs embody complex interlaced feedback relationships, leading to nonlinear, path-dependent dynamics. Note that the model's output is both the temporal patterns in the micro-behavior of agents as well as the emergent macro-level structures, relationships, and dynamics that result from correlated microlevel activity.

Because we can do controlled experiments in ABMs, they can be used to enhance our understanding of processes that can lead to the patterns we see in the systems being modeled. For example, we can explore the role of various factors that vary across people in their decision processes by explicitly representing those factors in our models at the microlevel and then observing, at the aggregate group level, how the model responds to variations in those factors. We also can explore alternative microlevel mechanisms, for example, agent decision rules, as well as alternative social network structures. By systematically exploring a variety of simple ABMs, constructed from different combinations of components and mechanisms, we can test hypotheses about the underlying processes that generate the data patterns we see, including both hypotheses that are suggested by theoretical assumptions about human behavior and hypotheses generated by analysis of data from the field.

Finally, ABMs can make predictions that can be tested about what to expect in situations and times not yet studied, as well as predictions about what aggregate patterns to expect in variables not yet observed, thus suggesting additional questions to ask and cases to study. Because ABMs are dynamic, we can examine behavior of the system over various time scales and track the state of the system during the "transients" rather than just looking at equilibrium or other "snapshot" states of the system. Understanding the dynamics of a system is critical for gaining a deeper understanding of how the system being represented gets to the snapshots

we obtain from empirical data. Because ABMs are computational models, they are formal, unambiguous, and thus replicable and testable (Axelrod 1997; Axelrod and Cohen 2000). However, they can be used to study aspects of systems that are difficult or impossible to study using traditional analytic (equation-based or game-theoretic) techniques (Parunak et al. 1998). All of these features of ABMs contribute to their suitability as a way to study the dynamical processes that characterize many psychological processes, both at the individual and the group levels.

In general, the development of an ABM consists of several stages, which are depicted in Fig. 2. For a more detailed treatment of this topic, a good paper to read is Macal and North (2010). First, and as with any model, we create an abstraction of the real system under study and write down an initial specification of the ABM. Then, we translate this specification into a computer code and make sure that the code reflects that specification, i.e., we verify that the code correctly implements the ABM. These two actions (codification and verification) are carried out until we are sure that there are no errors in the code. Now we have a verified ABM and proceed to validate it, i.e., make sure that the ABM represents the real system to a certain degree of detail/accuracy that allows us to use the ABM to study the real system. As Fig. 2 shows, validation is carried out by obtaining data from the ABM and from the real system and assess whether those data match to a desired degree. If that is the case, the ABM is validated. If not, we need to redo the specification of the ABM, by revising and accordingly changing the abstraction of the real system and performing again specification, codification, verification, and validation. After the ABM is validated, we execute controlled experiments, which allow us to get data from the ABM and analyze it to understand and/or explain the behavior of the real system, and perhaps predict it and make decisions regarding some relevant aspects of it. Regarding validation and doing controlled experiments with a model, note that



**Fig. 2** Stages in the development and analysis of a real system using ABM

there might not be a strict equivalent in psychological research. We will discuss those issues further along in the current chapter.

After the reader has gained a general vision of ABM, we next discuss the reasons why ABM is seldom used in psychological research.

## 2 Why Are ABMs so Seldom Used in Psychological Research?

Though ABM is widely used in conjunction with other analysis tools in many fields, its use in psychological research is much less frequent. To illustrate, here we give just a small glimpse of ABM's applications in different fields: in social sciences, economy (e.g., Tesfatsion 2002), anthropology (e.g., Kohler et al. 2005), business administration (e.g., North and Macal 2007), and sociology (e.g., Macy and Willer 2002); in life sciences, biology (e.g., Alber et al. 2003) and ecology (Mock and Testa 2007); and in engineering, traffic routing (e.g., Bazzan and Klügl 2014), optimization (e.g., Fikar et al. 2018), and socio-technical system design (e.g., Shah and Pritchett 2005). We believe that there are different reasons why ABMs have had little impact in psychology. In the next paragraphs, we explain those reasons, starting with the more general ones.

### 2.1 General Reasons of the Infrequent Use of ABM

*Codification of the ABM:* Like any computational model, an ABM is a computer program that needs to be specified at a macro- (e.g., how agents might aggregate in different groups or organizations) and microlevel (e.g., particular behaviors of the agents) and then coded in a certain programming language. As such, we need to express agents' behavior in detailed rules that must be translated into computer code. Hence, our model needs to be based on theory that can at least potentially be expressed in mathematical functions and/or logical rules. This may be a first hurdle in building an ABM if we do not know enough of the internal details of how the system works, which is typical for mental processes studied in psychology.

*Verification of the ABM:* Before any analysis is carried out with an ABM's output, we need to make sure that its specification is correctly translated into computer code. Fortunately, there exist useful software developing and testing practices that allow decreasing the probability of making errors when coding the ABM and afterward also checking that the code is free of errors and that it conforms to the model's specification (e.g., Pressman 2010). Given that many mental processes are very complex, representing them through rules, and then translating those rules to code, may require a detailed knowledge of those processes. However, in psychology, we frequently do not know sufficient details of the relevant processes. This may

preclude us from making a strict one to one mapping from general practices in ABM modeling in other fields to its use in psychology. We will discuss this topic in greater detail in the coming pages.

*Validation of the ABM:* While ABM and all formal models, in and of themselves, increase our knowledge about the behavior of any system consisting of similar processes, to use such models to make inferences about particular real-world systems requires model validation. Validation is the process by which we make sure that the ABM represents the real-world system closely enough, so that it enables us to answer our research questions, and thus, we can confidently draw conclusions that are scientifically informative. However, model validation for ABM is not trivial (Grimm et al. 2005; Grimm and Railsback 2005). ABMs generally have many input parameters, and thus the parameter space tends to be huge. Because interactions generally exist among parameters, validation requires simultaneously varying many parameters, which amounts to trying millions of possible combinations. Given that ABMs may effortlessly grow in size and number of parameters compared with other types of models, e.g., mathematical, this makes ABMs harder to validate. In general, it exists a relative consensus among ABM practitioners that a good procedure is to apply the *KISS* (Keep It Simple Stupid, Axelrod 1997) principle. This means that we should start with an as simple as possible model and then add details only if necessary (i.e., normally include more details only when the ABM does not reach validity). Finally, another possibility to ABM validation is to reach relational equivalence, i.e., matching patterns and relationships between the model and the system being modeled, rather than trying to match details (Axelrod 1997). That is, we assess whether the general behavior of the model matches that of the real system. For example, if the model shows that as we increase the value of a parameter, we observe an increase in an output variable of the model, then a similar change in the corresponding “parameter” in the real system should result in a similar change in a corresponding measured variable of the real system (Canessa and Riolo 2006).

*Reaching a useful trade-off between ABM’s detail and validation:* On the one hand, if we try to incorporate into an ABM a very detailed representation of the real-world system, this makes codification, verification, and validation extremely difficult. Contrarily, if we simplify the ABM too much, its codification and verification will be easier, but we may not achieve a sufficiently valid model. Thus, it is still an open issue how to proceed. In general, and as already mentioned, a good procedure is to apply the *KISS* principle, starting with a simple ABM and then incrementally add details only if necessary. However, this practice may also lead us to a *tinkering* pitfall (Murphy 2005). This problem arises when a researcher develops an ABM that in general successfully models a system (e.g., accounts for most of empirical data) but does not explain some other part of the same data or different data. Thus, the researcher incrementally modifies the ABM to better account for a larger portion of the data. However, in doing so, the researcher ends up with a model that naively puts together different processes (hopefully of the real-world system) without any theoretical consistency/coherence among them. This is similar to overfitting a regression model, where we can add more explanatory variables to



the regression and almost always obtain an increment in explained variance, but losing parsimony, generalization, and explanatory power.

*Replicability of ABM's results in the real system:* Closely related to validation issues, an ABM's results, though interesting and thought provoking, may not go beyond mere "thought experiments" (Axelrod 1997). Once an ABM is developed and verified, we may effortlessly begin doing many experiments, which is even an advantage of ABM. However, to draw sound conclusions from those experiments, we need to replicate them in the real system. Here, one may claim that after an ABM is validated, conclusions from ABM experimentation should also be valid. Nevertheless, as with all models, validity is always restricted to some subset of the parameter space (i.e., to some ranges of the values of the model's parameters) or to the context of the study. For example, in a decision-making study, if subjects must decide under time pressure (context of the study), the type of processing that could operate could be mainly associative. On the other hand, if the study does not involve decision-making under time pressure (a different context), a deliberative model might be more appropriate. The same happens with ABMs. Though we may establish an ABM's validity, this does not guarantee that we can always draw sound conclusions outside a close-enough range of parameter values within which we validated the ABM or, similarly, if we change the context of the phenomenon for which the ABM was developed. This is not to say that ABM's conclusions are not useful but that we always need to carefully ponder the use of ABM's results and, if doubts exist, try to replicate results in the real system. However, for several reasons, doing that might be difficult. Some examples of that are when there exist economical and/or ethical restrictions that prevents doing that (e.g., economical, test a vaccine against a disease in a whole continent in a short time period; ethical, letting a virus disease spread to see how many people die), or when it is impractical to do that (e.g., confine whole countries for years to see whether we can control the spread of a virus disease), or because it is very difficult to do some controlled experiments in reality (e.g., test whether people's specific patterns of movements in the real world correlate with the spread of a virus disease).

## ***2.2 Specific Psychological Research-Related Reasons of the Infrequent Use of ABM***

Additionally, to the above-discussed general reasons for why ABM is sometimes difficult to use in research, we believe that there are some other issues specific to psychology that hinder its utilization in the field.

*There might not be strict equivalents to model verification and validation in psychological research:* Throughout this chapter, we have hinted at general difficulties that may arise when trying to use ABMs to model psychological phenomena. Some of these difficulties relate to the issues of verification and validation. As we have discussed above, verifying and validating an ABM require access to the real-world system to a sufficient level of detail. In contrast, psychological models are

typically highly theoretical, such that many of their mechanisms are not directly measurable and remain hidden from the researcher (for practical reasons or even in principle). To illustrate, imagine a theory about logical reasoning that tries to explain how people draw conclusions from premises. Here, the experimenter would manipulate premises and measure conclusions, and the model (i.e., theory) offers a mechanism that might account for the observed regularities linking premises and conclusions. Contrary to a mostly transparent system that could be of interest in, e.g., engineering, systems under study in psychology typically remain opaque. Thus, the goal of using an ABM that is sufficiently close to the real system it models to carry out experiments that are not possible to carry out given practical or ethical considerations is a hard to achieve goal in psychological research. In consequence, we might argue that when ABMs are used in psychology, model fitting is what researchers can in practice do.

*Difficulty in fitting ABM's outputs to data:* In psychological research, and especially when developing computational models, there is an emphasis in using as a model's measure of success how well it fits the corresponding empirical data, i.e., model fitting. This is normally done by adjusting the free parameters of the model, so that its relevant outputs match as close as possible the empirical data. Models that are not amenable to such procedure may still be able to make qualitative predictions (e.g., predicting differences in means across experimental conditions) but are definitely less appealing. We already discussed validation and saw the difficulties in reaching them for ABMs. Hence, those difficulties can impact the extent to which model fitting can be done for ABMs and hinder a more straightforward application and acceptance of ABM in psychology.

*Difficulties in using model fitting as the main ABM's success measure:* There are some differences between validation, as described above, and model fitting as is generally done in psychology. In computational models in psychology, it is generally possible to describe model complexity by keeping track of the number of free parameters the model allows (i.e., more free parameters, more complexity). This means that models can be formally compared based on their number of free parameters (e.g., the Akaike Information Criterion, AIC, Akaike 1974). However, the number of free parameters is not necessarily a good proxy for an ABM's complexity. The researcher has many other decisions that increase the model's degrees of freedom and will not necessarily reflect in a model's free parameters. In ABMs, the model may involve many decisions about the simulated system's structure that are not strictly speaking considered parameters (e.g., agent's decision rules, environment structure). This is perhaps why the *KISS* principle (Axelrod 1997) is a mostly qualitative judgment regarding model complexity and is not strictly equivalent to the typical numerical model evaluation formulae available for computational models used in psychology (e.g., the AIC). This problem, together with the typical relative lack of access to the real-world system that wants to be modeled, makes evaluating an ABM's fit to data difficult to judge. Relatedly, the ideal of using an ABM to conduct controlled experiments with a validated model is generally difficult to achieve in psychological research, precisely because of the lack of access to the real-world system.

*Perception that ABMs are less formal than other types of models:* In psychology, there is a long tradition of modeling mental processes using mathematical models (e.g., see Murphy 2005) and also a more recent trend of using computational models, for example, based on neural networks (e.g., Murphy 2005). Although there is still controversy regarding the latter ones, when sensibly applied to psychological research, these computer-based models have helped to advance the knowledge of the field (Murphy 2005). However, ABMs are still perceived by many researchers as lacking the formality of mathematical models and even of computer-based models. Mathematical models tend to be tractable, unambiguous, and generally can be solved in closed form. Also, they are elegant and their internal consistency and external coherence with other models may be checked without too much effort. On the contrary, ABMs are harder to develop, verify, and especially validate and do not provide closed form solutions. Those ABM's characteristics create a perception that ABMs are not as formal as other types of models. However, we would like to argue that being a model makes ABMs similar to mathematical models and for some situations even better than mathematical models. In ABMs, the constituents of the model need not be homogeneous as generally assumed in mathematical models in order to solve them in closed form. Moreover, in ABM, we can represent systems with different feedback loops, local information processing parts, and adaptive mechanisms, leading to the nonlinear behavior of the system. This makes ABMs especially suited to analyzing and tracing the dynamics of the system's outputs, so as to learn how the system evolves in time. Finally, ABMs may be able to more simply express in computer code complex behaviors and interaction rules, freeing the researcher to better convey his/her theory in the model. Here, we must note that any model is an abstraction of reality, and thus the modeler needs to always leave out of it some aspects of reality. ABMs allows a researcher to more easily do that in incremental steps, pondering the consequences of his/her modeling choices and, if needed, include in the model mechanisms of the real system that would otherwise be difficult to express in mathematical form.

*Restricted vision of what an agent is:* It is intuitive to see a computational agent as a representation of an individual. In this view, an agent is an individual capable of executing autonomous actions based on some internal mental processes, i.e., based on some behavioral rules. Additionally, agents interact among them and with the environment based on those rules. Much of the use of the term agent in social sciences, and economy in particular, reinforces such preconceptions. Given that in psychology the focus is on studying the "internal processes" of an individual, that restricted view of an agent, which emphasizes the interaction among agents (i.e., individuals), does not match the psychological mainstream research. However, an agent may be much more general than just representing individuals. For example, an agent may be an institution or some other form of social group (e.g., schools and families, Canals et al. 2022). Additionally, even if agents represent individuals, there are psychological research to which that paradigm is applicable. For example, in our own work, we have applied ABM to study the listing process in a Property Listing Task (PLT), where agents list properties for a concept (Canessa et al. 2018; cf. Canessa et al. 2021). In that ABM, agents have a mental retrieval process based

on Atkinson and Shiffrin (1968) human memory and recall process (more recently revised in Malmberg et al. 2019), and we have been able to model the listing process dynamics for concrete vs. abstract concepts, obtaining relational equivalence in the listing process dynamics for such types of concepts (cf. Canessa et al. 2021). In another ABM, we have modeled the dynamics of the salience of stereotypes and negative stereotypes among groups of people that have power asymmetries between those groups (Lagos et al. 2019). That ABM allowed us to explore that phenomenon and preliminarily explain why negative stereotypes, although detrimental to the group that keeps them, are nevertheless preserved by that group. Finally, note that agents may also represent the mental components of an individual, e.g., the different executive processes of the retrieval process and its interaction with long-term memory, working memory, and so on. Hence, agents are not necessarily restricted to characterizing only individuals and the interactions among them.

### 3 Use of ABM in Psychological Research and Its Advantages

Having discussed the principal difficulties of applying ABM to psychological research, we end this chapter by considering why ABM is still a valuable and viable tool for the field. We concede that as part of our previous discussion, we have already presented some of the advantages of using ABM. However, we believe that a more explicit treatment of this matter will allow the reader to more sensibly decide whether his/her research may benefit from using ABM. We think that the reasons to use ABM and the related advantages may be divided into more conceptual and more practical/pragmatic ones. Thus, we begin presenting the conceptual reasons and then the practical ones.

#### 3.1 *Conceptual Benefits of Using ABM*

*Analytical vs. synthetic approach:* In general, in science, researchers normally use the “divide and conquer strategy,” i.e., given that a system under study may be huge and encompass many complex processes, we simply divide it into pieces and study those parts in relative isolation from each other. After we have obtained a reasonably good understanding of the parts, we try to connect those parts in a hopefully coherent whole model to analyze the real system. However, this approach assumes that those parts of the system do not strongly interact among them, so that they can be studied relatively independently from each other. In many systems, that is generally not the case (Banks 2002), and thus we need to explicitly analyze the interactions among the various components of the system. Hence, in an ABM approach, which we label *synthetic*, because we synthesize (i.e., build) the artificial world (system), it is possible to explicitly include and study those interactions.

*Equilibrium vs. Nonequilibrium supposition:* That social systems always get to an equilibrium state is a typical supposition in social sciences (e.g., in economy, markets get to equilibrium between the offer and demand of goods by adjusting the price of goods). Unfortunately, in many systems, the evolution and coadaptation of the system's entities lead to positive feedback loops, undermining such assumption. Under such situation, systems exhibit statistical macro regularities, although they show disorder at the microlevel. The abovementioned example of economic markets is an instance of that. Although we may see that markets somewhat regulate prices at the macro level (i.e., at the level of aggregate institutions), the behavior of prices at the microlevel (i.e., individuals) is far from equilibrium. Thus, to better understand and explain economic behavior and dynamics, we need to study nonequilibrium dynamics of the system, which is more easily accomplished by using ABM.

*Nomothetic vs. generative method:* In the nomothetic method of science, a researcher experiments with the real system, hopefully doing controlled experiments, obtains data and analyzes those data to study the system. But, what can we do if we cannot experiment with the system, as we already saw? Using ABM, we can explain the system's regularities by deriving the mechanisms that generate those regularities, generally at the microlevel. Axelrod labels this a *generative* approach (given that it generates the artificial system) and regards it as the "Third way of doing science" (Axelrod 1997). In the same vein, as Epstein says "If you haven't created it, you haven't explained it" (Epstein 1999). Of course, no practitioner of ABM claims that this generative method is always easy to do, but at least, it is another possibility to explore.

*Models based on variables vs. configurative ontology:* Generally, most models and especially mathematical models are based on variables and equations that relate those variables to each other. Nevertheless, using variables and equations, it is sometimes very difficult to capture the complex behavior of social systems. Moreover, generally equations assume homogeneity of the different parts of the system, so that those parts can be represented by variables. Given that a system may have a configuration of interactions among actors and aggregate level structures, a sensible model of the system needs to account for those characteristics. With ABM, it is relatively easy to endogenize those actors and interactions, leading to a more sensible representation of reality.

### **3.2 Practical Benefits of Using ABM**

We hope that by now, the reader may have appreciated some of the advantages of considering using ABM in his/her research. Now we complement them with the next more practical/pragmatic advantages of applying ABM.

*Use a simulated system when it is impossible to study the real system:* Oftentimes, we cannot experiment with the real system because of ethical, economical, and/or practical issues. Thus, with an ABM, we can at least open a window to the system

and perform controlled experiments. Of course, that assumes that the ABM was verified and validated. Regarding validation, which might be the most difficult part of the process, remember that validation may be done at various levels of detail. Here, achieving relational equivalence might be a good and valuable alternative. However, we recognize that in psychology, mainstream research uses model fitting to assess validity, i.e., the model is valid as long as it replicates empirical data to a certain degree of accuracy. Thus, given that relational equivalence only replicates patterns in data, not the exact data points, some researchers might regard relational equivalence as not a formal enough method for assessing validity. However, even if model fitting is not completely possible, the ABM may be used to gain insights about the real system and do “thought experiments.” These “thought experiments” may be especially valuable when they lead to find counterintuitive results that may be considered critical predictions of the model.

*Attrition of sample in longitudinal studies:* In social sciences, longitudinal studies are based on small sample sizes and a few measurements of variables over time (Quezada and Canessa 2010). Additionally, since in many studies the analysis relies on the specific context of the situation, it is difficult to generalize the findings. Those problems are also reinforced by the attrition of the sample along the time span of the study. With ABM, we can get as much data as needed, which allows a very precise characterization of the dynamics of the system (Quezada and Canessa 2010). And we already explained why the analysis of real-world systems needs to include the dynamics of such systems: because of nonequilibrium and to assess how the system gets to the states/equilibria we are observing.

*Documentation and formalization of the model:* Given that we must translate the specification of the ABM to computer code, all of the processes, behavioral rules, and components of the model must be precisely defined, without any doubt regarding how the description of the various parts is implemented. That formalizes the model and helps assess whether we must think in greater detail the verbal specification of the model. Anything that is not amenable to be coded must be specified with greater detail. Moreover, the computer code provides for a deeper documentation of the model, and thus, the ABM’s details are more transparent to researchers. This allows a better and more in-depth scrutiny and critique of the model.

*Explain and understand the model in a research group:* Most research is conducted by groups of scholars and always the mutual understanding, and discussion of the model among group members is a concern. How can we make sure that all the group members share the same understanding of the model, especially if the model is rather huge and/or complex? Given that the ABM is finally implemented in computer code, that code is an unambiguous and detailed description of it. That allows exposing the different views about the model and better get an informed consensus about the model. Additionally, if someone disagrees with some parts of the model, it is relatively easy to change the corresponding computer code and see the consequences of those changes on the outputs of the model and corresponding conclusions. Closely related to that aspect is that we can perform sensitivity analysis of the various parts of the ABM. If we change some parts of the ABM and the outputs do not vary too much, or if that variation does not alter the conclusions,

then we can say that those parts are not so important and the model is robust to those variations and that the conclusions are more generalizable.

*Effectively and efficiently communicate analyses and results:* Closely related to the preceding point is that an ABM allows for a more effective and efficient communication of analyses and results to the research community. First, results are easily conveyed to an audience given that ABM's development platforms have nice graphic user interfaces (GUI), which allow building many different views of the artificial world and plots of many ABM's outputs of interest. That helps visualize the behavior of agents and the dynamics of the outputs in real time. Moreover, all of that can be done at various levels of analysis at the same time. Second, if someone wants to do "what if" type of analysis, that can be done by simply changing the settings of the corresponding parameters of the ABM (by moving the respective sliders of the GUI) and immediately observing the impact of those changes on the relevant outputs of the model. Third, if someone questions parts of the ABM, and the associated computer code is rather easy to modify to accommodate that critique, we can immediately deal with those concerns, assessing whether the criticism is really warranted or not. A substantial change in ABM's output and/or conclusions will tell us that the observation is worth further consideration.

Finally, we would like to say that this chapter does not intend to oversell ABM as the sole cure for all the problems empirical and other types of research exhibit but honestly present the benefits and disadvantages of ABM. By having those antecedents, the reader must ponder them and assess whether the advantages outweigh the drawbacks for the research he/she wants to conduct. Our own experience and work with ABM show that indeed ABM is worth considering in psychological research.

## References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723.
- Alber, M.S., Kiskowski, M.A., Glazier, J.A. & Jiang, Y. (2003). On cellular automaton approaches to modeling biological cells. In: Rosenthal J and Gilliam DS (eds). *Mathematical Systems Theory in Biology, Communication, and Finance*, IMA Vol. 134, Springer: New York, pp 1–39.
- Atkinson, R.C. & Shiffrin, R.M. (1968). Human memory: a proposed system and its control processes. In K.W. Spence & J.T. Spence (Eds.), *The psychology of learning and motivation: Advances in research and theory*. (Vol. 2). New York: Academic Press. 89–195.
- Axelrod, R. & Cohen, M.D. (2000). *Harnessing Complexity*, Chicago: The Free Press.
- Axelrod, R. (1997). *Advancing the Art of Simulation in the Social Sciences*, in R. Conte, R. Hegselmann and P. Terna (eds.) *Lecture Notes in Economics and Mathematical Systems: Simulating Social Phenomena*, Berlin: Springer-Verlag.
- Bankes, S.C. (2002). Tools and Techniques for Developing Policies for Complex and Uncertain Systems, *PNAS* 99(3): 7263–7266.
- Bazzan, A.L.C. & Klügl, F. (2014). A review on agent-based technology for traffic and transportation. *The Knowledge Engineering Review*, 29, 3, 375–403 <https://doi.org/10.1017/S0269888913000118>



- Canals, C., Maroulis, S., Canessa, E., Chaigneau, S. & Mizala, A. (2022). Mechanisms Underlying Choice-Set Formation: The Case of School Choice in Chile. *Social Science Computer Review*. <https://doi.org/10.1177/08944393221088659>
- Canessa, E., Chaigneau, S. & Moreno, S. (2021). Language processing differences between blind and sighted individuals and the abstract versus concrete concept difference. *Cognitive Science*, 45, 2021, <https://doi.org/10.1111/cogs.13044>
- Canessa, E., Chaigneau, S. & Barra, C. (2018). Developing and calibrating an ABM of the Property Listing Task. *Proceedings of the 32nd European Conference on Modelling and Simulation, ECMS 2018, Wilhelmshaven, Germany, May 22nd to May 25th, 2018*, pp. 13–19.
- Canessa, E. & Riolo, R. (2006). An agent-based model of the impact of computer-mediated communication on organizational culture and performance: an example of the application of complex systems analysis tools to the study of CIS. *Journal of Information Technology*, 21, 272–283. <https://doi.org/10.1057/palgrave.jit.2000078>
- Epstein, J.M. (1999). Agent-based computational models and generative social science. *Complexity*, 4, 5, 41–60. [https://doi.org/10.1002/\(SICI\)1099-0526\(199905/06\)4:53.0.CO;2-F](https://doi.org/10.1002/(SICI)1099-0526(199905/06)4:53.0.CO;2-F)
- Fikar, C., Hirsch, P. & Nolz, P.C. (2018). Agent-based simulation optimization for dynamic disaster relief distribution. *Central European Journal of Operations Research* 26, 423–442. <https://doi.org/10.1007/s10100-017-0518-3>
- Grimm, V. & Railsback, S.F. (2005). *Individual-Based Modeling and Ecology*, Princeton, NJ: Princeton University Press.
- Grimm, V., Revilla, E., Berger, U., Jeltsch, F., Mooij, W.M., Railsback, S.F., Thulke, H.-H., Weiner, J., Wiegand, T. & DeAngelis, D.L. (2005). Pattern-Oriented Modeling of Agent-Based Complex Systems: Lessons from ecology, *Science* 310(5750): 987–991.
- Kohler, T.A., Gumerman, G.J. & Reynolds, R.G. (2005). Simulating ancient societies. *Scientific American* 293(1): 77–84.
- Lagos, R., Canessa, E., Chaigneau, S. (2019). Modeling stereotypes and negative self-stereotypes as a function of interactions among groups with power asymmetries. *Journal of Theory Social Behavior*, 2019, pp. 1–22, <https://doi.org/10.1111/jtsb.12207>
- Macal, C.M. & North, M.J. (2010). Tutorial on agent-based modelling and simulation. *Journal of Simulation* 4, 151–162. <https://doi.org/10.1057/jos.2010.3>
- Macy, M.W. & Willer, R. (2002). From factors to actors: Computational sociology and agent-based modeling. *Annual Review Sociology* 28: 143–166.
- Malmberg, K.J., Raaijmakers, G.W.J. & Shiffrin, R.M. (2019). 50 years of research sparked by Atkinson and Shiffrin (1968). *Memory & Cognition*. <https://doi.org/10.3758/s13421-019-00896-7>
- Mock, K.J. & Testa, J.W. (2007). *An Agent-based Model of Predator-Prey Relationships between Transient Killer Whales and Other Marine Mammals*. University of Alaska Anchorage, Anchorage, AK, 31 May 2007. <http://www.math.uaa.alaska.edu/Borca/>.
- Murphy, G. L. ( 2005 ). The study of concepts inside and outside the lab: Medin vs. Medin. In W. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, & P. Wolff (eds.), *Categorization Inside and Outside the Lab: Essays in Honor of Douglas Medin* (pp. 179–195). Washington, DC : APA .
- North, M.J. & Macal, C.M. (2007). *Managing Business Complexity: Discovering Strategic Solutions with Agent-Based Modeling and Simulation*. Oxford University Press: Oxford, UK.
- Parunak, V., Savit, R. & Riolo, R. (1998). Agent-based Modeling and Equation based Modeling: A case study and users guide, in *Proceedings of Workshop on Multi-Agent Systems and Agent-based Simulation*; Berlin: Springer.
- Pressman, R. (2010). *Software Engineering: a practitioner’s approach*, 7th ed., McGraw-Hill: New York, NY.
- Quezada, A. & Canessa, E. (2010). Modelado basado en agentes: una herramienta para complementar el análisis de fenómenos sociales. *Avances en Psicología Latinoamericana*, 28, 2, 226–238
- Shah, A. P., & Pritchett, A. R. (2005). Agent-Based Modeling and Simulation of Socio-Technical Systems. In *Organizational Simulation* (pp. 323–367). John Wiley & Sons, Inc. <https://doi.org/10.1002/0471739448.ch12>
- Tesfatsion, L. (2002). Agent-based computational economics: Growing economies from the bottom up. *Artificial Life* 8(1), 55–82.



# *Nyāyasūtra* Proof Pattern: An Interpretation of Similarity as the Fact of Sharing Two Properties



Miguel López-Astorga 

**Abstract** In the book *Nyāyasūtra*, there is an inference that Schayer related to first-order predicate logic. The relation was challenged, because it seemed to ignore important components of that Indian inference. The present paper also proposes a relation to first-order predicate calculus. However, it comes to the new relation following contemporary approaches about human reasoning: the theory of mental models and case-based reasoning. The main idea is that the chief concept in the inference in the *Nyāyasūtra* is that of similarity. If that concept is understood as the fact of sharing two properties, it is possible to link the inference to Carnap's reduction sentences. This allows thinking that, although they belong to different intellectual traditions, the Indian inference and Carnap's reduction sentences have a similar potential for the analysis of scientific definitions.

**Keywords** Carnap · Inference · *Nyāyasūtra* · Reasoning · Reduction sentence

## 1 Introduction

In 1824, an Indian inference coming from Colebrooke's translation of the *Nyāyasūtra* was presented (e.g., Ganeri 2004). From then on, different accounts of the inference, which will be called here NPP (from the expression *Nyāyasūtra* proof pattern), have been given. The intention of those accounts was to understand the real nature of NPP. One of the accounts has interpreted NPP based on first-order predicate logic (Schayer 1933). Another example is the explanation that has related NPP to a contemporary psychological framework: the theory of mental models (López-Astorga 2016, 2022).

---

M. López-Astorga (✉)

Institute of Humanistic Studies, Research Center on Cognitive Sciences, University of Talca, Talca Campus, Talca, Chile  
e-mail: [milopez@utalca.cl](mailto:milopez@utalca.cl)

Those two accounts will be described below. However, the present paper will also consider the link provided between an Indian inference such as NPP and “case-based reasoning” (Ganeri 2004). What is important about this link is that it allows capturing an essential element in NPP: similarity. NPP establishes that, if two elements, element  $a$  and element  $b$ , are similar, both of them have property  $P_1$ , and  $b$  has property  $P_2$ , and  $a$  must have property  $P_2$  too. As indicated below, the relation between NPP and case-based reasoning enables to show that similarity consists of sharing one more property. On this basis, this paper deems the mentioned properties as predicates in first-order predicate calculus. Thus, it presents an interpretation within first-order predicate logic different from the previous one (i.e., different from that of Schayer 1933). The advantage of this new interpretation is that it permits similarity in NPP to be expressed as a “reduction sentence” (Carnap 1936, 1937). Because Carnap proposed reduction sentences to draw up exact scientific definitions, this reveals the potential that NPP has in science.

A relevant point to clarify is that the present paper is not intended to support the idea that NPP was an anticipation of the concept of reduction. It does not even claim Schayer’s (1933) idea that NPP anticipated modern predicate logic. What will be assumed here is that NPP has characteristics and a context that are not necessarily compatible with current modern logic and philosophy of science. But the paper will try to argue that both Eastern and Western traditions have intellectual tools to develop science. If reduction sentences can be built from NPP, the role of reduction sentences can be played by NPP as well. Carnap (1936, 1937) proposed reduction sentences to improve scientific definitions. Therefore, if the relations between reduction sentences and NPP are correct, NPP can also improve those definitions. The cognitive consequences of all of this will be also indicated.

The first section will describe what NPP is. The second one will be devoted to the interpretation of NPP from first-order predicate calculus that was presented in 1933 (Schayer 1933) and its limitations. The third one will explain the link provided between NPP and a current psychological theory: the theory of mental models (López-Astorga 2016, 2022). Then, Ganeri’s (2004) account relating NPP to case-based reasoning will be shown. Lastly, it will be argued that the link to case-based reasoning allows deriving reduction sentences from the schema of NPP.

## 2 The Five Steps of NPP

There are several ways to explain the general structure of NPP. The present paper will follow that of Ganeri (2004). According to that account, NPP has five steps: *Thesis*, *Reason*, *Example*, *Application*, and *Conclusion*. *Thesis* expresses what is wished to prove, for example, that  $a$  has property  $P_2$ , that is, (1).

$$P_2a \tag{1}$$

*Reason* shows why it is thought that (1) holds: because  $a$  has property  $P_1$  too, that is, because (2) holds too.

$$P_1 a \quad (2)$$

*Example* supports the derivation referring to an element similar to  $a$ , element  $b$ , which has both property  $P_1$  and property  $P_2$ , that is, referring to the fact that (3) and (4) are the case as well.

$$P_1 b \quad (3)$$

$$P_2 b \quad (4)$$

In *Application*, it is checked that (2) is true, which leads to *Conclusion*, where (1) is admitted (see Ganeri 2004).

In other words, the assumption is (1) (*Thesis*). That is proposed because of (2) (*Reason*). There are facts supporting this idea: (3) and (4) (*Example*). Given that (2) is the case (*Application*), (1) can be accepted (*Conclusion*).

From Western perspective, NPP describes an induction process, or, perhaps better still, a generalization process. From the characteristics of one element ( $b$ ), it is concluded that elements similar to that element (e.g.,  $a$ ) should have the same characteristics. If this is the manner NPP is understood, it seems to move away from Western classical logic. Nonetheless, an interpretation of NPP within first-order predicate logic was presented in 1933.

### 3 NPP from First-Order Predicate Logic

Schayer (1933) gave that interpretation. As Ganeri (2004) reminds, following Schayer, two basic rules in first-order predicate calculus underlie NPP: the universal quantifier elimination rule and *Modus Ponendo Ponens*. If  $P_1$  and  $P_2$  are deemed as two predicates in first-order predicate logic, what Schayer proposes is that the inference implicit in NPP is (5).

$$\{P_1 a, \forall x (P_1 x \rightarrow P_2 x)\} \therefore P_2 a \quad (5)$$

In (5), “ $\forall$ ” is the universal quantifier, “ $\rightarrow$ ” represents implication, and “ $\therefore$ ” stands for logical deduction.

Deduction (5) is trivial in first-order predicate calculus. Premise  $\forall x (P_1x \rightarrow P_2x)$  can be transformed into (6) by virtue of the universal quantifier elimination rule.

$$P_1a \rightarrow P_2a \tag{6}$$

And *Modus Ponendo Ponens* allows deriving (1), which is the conclusion in (5), from (2), which is the first premise in (5), and (6).

However, a possible criticism against this explanation is that it ignores that element  $b$  is an essential component in NPP. Conclusion (1) is obtained because of  $b$  and its similarity to  $a$ . If this is ignored, the essence of NPP might be being ignored too (see also, e.g., López-Astorga 2016).

This problem does not appear to be solved even resorting to frameworks offered by contemporary philosophy of science and language. For example, the “semantic method of extension and intension” (Carnap 1947) does not eliminate the difficulties. One might think that method can complement Schayer’s (1933) explanation: it can include the role of element  $b$  by means of the concept of L-equivalence. But this is not clear. The method is based on modal logic and uses the concept of state description, that is, possible world. A state description is “. . . a complete description of a possible state of the universe of individuals with respect to all properties and relations . . .” (Carnap 1947, p. 9). This enables to define L-equivalence:  $a$  and  $b$  are L-equivalent if and only if its properties and relations are the same in all state descriptions. So, if  $a$  and  $b$  are L-equivalent and  $b$  has a particular property,  $a$  must have that very property.

Nevertheless, the difficulties here are at least two. On the one hand, to know that (1) is the case, that is, that it has  $P_2$ , it would be necessary only to know that (4) is the case, that is, that  $b$  also has  $P_2$ . L-equivalence means that if one of the L-equivalent elements has a property, that property should be had by the other L-equivalent elements as well. Hence, it would be necessary to consider neither (2) nor (3). But both (2) and (3) are key components of NPP, which reveals that this explanation does not capture what NPP tries to provide. On the other hand, according to the method of extension and intension, if two elements are L-equivalent, those elements have the same intension (Carnap 1947, p. 23). This implies that the two elements are identical. But the relation in NPP is not identity; it is similarity (e.g., López-Astorga 2016).

Therefore, Schayer’s (1933) interpretation from first-order predicate logic does not appear to be the most suitable interpretation. Another interpretation from that very logic will be presented below. First, it is important to note that NPP has been related to contemporary psychological approaches too. The intention has been to show that NPP is closer to the manner people think than Western logic.

## 4 NPP and the Theory of Mental Models

One of those relations is that between NPP and the theory of mental models (López-Astorga 2016, 2022). The theory of mental models is a general theory about the human mind and language (e.g., Khemlani et al. 2018). Only a small part of it is relevant here: the way it deals with induction (e.g., Johnson-Laird 2012).

The latest versions of the theory claim that, given a sentence expressing information, people analyze a “conjunction of possibilities”: the conjunction of possibilities that can be admitted if the sentence holds (see also, e.g., Johnson-Laird et al. 2021). Reviewing the possibilities can lead to make different inferences, including inductions. If Person<sub>1</sub> is a person that finds a rabbit and wants to feed that rabbit, Person<sub>1</sub> needs to know what food is good for a rabbit. An alternative can be carrots. There are four possibilities that can relate the fact that an animal is a rabbit to the fact that the animal likes carrots:

Possible (this animal is a rabbit and this animal likes carrots) (7)

And possible (this animal is a rabbit and this animal does not like carrots) (8)

And possible (this animal is not a rabbit and this animal likes carrots) (9)

And possible (this animal is not a rabbit and this animal does not like carrots.) (10)

In conjunction of possibilities (7) to (10), the key is conjuncts (7) and (8). (7), (9), and (10) allow coming to conditional (11).

If this animal is a rabbit, then this animal likes carrots. (11)

On the other hand, (8), (9), and (10) lead to conditional (12).

If this animal is a rabbit, then this animal does not like carrots. (12)

Possibilities (9) and (10) are not relevant. They are presuppositions (see also, e.g., Espino et al. 2020). They hold whether (11) or (12) are the case. The key point is the choice between (7) and (8). Following the theory of mental models, Person<sub>1</sub> can prefer (7) because, according to general knowledge, it is the most probable scenario.

This is the mental process generally characterizing inductions (for similar accounts with other examples, see, e.g., Johnson-Laird 2012; López-Astorga 2022).

It has been stated that the mental process described in NPP is akin to this one. The basic idea is that, if presuppositions are ignored (they will be always admissible), the chief possibilities in NPP are two (López-Astorga 2016, 2022):

$$\text{Possible } [P_1a \ \& \ P_2a \ \& \ P_1b \ \& \ P_2b \ \& \ (a \cong b)] \quad (13)$$

where “ $x \cong y$ ” means that  $x$  is similar to  $y$ .

$$\text{And possible } [P_1a \ \& \ \neg P_2a \ \& \ P_1b \ \& \ P_2b \ \& \ (a \cong b)] \quad (14)$$

where ‘ $\neg$ ’ indicates negation.

The difference between (13) and (14) is just that (1) does not hold in (14). It can be thought that, by virtue of a process akin to that explained by the theory of mental models, individuals choose (13) because it is more likely (see López-Astorga 2016, 2022).

More links between NPP and daily reasoning are possible. For instance, NPP can represent the way doctors reach diagnoses.

## 5 NPP and Case-Based Reasoning

The relation between NPP and case-based reasoning has been explained linking Indian inferential processes in the *Nyāyasūtra* to medical diagnosis processes:

The physician observing a patient  $A$  who has, for example, eaten a certain kind of poisonous mushroom, sees a number of associated symptoms displayed, among them  $F$  and  $G$ , say. He or she now encounters a second patient  $B$  displaying a symptom at least superficially resembling  $F$ . The physician thinks back over her past case histories in search of cases with similar symptoms. She now seeks to establish if any of those past cases resembles  $B$ , and on inquiry into  $B$ 's medical history, discovers that  $B$  too has consumed the same kind of poisonous mushroom. These are her grounds for inferring that  $B$  too will develop the symptom  $G$ , a symptom that had been found to be associated with  $F$  in  $A$  (Ganeri 2004, p. 328; italics in text).

What is interesting about this passage is that it shows how the relation of similarity can be understood. To note that, only the following equivalences require to be assumed:

$$A = a$$

$$F = P_1$$

$$G = P_2$$

$$B = b$$

The quotation reveals that what leads to relate  $a$  to  $b$  is not just the fact of sharing  $P_1$  ( $F$  in text) but also the fact of taking the same action: to eat certain type of poisonous mushroom. The act of eating the poisonous mushroom can be deemed as property  $P_3$ . Thus, it can be said that what makes  $a$  similar to  $b$  is that both of them have properties  $P_1$  and  $P_3$ , that is, that (2), (3), (15), and (16) hold.

$$P_3a \tag{15}$$

$$P_3b \tag{16}$$

If all of this is taken together with the fact that (4) is also the case, (1) can be inferred. From this point of view, the relation of similarity is clarified: it is sharing two properties (e.g.,  $P_1$  and  $P_3$ ).

## 6 NPP and Reduction Sentences

If similarity is sharing two properties, a new account based on first-order predicate logic can be given. Possibilities (13) and (14) have a dark point: what their last conjunct (i.e.,  $a \cong b$ ) means. The link in the previous section between case-based reasoning and NPP helps understand this conjunct, that is, what similarity between  $a$  and  $b$  is. Similarity is not sharing only property  $P_1$  but also property  $P_3$ . From this perspective, (13) and (14) can be better expressed as (17) and (18).

$$\text{Possible } (P_1a \ \& \ P_2a \ \& \ P_3a \ \& \ P_1b \ \& \ P_2b \ \& \ P_3b) \tag{17}$$

$$\text{And possible } (P_1a \ \& \ \neg P_2a \ \& \ P_3a \ \& \ P_1b \ \& \ P_2b \ \& \ P_3b) \tag{18}$$

Now, the fact that the properties shared by  $a$  and  $b$  are two, and not just one, justifies to a greater extent the election of (17) as the most probable alternative, and hence the inductive process.

On the other hand, if  $P_1$ ,  $P_2$ , and  $P_3$  are deemed as predicates in first-order predicate logic, this psychological process can lead to think that the correct inference is not (5) but (19).

$$\{P_1a, P_3a, \forall x [(P_1x \wedge P_3x) \rightarrow P_2x]\} \therefore P_2a \quad (19)$$

In (19), “ $\wedge$ ” stands for conjunction.

The third premise in (19), which is valid in first-order predicate calculus, that is, (20), captures the idea that the properties that should be shared in similarity are two. That is what allows assigning the third property to every element having those two properties.

$$\forall x [(P_1x \wedge P_3x) \rightarrow P_2x] \quad (20)$$

Against this, an argumentation is possible. It can be stated that the theory of mental models is not classical logic. In fact, it is not even a logic (see also, e.g., Johnson-Laird 2010). For this reason, the relation between formulae such as (20) to possibilities such as (17) or (13) may not be right. However, what the present paper is supporting is not that formulae such as (20) logically derive from possibilities such as (17). The idea is that, if the theory of mental models is deemed as a correct description of inductive processes in human beings, logicians, who are human beings, can come, via induction, to think that the most appropriate formula for situations such as that described in NPP is (20).

Another interesting point with regard to (20) is that it can be transformed into (21).

$$\forall x [P_1x \rightarrow (P_3x \rightarrow P_2x)] \quad (21)$$

This is interesting because (21) has the logical form of reduction sentences in Carnap (1936). Those sentences were established to build and improve scientific definitions. So, if the link provided between (21) and NPP is admissible, that means that, beyond its initial aim, NPP already had the potential to create and qualify scientific definitions. This does not imply that NPP can be understood as a precedent of the concept of reduction. It only shows that Indian logical tradition also has resources in the field of philosophy of science. It has had those resources for a long time.

Finally, one more objection could be that a condition must hold in order that (21) is a reduction sentence. (22) must be the case as well.

$$\exists x (P_1x \wedge P_3x) \quad (22)$$



In (22) “ $\exists$ ” is the existential quantifier.

Nevertheless, if the relation between NPP and case-based reasoning is correct, both  $a$  and  $b$  satisfy (22): (2), (15), (3), and (16) hold.

## 7 Conclusions

There is an inference within Indian logic with the same potential as Carnap’s reduction sentences. It is NPP. This does not imply that NPP provided the basis for the concept of reduction. NPP and Carnap’s reduction sentences come from different cultural contexts. That only means that the same results can be obtained from both NPP and reduction sentences.

A link can be established resorting to well-formed formulae in first-order predicate logic. But Schayer’s framework does not seem to be the best approach in this way. Indian logic in general and NPP in particular should be considered as proposals trying to describe natural processes of human thinking, not as pieces included in systems akin to current logical systems. Following the accounts in the literature, this appears to be the most important point in the relations between NPP and both the theory of mental models and case-based reasoning.

The induction process as understood by the theory of mental models seems to be pretty close to NPP. Nonetheless, the account relating NPP and the theory of mental models causes one difficulty: it is not clear what similarity between two elements sharing properties is. That difficulty is removed by the description referring to case-based reasoning. Following the latter, it is possible to consider similarity to be the fact of sharing two properties. Thereby, NPP can be seen as an inferential process that enables to conclude that an element has a property because that very element has other two properties. The grounds are that the same happens with other elements: they have the three properties, that is, the property that is concluded and the two properties leading to the third one.

Both the theory of mental models and case-based reasoning are approaches trying to capture natural inferential activity in human beings. Given that logicians are human beings, one might assume that, if those approaches are right, logicians should think as those very approaches indicate. This in turn enables to suppose that logicians can consider the inference corresponding to NPP in first-order predicate logic not to be that proposed by Schayer, but another one including three predicates. After this point, it is not hard to come to a reduction sentence as proposed by Carnap.

This is the process allowing indirectly linking NPP and Carnap’s concept of reduction. As indicated, this does not mean that NPP and reduction sentences come from similar frameworks or that there are theoretical commonalities between them. Their cultural, philosophical, and historic contexts are different. This paper makes only one point: if certain contemporary approaches about human reasoning can be accepted, NPP and reduction sentences can be used in a similar way in the analysis and review of scientific definitions.

Thus, it can be thought that both NPP and reduction sentences can be used to create scientific definitions and to improve those definitions. If this is the case, both of them can also help build all kinds of definitions. The process to construct definitions, whether scientific or not, should always take the same steps.

Besides, the potential of NPP in cognitive science cannot be forgotten. As shown in the literature, it can be related to both the theory of mental models and case-based reasoning. The relation to the theory of mental models corresponds to the induction process. As pointed out above and in the literature, that relation reveals an important point: if the theory of mental models actually describes human reasoning, NPP also offers a description of human reasoning (at least, as far as induction is concerned). The link to case-based reasoning also accounts for in this sense. Ultimately, it shows that cases such as those in which physicians make diagnoses are captured by NPP too. Furthermore, the present paper can lead NPP to the field of communication as well. The paper argues that NPP can be useful with regard to scientific definitions. But, as also indicated, the way all types of definitions are understood and processed by people should not be very different from the way human beings understand and process scientific definitions.

**Acknowledgments** Programa de Investigación Asociativa en Ciencias Cognitivas, Centro de Investigación en Ciencias Cognitivas, Instituto de Estudios Humanísticos, Universidad de Talca.

Fondo Fondecyt de Continuidad para Investigadores Senior, código FCSN2102, Universidad de Talca.

Fondequip (Programa de Equipamiento Científico y Tecnológico) 2019, código EQM190153.

## References

- Carnap, R. (1936). Testability and meaning. *Philosophy of Science*, 3(4), 419–471. <https://doi.org/10.1086/286432>
- Carnap, R. (1937). Testability and meaning – Continued. *Philosophy of Science*, 4(1), 1–40. <https://doi.org/10.1086/286443>
- Carnap, R. (1947). *Meaning and Necessity: A Study in Semantics and Modal Logic*. Chicago, IL: The University of Chicago Press.
- Espino, O., Byrne, R. M. J., & Johnson-Laird, P. N. (2020). Possibilities and the parallel meanings of factual and counterfactual conditionals. *Memory & Cognition*, 48, 1263–1280. <https://doi.org/10.3758/s13421-020-01040-6>
- Ganeri, J. (2004). Indian logic. In D. M. Gabbay & J. Woods (Eds.), *Handbook of the History of Logic, Volume 1. Greek, Indian and Arabic Logic* (pp. 309–395). Amsterdam, The Netherlands: Elsevier. [https://doi.org/10.1016/S1874-5857\(04\)80007-4](https://doi.org/10.1016/S1874-5857(04)80007-4)
- Johnson-Laird, P. N. (2010). Against logical form. *Psychologica Belgica*, 50(3/4), 193–221. <https://doi.org/10.5334/pb-50-3-4-193>
- Johnson-Laird, P. N. (2012). Inference with mental models. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford Handbook of Thinking and Reasoning* (pp. 134–145). New York, NY: Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199734689.013.0009>
- Johnson-Laird, P. N., Quelhas, A. C., & Rasga, C. (2021). The mental model theory of free choice permissions and paradoxical disjunctive inferences. *Journal of Cognitive Psychology*, 33(8), 951–973. <https://doi.org/10.1080/20445911.2021.1967963>

- Khemlani, S., Byrne, R. M. J., & Johnson-Laird, P. N. (2018). Facts and possibilities: A model-based theory of sentential reasoning. *Cognitive Science*, 42(6), 1887–1924. <https://doi.org/10.1111/cogs.12634>
- López-Astorga, M. (2016). The Hindu Syllogism, iconic representations, and human thought. *Revue Roumaine de Philosophie*, 60(2), 351–358.
- López-Astorga, M. (2022). Induction in human reasoning: Gautama's Syllogism and system K. *Philosophia: International Journal of Philosophy*, 23(2), 355–365. <https://doi.org/10.46992/pijp.23.2.a.7>
- Schayer, S. (1933). Altindische Antizipationen der Aussagenlogik. *Bulletin International de l'Academie Polonaise des Sciences et des Lettres, Classe de Philologies*, 90–96.

# Using Pheromone Trail Algorithm to Model Analog Memory



Trung T. Pham, Ramón D. Castillo, Xiaojing Yuan, and Heidi Kloos

**Abstract** This chapter proposes the use of pheromone trail algorithm to model how the human mind registers data into the memory in the brain. In this approach, each time data is registered, a representation in the memory is marked with an additive quantity of pheromone to represent the reinforcement in the memory. When a new data is registered, the pheromone of the previously registered data is evaporated to represent the fading in the memory. While this work is intended to support the modeling of human memory where data are stacked up on each other, this model is extended with the algorithms to register and recall data embedded in an overlaid manner to represent the analog memory of a theoretical quantum computer. Numerical simulations are provided to illustrate the concept and to demonstrate the workability of the algorithms.

---

This work was supported by the Associative Research Program in Cognitive Science financed by the University of Talca, Talca, Chile.

---

T. T. Pham (✉)

United States Air Force Academy, Colorado Springs, CO, USA

University of Talca, Talca, Chile

e-mail: [trung.pham@usafa.edu](mailto:trung.pham@usafa.edu); [tpham@utalca.cl](mailto:tpham@utalca.cl)

R. D. Castillo

Centro de Investigación en Ciencias Cognitivas of the University of Talca, Talca, Chile

e-mail: [racastillo@utalca.cl](mailto:racastillo@utalca.cl)

X. Yuan

University of Houston, Houston, TX, USA

e-mail: [xyuan@Central.UH.EDU](mailto:xyuan@Central.UH.EDU)

H. Kloos

University of Cincinnati, Cincinnati, OH, USA

e-mail: [heidi.kloos@uc.edu](mailto:heidi.kloos@uc.edu)

## 1 Introduction

Memory is the storage area that retains data that have been placed there in a process of registering (Patterson and Hennessy 2016; Null 2018). The data registered in the memory can be recovered for later use in another context (Kokosa 2018; Hall and Slonka 2018). In computer sciences, digital memory has been developed electronically to imitate the functionalities of storing and recovering data in human memory (Milton 2015). However, beyond the functional level of storing and recovering, the internal details of how to register and to recover data in the memory are completely different between digital memory and human memory. While human memory seems to have no limit (Baddeley 2013), digital memory always has physical limitation in its capacity and will stop registering new data when it is filled up to its capacity. With the purpose of extending the capacity of digital memory beyond its physical limitation, the modeling of human memory has always been an inspiring topic for studying in the discipline of computer sciences.

Digital memory has been designed around the digital circuit unit of a flip-flop (LaMeres 2019; Harris and Harris 2012) that can alternate between two clearly different voltage values. This flip-flop is used as a memory unit to store one bit of binary data (Agresti 2018; Andriesse 2018). A memory bank consists of as many flip-flop units as the manufacturing process can squeeze into a physical form factor. The data are saved in a block and recovered according to the information of that block. Up to certain point, the space in the memory bank is completely used up, and the storing function will stop working. With recent development in quantum physics, a theoretical model of a quantum computer has been outlined (Liu et al. 2019; Hidary 2019) where a unit of analog memory capable of containing real values between zero and one is used to construct a memory bank. This model suggests potential for an unlimited capacity of memory but at the same time implies that the traditional methods of registering data in the memory and recovering them from the memory no longer apply.

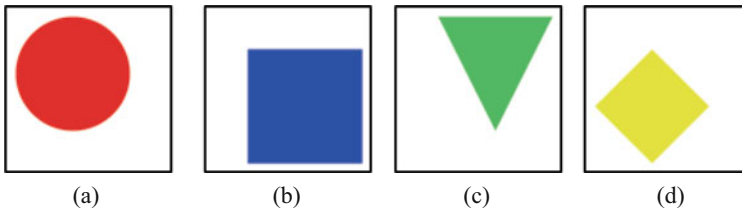
This paper proposes the modeling of a phenomenon commonly observed about the human memory: the more often an object is seen, the more vivid it appears in the memory later (Elsey et al. 2018; Kahana et al. 2018). In this memory model, images are stored in the memory in the order of observation, in the manner of being superimposed over the existing data in the memory. In this model, the memory is an image consisting of many images in superposition. When an image is superimposed in the memory, it is marked with a quantity of pheromone (Lv et al. 2017; Yang et al. 2019) that will evaporate with time. The image with more pheromone will appear more vivid in the superposition. This model is used by psychologists to visualize human memory in their study of the memorization process in the human minds. In this work, an algorithm to assign pheromone to each image when it is registered in the memory is derived in the manner that reflects how vivid it appears in the memory as a function of time. Furthermore, this work considers the research question: is it possible to recover an image stored in this type of memory based on some vague description of the image? In the context of this study, images of simple nature that

can be described by their colors are used to develop a concept base, and images of more complex nature will be used in future work in an incremental approach (Hibbs et al. 2009; Singh and Kaur 1998).

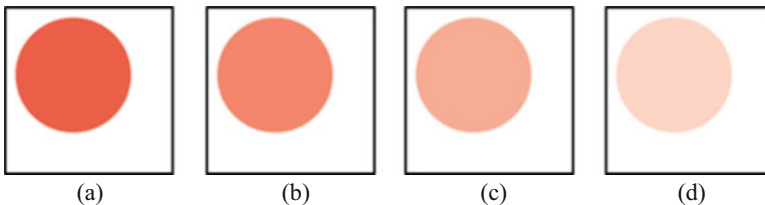
## 2 Background

Consider the simple images shown in Fig. 1 where each image can be uniquely characterized by its shape and color. The RGB (red, green, blue) coding is used in this work due to its additive characteristic that is consistent with the additive characteristic of the pheromone proposed in this study. Considering the observation that the memory of an image is faded with time (Wingfield and Byrnes 2013; Bliss et al. 2017) as shown in Fig. 2, combined with the testimony that images seem to be blended together (Radvansky 2017; Murray et al. 2020) in the memory as shown in Fig. 3, the modeling of the memory in this study is to construct an image that can contain data of other images in a superposition according to the order of their registrations.

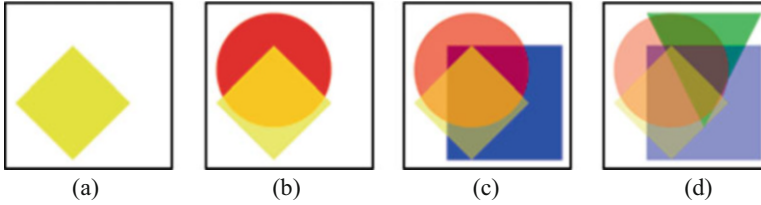
In the proposed model, time is an additional dimension and the time mark when an image is registered in the memory will determine the level of how vivid it will appear afterward. For this reason, a convention of fading the color as a function of time is needed. Furthermore, a convention of blending the colors faded in different



**Fig. 1** Images of different shapes and colors that can be uniquely described are considered for the initial stage of the study. **(a)** red circle. **(b)** blue square. **(c)** green triangle. **(d)** yellow rhombus



**Fig. 2** Image can be faded with time in human memory in the manner that some portion of its original image is retained. **(a)** 80% retained. **(b)** 60% retained. **(c)** 40% retained. **(d)** 20% retained



**Fig. 3** Images can be blended into a memory in a superposition while being faded. (a) 1st image. (b) 2nd image. (c) 3rd image. (d) 4th image

levels in a superposition to construct a visualization of the model that permits more profound study of human memory is needed. However, due to an additional objective in this study that answers the research question, “is it possible to recover an image stored in this type of memory based on some vague description of the image?,” the model is designed in a manner so that an algorithm can be developed for recovering an image based on its color that represents a vague and incomplete description of the image.

The memory model that permits squeezing more data into a limited memory area also has application in quantum computing that tries to extend the capacity of the storage while maintaining the physical size of the hardware. In this endeavor, it is important that an image, when being faded, is retained in some manner without being completely eliminated from the memory. In this aspect, the function that describes the rate of fading must be decreased in its value with the advancement of time but must not be equal to zero. Furthermore, when an image is stored in the memory, it implies that in any moment later, the image can be recovered in its original format without any loss of information.

In the context of this work, an image is characterized by two attributes: color and shape. However, when a person recalls an image stored his/her memory, the query normally contains only a portion of the attributes, and the recovered image reveals the rest of the attributes. For replicate this phenomenon in the model of the analog memory, the query for recovering an image shown in Fig. 1 that is already stored in the memory shown in Fig. 3 will contain a color, and the expected result is an original image that reveals the other attribute of the image.

### 3 Mathematical Model

In this study, the model of the quantum memory consists of a sequence of analog bits in which each bit is capable of containing a floating number with real value between zero and one. For storing an image of dimension  $N \times M$  in which each pixel indexed by the integers  $n$  and  $m$ ,  $1 \leq n \leq N$  y  $1 \leq m \leq M$  consists of three components  $(r_{n,m}, g_{n,m}, b_{n,m})$  of the colors red, green, and blue occupying 24 digital bits, 3 analog bits  $(\rho_{n,m}, \gamma_{n,m}, \beta_{n,m})$  are assigned for a pixel.

In the digital memory bank, each image occupies a block of  $24 \times N \times M$  digital bits (the header containing more information about the coding is negligible and is not important in the context of this study), and this block is not usable for other images. In the model of the quantum memory, a block of  $3 \times N \times M$  analog bits can be used for storing many images in the superimposed manner illustrated by Fig. 3. For the reason explained later for the purpose of recovering data, each image must be faded differently.

There are three main parts in the development of the analog memory: fading function, algorithm to store an image, and algorithm to recover an image. In this section, a sequence of images  $I_1, I_2, \dots, I_K$  is considered, with each image  $I_k$ ,  $1 \leq k \leq K$ , of dimension  $N \times M$ . Each pixel  $p_k(n,m)$  of the image  $I_k$  has three components  $(\rho_{k,n,m}, \gamma_{k,n,m}, \beta_{k,n,m})$  representing the colors red, green, and blue, respectively. When a color is faded, its value is reduced by a multiplicative factor  $\lambda$ , where  $0 \leq \lambda \leq 1$ , in the manner that  $\rho_{k,n,m}(\text{faded}) = \lambda\rho_{k,n,m}(\text{original})$ ,  $\gamma_{k,n,m}(\text{faded}) = \lambda\gamma_{k,n,m}(\text{original})$ , and  $\beta_{k,n,m}(\text{faded}) = \lambda\beta_{k,n,m}(\text{original})$ .

#### A. Algorithm to Store an Image

Considering a sequence of images  $I_1, I_2, \dots, I_K$  being registered in an analog memory  $\Omega$  of dimension  $N \times M$ , the registration is done sequentially according to the order of appearance. The empty memory is denoted  $\Omega(0)$ , and the memory that contains  $k$  images is denoted  $\Omega(k)$ . In general, the memory  $\Omega(k)$  must contain data of the images  $I_1, I_2, \dots, I_K$  in the manner

$$\Omega(k) = \sum_{q=1}^k \lambda_q I_q \quad (1)$$

so that each of  $I_1, I_2, \dots, I_k$  can be recovered. In this context, each of the constants  $\lambda_1, \lambda_2, \dots, \lambda_k$  must be uniquely selected in a sense that

$$\lambda_1 \neq \lambda_2, \lambda_2 \neq \lambda_3, \dots, \lambda_{k-1} \neq \lambda_k. \quad (2)$$

For the purpose of visualizing the memory, the constants  $\lambda_1, \lambda_2, \dots, \lambda_k$  must reflect the concept that the image registered more recently is more vivid than the image registered in the distant past. This requirement is converted into the condition:

$$\lambda_1 < \lambda_2 < \dots < \lambda_k, \quad (3)$$

where the image  $I_k$  is the image registered most recently, and the image  $I_1$  is the first image registered in the memory.

For the computational purpose of satisfying the property of the analog memory where a bit of data used for a color of a pixel, or the value of each pixel color, must be equal to or less than one:



$$0 \leq \Omega_\rho(k)(m, n) = \sum_{q=1}^k \lambda_q \rho_q(m, n) \leq 1, \quad (4a)$$

$$0 \leq \Omega_\gamma(k)(m, n) = \sum_{q=1}^k \lambda_q \gamma_q(m, n) \leq 1, \quad (4b)$$

$$0 \leq \Omega_\beta(k)(m, n) = \sum_{q=1}^k \lambda_q \beta_q(m, n) \leq 1, \quad (4c)$$

where  $\Omega_\rho(k)(m, n)$ ,  $\Omega_\gamma(k)(m, n)$ , and  $\Omega_\beta(k)(m, n)$  are the components red, green, and blue of the pixel indexed by  $(n, m)$  of the memory  $\Omega(k)$ , and  $\rho_q(m, n)$ ,  $\gamma_q(m, n)$ , and  $\beta_q(m, n)$  are the components red, green, and blue of the pixel  $(n, m)$  of the image  $I_q$ . Due to the requirement that the results of (4) must be equal to or less than one, the constants  $\lambda_1, \lambda_2, \dots, \lambda_k$  must satisfy:

$$\lambda_1 + \lambda_2 + \dots + \lambda_k = 1. \quad (5)$$

Furthermore, each time an image is registered in the memory, these constants must be reduced and still satisfying (3) and (5).

Due to the reducing nature of the constants  $\lambda_1, \lambda_2, \dots, \lambda_k$ , the use of the pheromone trail algorithm, commonly used in the ant colony optimization (Patnaik et al. 2017; Solnon 2013), is proposed for assigning values to these constants. In this sense, the values of  $\lambda_1, \lambda_2, \dots, \lambda_k$  represent the quantities of pheromone used to mark the images  $I_1, I_2, \dots, I_k$  when they are registered in the memory. Then, the memory is calculated as

$$\Omega(k) = (1 - \varepsilon(k)) \Omega(k - 1) + \lambda_k I_k, \quad (6)$$

where  $\varepsilon(k)$  is the function representing the evaporation of the existing pheromone at the moment that the image  $I_k$  is marked with the quantity  $\lambda_k$  of pheromone. This function is developed in the next subsection.

The memory model (6) permits the simulation of human memory under the hypothesis that the memory registered most recently or the image registered most repeatedly will be the image marked with the highest quantity of pheromone that will appear most vividly in the visualization of such memory.

### B. Evaporation Function

In computer programming, a computational cycle is the period  $\Delta t$  of time in which the central processing unit can perform a basic task. Therefore, the evaporation function  $\varepsilon(\cdot)$  is developed as a function of time that is changed with the advancement a period  $\Delta t$  in the timeline. According to (3), it is proposed that

$$\lambda_1 = \theta_k, \lambda_2 = 2\theta_k, \dots, \lambda_k = k\theta_k \quad (7)$$

for  $\Omega(k)$ , therefore (5) becomes

$$\theta_k + 2\theta_k + \dots + k\theta_k = 1, \quad (8)$$

where  $\theta_k > 0$  is a constant with value

$$\theta_k = 2/(k)(k+1). \quad (9)$$

According to (9),  $\theta_{k-1} = 2/(k-1)(k)$ . Furthermore, (6) implies that

$$\theta_k = (1 - \varepsilon(k)) \theta_{k-1}, \quad (10)$$

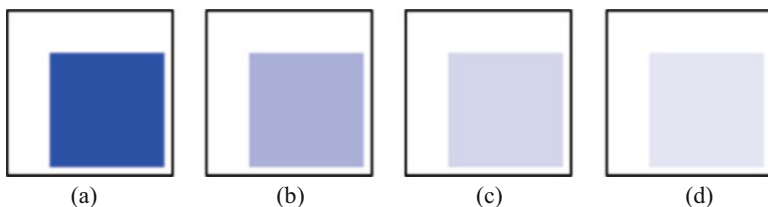
that, when substituted with (9) and its derived  $\theta_{k-1} = 2/(k-1)(k)$ , yields

$$\varepsilon(k) = 1 - \frac{k-1}{k+1} = \frac{2}{k+1}. \quad (11)$$

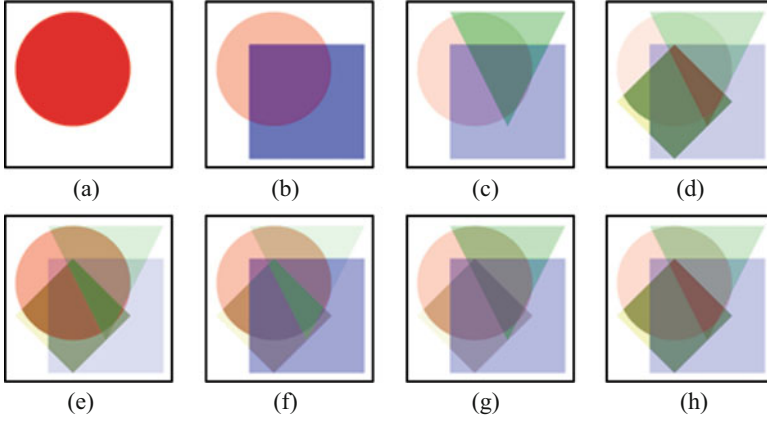
The evaporation function  $\varepsilon(k)$  in (11) is based on the quantities of pheromone defined in (7) that implies the level of fading of the images in the memory  $\Omega(k)$  is spread out on the equidistant basis. It is important to note that when  $k = 1$ , the memory is empty and, therefore, there is no existing pheromone, and in this case, the evaporation function  $\varepsilon(k)$  can take any value without loss of generality. Equation (11) implies the value of 1 for  $\varepsilon(1)$ , meaning the total elimination of the residual data that might exist at the beginning. Figure 4 visually demonstrates how an image fades according to (11), and Fig. 5 demonstrates how the memory  $\Omega(k)$  is visualized with four images from Fig. 1 being registered two times according to (6) with the evaporation function (11).

### C. Algorithm to Recover an Image

The problem of recovering an image embedded in the memory  $\Omega(k)$  containing  $k$  images is defined based on the partial description of the image. In this stage



**Fig. 4** Image may fade according to the evaporation function in the scheme of using pheromone to mark its level of fading in the memory. (a)  $\varepsilon(1) = 1$ . (b)  $\varepsilon(2) = 2/3$ . (c)  $\varepsilon(3) = 1/2$ . (d)  $\varepsilon(4) = 2/5$



**Fig. 5** The analog memory containing images superimposed by the scheme of the trace algorithm of pheromone can be visualized. (a)  $t = \Delta t$ . (b)  $t = 2\Delta t$ . (c)  $t = 3\Delta t$ . (d)  $t = 4\Delta t$ . (e)  $t = 5\Delta t$ . (f)  $t = 6\Delta t$ . (g)  $t = 7\Delta t$ . (h)  $t = 8\Delta t$

of proving the concept, each of the images in the memory has two descriptions: shape and color. In the context of this section, the color of an image is given for the recovering of the image. This problem is equivalent to the problem in which a human tries to recall something with a specific color in his/her memory. The result of the recovering should be an image in its original entity that reveals the other description of its shape.

The problem of recovering an image can be defined as given an image with color of RGB values  $(\zeta_r, \zeta_g, \zeta_b) \in \{(\iota_1, \iota_2, \iota_3) \mid \iota_1, \iota_2, \iota_3 \in \{1, 0\}\}$ , find the image embedded in the memory  $\Omega(k)$  that has this color, and recover it in its original entity.

Considering the three components in each pixel of the analog memory, there is a total of 8 possible unique combinations:  $(0, 0, 0)$ ,  $(0, 0, 1)$ ,  $(0, 1, 0)$ ,  $(0, 1, 1)$ ,  $(1, 0, 0)$ ,  $(1, 0, 1)$ ,  $(1, 1, 0)$ , and  $(1, 1, 1)$ . The combination  $(0, 0, 0)$  cannot be used in (4) because the value  $\lambda_q$  cannot be registered when multiplied by zero. The combination  $(1, 1, 1)$  cannot be used because it is the color white reserved for the background of each image. In the memory  $\Omega(k)$ , each pixel is represented by

$$\begin{bmatrix} \rho(n, m) \\ \gamma(n, m) \\ \beta(n, m) \end{bmatrix} = \iota_a \eta_a \theta_k \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + \iota_b \eta_b \theta_k \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + \iota_c \eta_c \theta_k \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} + \iota_d \eta_d \theta_k \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + \iota_e \eta_e \theta_k \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} + \iota_f \eta_f \theta_k \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \quad (6)$$

where  $\eta_a\theta_k, \eta_b\theta_k, \dots, \eta_f\theta_k$  are the quantities of pheromone assigned to the basic colors, and  $\iota_a, \iota_b, \dots, \iota_f$  are the binary indicators defining the presence of a color in the pixel. When an indicator is 1, it means that the pixel contains the corresponding color. Equation (6) is a set of three equations with twelve variables that has an infinite number of solutions. However, with the knowledge of the six basic colors, it is possible searching in the memory  $\Omega(k)$  for the pixels that only contain a basic color and discover their associated quantities of pheromone  $\eta_a\theta_k, \eta_b\theta_k, \dots, \eta_f\theta_k$ . In doing this task, Eq. (6) is reduced to contain six binary variables  $\iota_a, \iota_b, \dots, \iota_f$ . Furthermore, Eq. (6) becomes three scalar integer equations

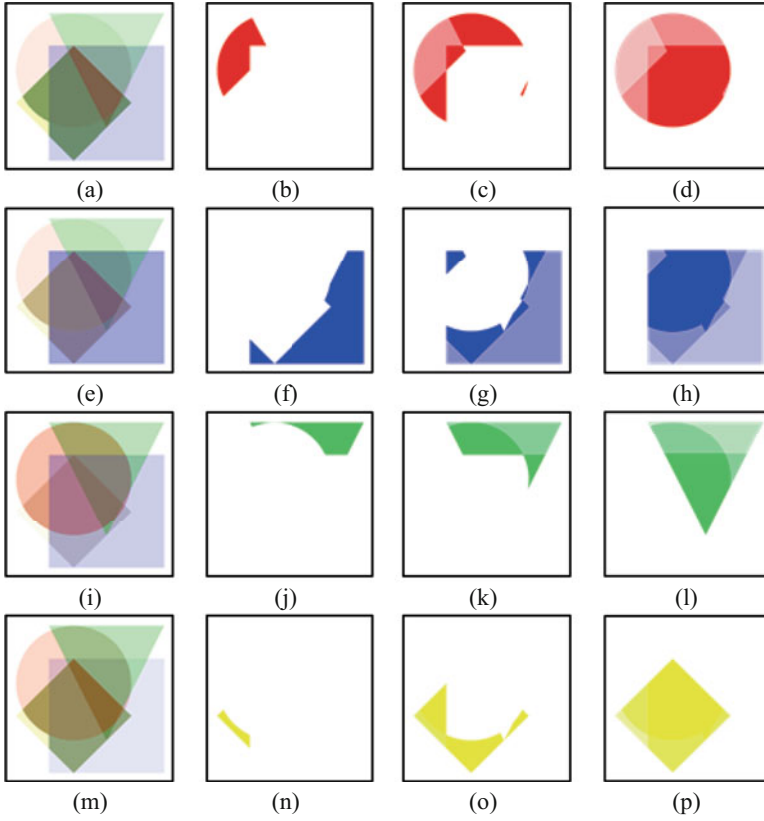
$$\alpha_r = \eta_1\iota_1 + \eta_4\iota_4 + \eta_6\iota_6, \quad (7a)$$

$$\alpha_g = \eta_2\iota_2 + \eta_4\iota_4 + \eta_5\iota_5, \quad (7b)$$

$$\alpha_b = \eta_3\iota_3 + \eta_5\iota_5 + \eta_6\iota_6, \quad (7c)$$

where they can be tested for the presence of a specific color  $q, 1 \leq q \leq 6$ , by setting  $\iota_q = 1$  and evaluate (7a), (7b), and (7c) with the 32 possible combinations of the remaining indicators  $\iota_1, \iota_2, \dots, \iota_{q-1}, \iota_{q+1}, \dots, \iota_6$ . If there exists a combination in which (7a), (7b), and (7c) are true, the presence of the color  $q$  in the pixel  $(m, n)$  of the memory  $\Omega(k)$  is detected.

Figure 6 shows the detection of a color in each pixel of the memory  $\Omega(4)$  that contains data shown in Fig. 5 in different configurations. The recovered pixels are shown in different scenarios: when only one color exists, when a color coexists with another color, and when a color coexists with many colors in the memory. These scenarios permit a more profound analysis of the algorithm of recovering data by identifying if there is a situation of uncertainty in which there are more than one combination that can satisfy (7a), (7b), and (7c). The results in Fig. 6 show the complete recovering of an image when embedded in a memory containing a total of four images. Figure 7a–d are additional data to test the recovering algorithm: Fig. 7a, b for increasing the capacity of the memory to contain six images. Figure 7e–l show complete recovery in this situation. Figure 7c, d are for testing the situation when a color (red or blue) is used for two different images. Figure 7m–p show the result in which only a portion of the image is recovered. In common sense, this situation is expected because the algorithm depends on the value of pheromone marked in an image, and here there are two values of pheromone for a color. For this reason, Eq. (6) is modified when more than a value of pheromone is detected for a color: the values  $\iota_a, \eta_a$ , and  $\theta_k$  of a color vector are split into sub-values, and the algorithms will examine (6) as containing more than six basic color vectors. This modification is reviewed in the next section of numerical simulation.

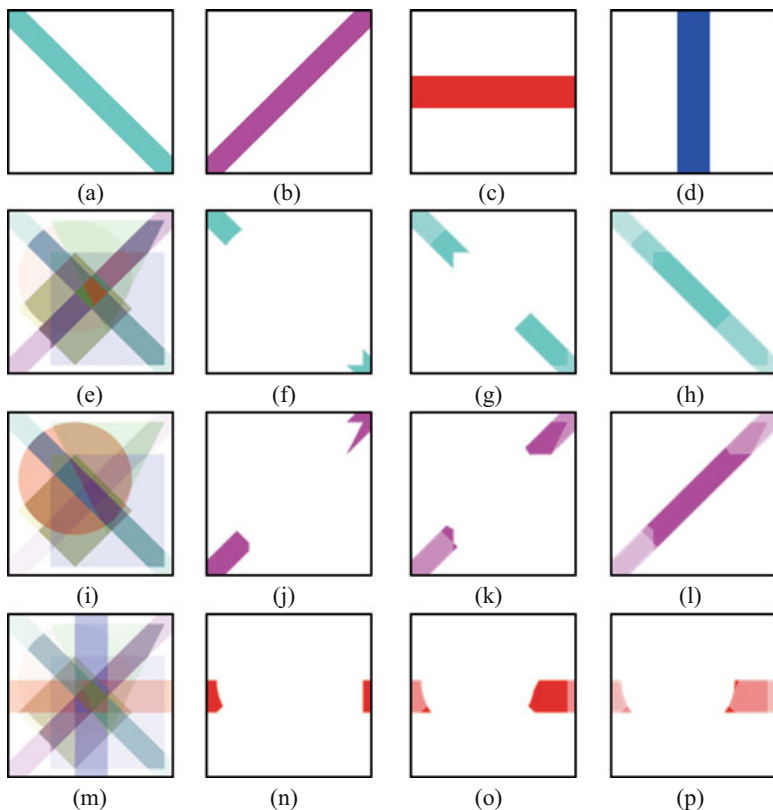


**Fig. 6** An image embedded in the analog memory can be recovered based on the quantity of marked pheromone. **(a)**  $\Omega(4)$ . **(b)**  $n_{\text{color}} = 1$ . **(c)**  $n_{\text{color}} = 1$  and 2. **(d)** all  $n_{\text{color}} \geq 1$ . **(e)**  $\Omega(4)$ . **(f)**  $n_{\text{color}} = 1$ . **(g)**  $n_{\text{color}} = 1$  and 2. **(h)** all  $n_{\text{color}} \geq 1$ . **(i)**  $\Omega(4)$ . **(j)**  $n_{\text{color}} = 1$ . **(k)**  $n_{\text{color}} = 1$  and 2. **(l)** all  $n_{\text{color}} \geq 1$ . **(m)**  $\Omega(4)$ . **(n)**  $n_{\text{color}} = 1$ . **(o)**  $n_{\text{color}} = 1$  and 2. **(p)** all  $n_{\text{color}} \geq 1$

## 4 Numerical Simulation

In this section, the algorithms of storing many images (dimension  $500 \times 500$ ) in an analog memory and of recovering an image from this memory are testing with the images shown in Fig. 8. These images now have the shape of an animal instead of a geometric shape used in the previous section. They represent an increase in the complexity of the data. For each shape of an animal, six images are generated for the six basic colors in Eq. (6). Images are selected randomly for testing the algorithms of storing and recovering images.

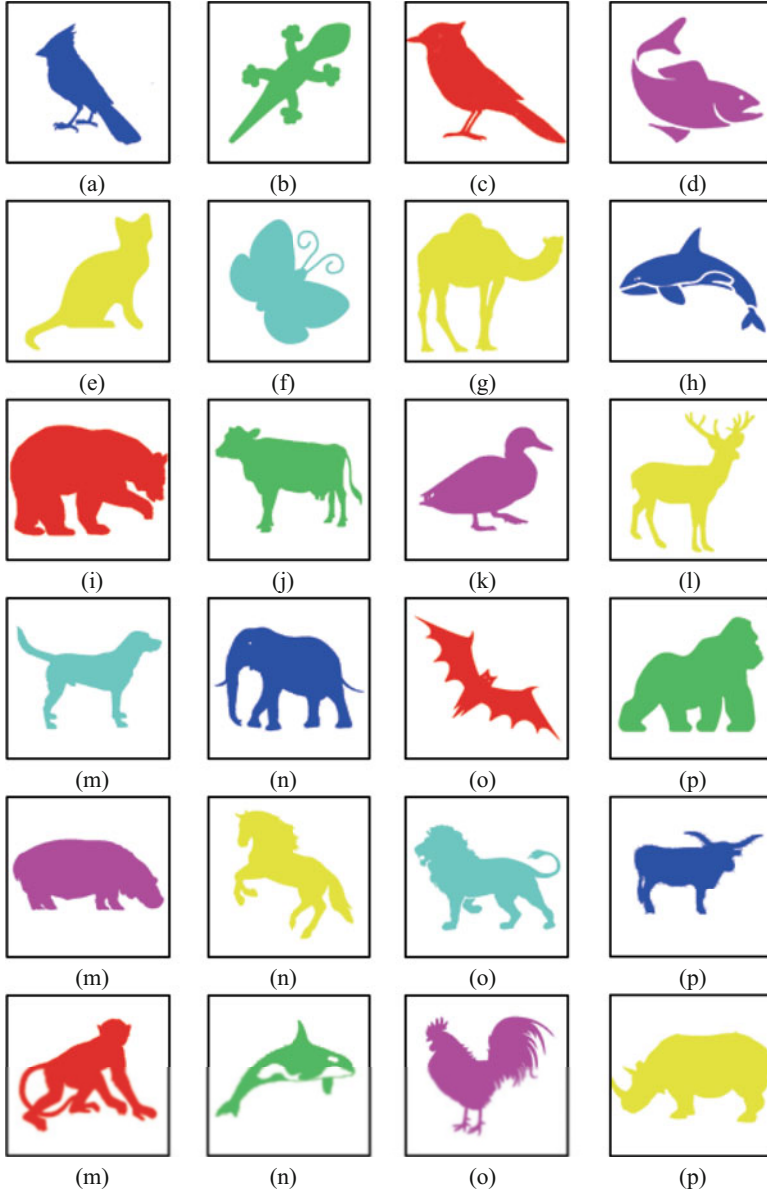
Figure 9 shows the results of the first test where six images of six different colors are selected randomly and stored in the analog memory. Each row of Fig. 9 represents an experiment, with the first image representing the constructed memory, the second image representing the pixels recovered from the region of the memory



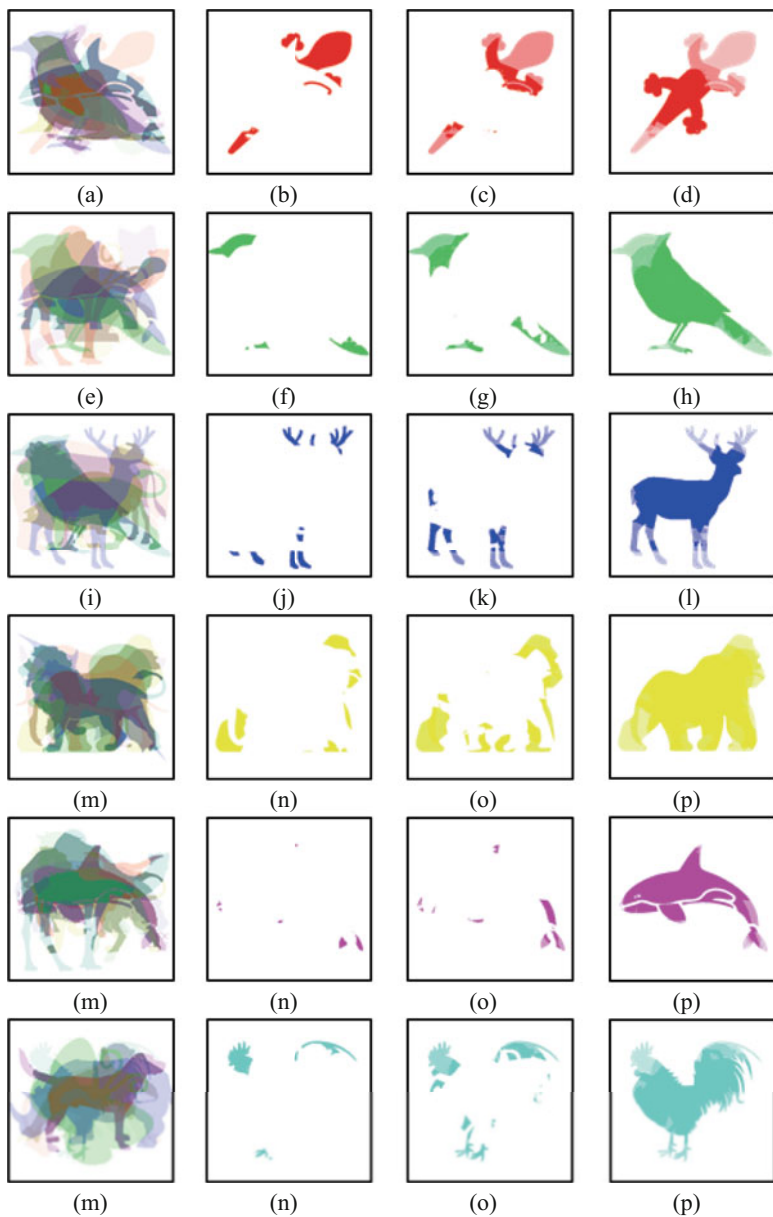
**Fig. 7** An image embedded in the analog memory can be recovered based on the quantity of marked pheromone. **(a)** cyan strip. **(b)** magenta strip. **(c)** red strip. **(d)** blue strip. **(e)**  $\Omega(6)$ . **(f)**  $n_{\text{color}} = 1$ . **(g)**  $n_{\text{color}} = 1$  and 2. **(h)** all  $n_{\text{color}} \geq 1$ . **(i)**  $\Omega(6)$ . **(j)**  $n_{\text{color}} = 1$ . **(k)**  $n_{\text{color}} = 1$  and 2. **(l)** all  $n_{\text{color}} \geq 1$ . **(m)**  $\Omega(8)$ . **(n)**  $n_{\text{color}} = 1$ . **(o)**  $n_{\text{color}} = 1$  and 2. **(p)** all  $n_{\text{color}} \geq 1$

that contains only one color, the third image representing the pixels recovered from the region of the memory that contains two colors (the pixels of the second image is superimposed in the lighter shade of its color), and the fourth image representing the pixels recovered in the region of the memory that contains three or more colors (the pixels of the third image on are superimposed in the lighter shade of its color). The fourth image in each row shows the complete recovery of an image based on a given color. The complete recovery can be visually verified for each result by comparing it with the original shape in Fig. 8. It is important to note that an image can be stored in the memory many times, and the recovery is still complete.

Figure 10 shows the results of the second test where ten images are selected randomly, with six images containing six different basic colors, and another four images containing four colors already used in the first six images, and these ten images are stored in the analog memory in the order of selection. Similar to

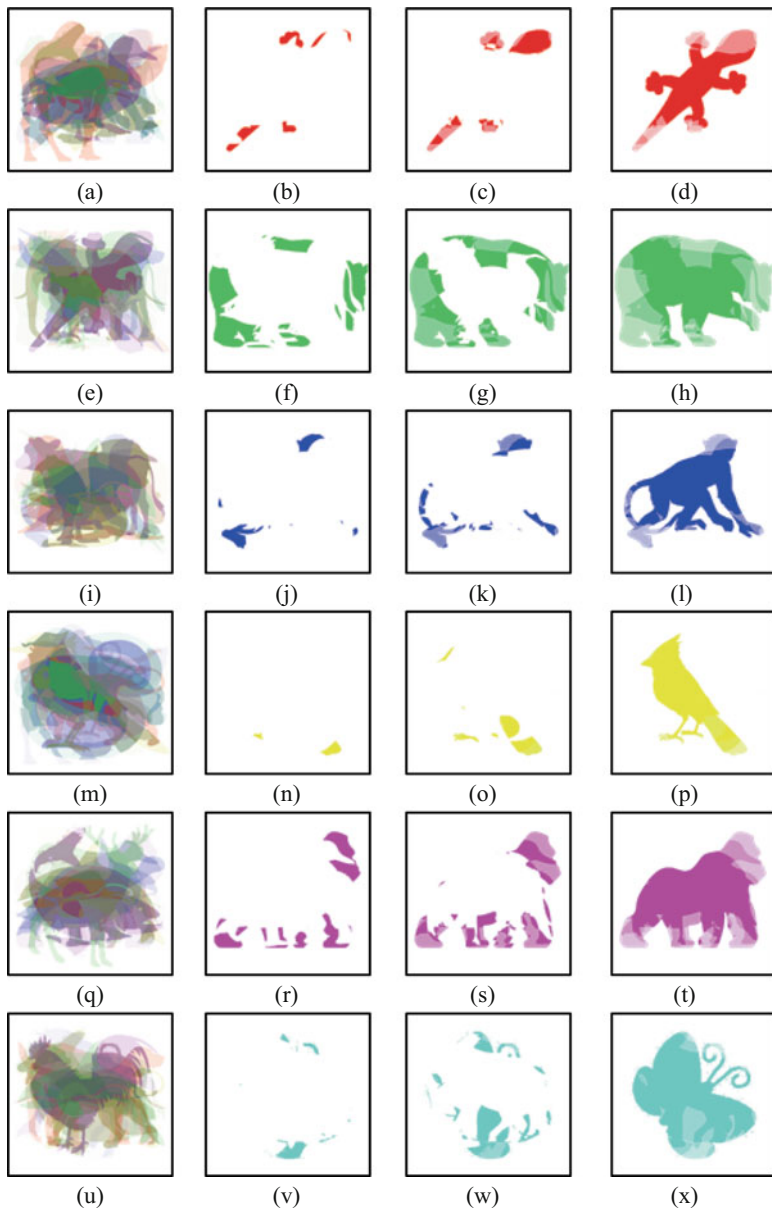


**Fig. 8** Images used for the numerical simulation: each of an animal is generated by one of the six basic colors. **(a)**  $\Omega(4)$ . **(b)**  $n_{\text{color}} = 1$ . **(c)**  $n_{\text{color}} = 1$  and 2. **(d)** all  $n_{\text{color}} \geq 1$ . **(e)**  $\Omega(4)$ . **(f)**  $n_{\text{color}} = 1$ . **(g)**  $n_{\text{color}} = 1$  and 2. **(h)** all  $n_{\text{color}} \geq 1$ . **(i)**  $\Omega(4)$ . **(j)**  $n_{\text{color}} = 1$ . **(k)** duck. **(l)** all  $n_{\text{color}} \geq 1$ . **(m)**  $\Omega(4)$ . **(n)**  $n_{\text{color}} = 1$ . **(o)**  $n_{\text{color}} = 1$  and 2. **(p)** all  $n_{\text{color}} \geq 1$

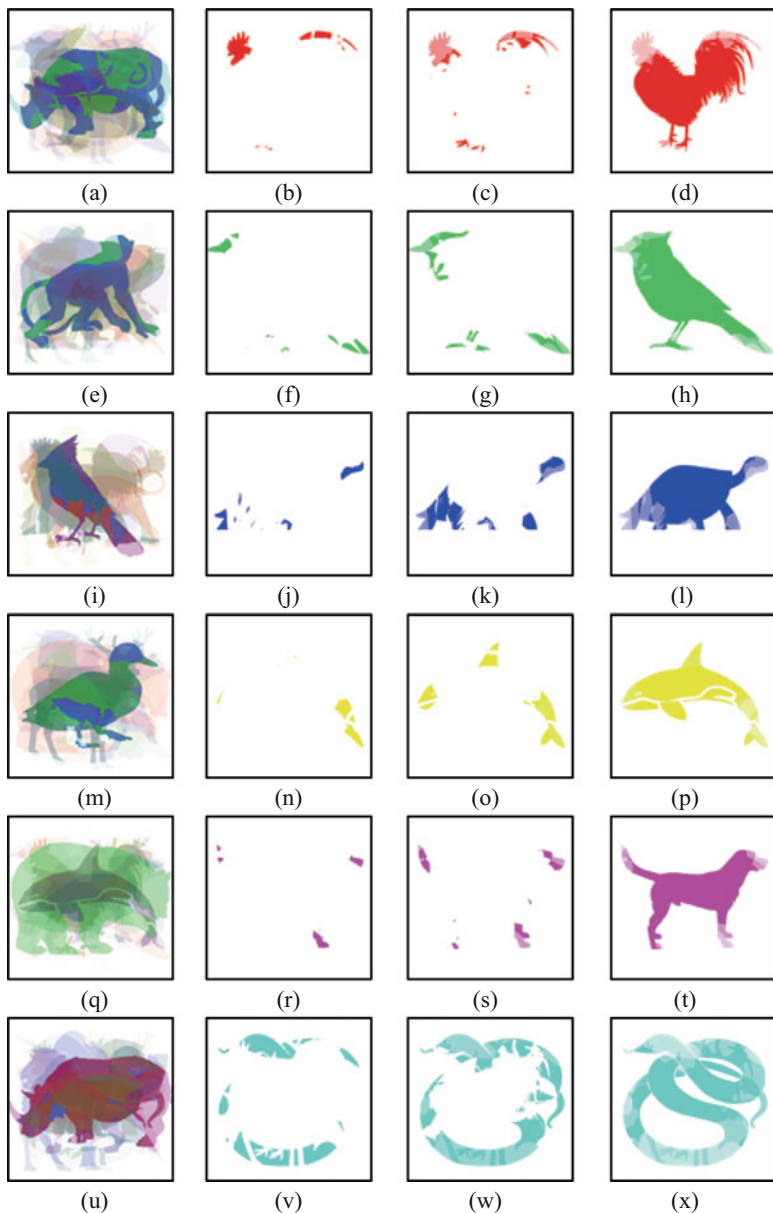


**Fig. 9** Recovery of an image incrustated in the memory containing six images based on a given color. **(a)**  $\Omega(6)$ . **(b)**  $n_{\text{color}} = 1$ . **(c)**  $n_{\text{color}} = 1$  and 2. **(d)** all  $n_{\text{color}} \geq 1$ . **(e)**  $\Omega(6)$ . **(f)**  $n_{\text{color}} = 1$ . **(g)**  $n_{\text{color}} = 1$  and 2. **(h)** all  $n_{\text{color}} \geq 1$ . **(i)**  $\Omega(6)$ . **(j)**  $n_{\text{color}} = 1$ . **(k)** duck. **(l)** all  $n_{\text{color}} \geq 1$ . **(m)**  $\Omega(6)$ . **(n)**  $n_{\text{color}} = 1$ . **(o)**  $n_{\text{color}} = 1$  and 2. **(p)** all  $n_{\text{color}} \geq 1$





**Fig. 10** Recovery of an image incrustated in the memory containing six images based on a given color. (a)  $\Omega(6)$ . (b)  $n_{\text{color}} = 1$ . (c)  $n_{\text{color}} = 1$  and 2. (d) all  $n_{\text{color}} \geq 1$ . (e)  $\Omega(1)$ . (f)  $n_{\text{color}} = 1$ . (g)  $n_{\text{color}} = 1$  and 2. (h) all  $n_{\text{color}} \geq 1$ . (i)  $\Omega(10)$ . (j)  $n_{\text{color}} = 1$ . (k) duck. (l) all  $n_{\text{color}} \geq 1$ . (m)  $\Omega(10)$ . (n)  $n_{\text{color}} = 1$ . (o)  $n_{\text{color}} = 1$  and 2. (p) all  $n_{\text{color}} \geq 1$ . (q)  $\Omega(10)$ . (r)  $n_{\text{color}} = 1$ . (s)  $n_{\text{color}} = 1$  and 2. (t) all  $n_{\text{color}} \geq 1$ . (u)  $\Omega(10)$ . (v)  $n_{\text{color}} = 1$ . (w)  $n_{\text{color}} = 1$  and 2. (x) all  $n_{\text{color}} \geq 1$



**Fig. 11** Recovery of an image incrustrated in the memory containing six images based on a given color. (a)  $\Omega(14)$ . (b)  $n_{\text{color}} = 1$ . (c)  $n_{\text{color}} = 1$  and 2. (d) all  $n_{\text{color}} \geq 1$ . (e)  $\Omega(14)$ . (f)  $n_{\text{color}} = 1$ . (g)  $n_{\text{color}} = 1$  and 2. (h) all  $n_{\text{color}} \geq 1$ . (i)  $\Omega(14)$ . (j)  $n_{\text{color}} = 1$ . (k) duck. (l) all  $n_{\text{color}} \geq 1$ . (m)  $\Omega(14)$ . (n)  $n_{\text{color}} = 1$ . (o)  $n_{\text{color}} = 1$  and 2. (p) all  $n_{\text{color}} \geq 1$ . (q)  $\Omega(14)$ . (r)  $n_{\text{color}} = 1$ . (s)  $n_{\text{color}} = 1$  and 2. (t) all  $n_{\text{color}} \geq 1$ . (u)  $\Omega(14)$ . (v)  $n_{\text{color}} = 1$ . (w)  $n_{\text{color}} = 1$  and 2. (x) all  $n_{\text{color}} \geq 1$

Fig. 9, Similar to Fig. 9, each row represents an experiment, with the first image representing the constructed memory, the second image representing the pixels recovered from the region of the memory that contains only one color, the third image representing the pixels recovered from the region of the memory that contains two colors (the pixels of the second image is superimposed in the lighter shade of its color), and the fourth image representing the pixels recovered in the region of the memory that contains three or more colors (the pixels of the third image on are superimposed in the lighter shade of its color). The fourth image in each row shows the complete recovery of an image based on a given color. Again, the complete recovery can be visually verified for each result by comparing it with the original shape in Fig. 8.

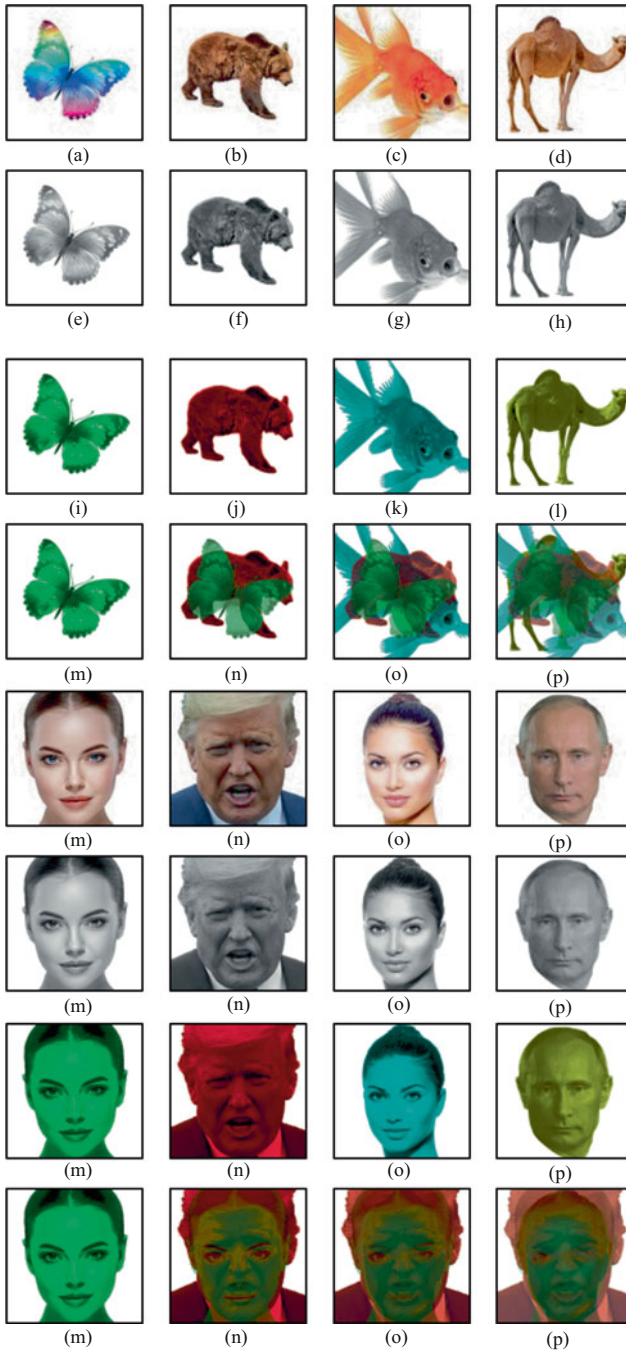
Figure 11 shows the results of the third test where the second test is done again with a change in which an image is stored repeatedly in various instances. This action changes the quantity of pheromone marked in each image stored in the analog memory. Similar to Fig. 9, each row represents an experiment, with the first image representing the constructed memory, the second image representing the pixels recovered from the region of the memory that contains only one color, the third image representing the pixels recovered from the region of the memory that contains two colors (the pixels of the second image is superimposed in the lighter shade of its color), and the fourth image representing the pixels recovered in the region of the memory that contains three or more colors (the pixels of the third image on are superimposed in the lighter shade of its color). The fourth image in each row shows the complete recovery of an image based on a given color. Again, the complete recovery can be visually verified for each result by comparing it with the original shape in Fig. 8.

More simulation will be provided in the supplemental materials.

## 5 Discussion of Future Work

In this project, it was demonstrated that unicolor images can be stored in an analog memory in the superimposed manner in a theoretical quantum computer and each image can be completely recovered when given a partial description of its color. This model of memory can contain an unlimited number of images. However, unicolor images are not common in realistic applications and therefore the concept presented in this paper must be extended for the capacity of storing more complex images.

In the next phase of the project, monochrome images that are a little more complex than unicolor images are considered. Figure 12 shows the scheme of this work in the immediate future where monochrome images can be created from color images in two steps: a color image can be converted into a black and white image by averaging the three red, green, and blue components and assigning this average as the gray level intensity, and the black and white image is converted into a monochrome image by using the gray level intensity as either red, green, or blue intensity. The second row of Fig. 12 shows the conversion of color images



**Fig. 12** Recovery of an image incusted in the memory containing six images based on a given color. **(a)** butterfly. **(b)** bear. **(c)** fish. **(d)** camel. **(e)** black and white. **(f)** black and white. **(g)** black and white. **(h)** black and white. **(i)** monochrome. **(j)** monochrome. **(k)** monochrome. **(l)** monochrome. **(m)**  $\Omega(1)$ . **(n)**  $\Omega(2)$ . **(o)**  $\Omega(3)$ . **(p)**  $\Omega(4)$ . **(m)** celebrity. **(n)** president. **(o)** celebrity. **(p)** president. **(m)** black and white. **(n)** black and white. **(o)** black and white. **(p)** black and white. **(m)** monochrome. **(n)** monochrome. **(o)** monochrome. **(p)** monochrome. **(m)**  $\Omega(6)$ . **(n)**  $\Omega(2)$ . **(o)**  $\Omega(3)$ . **(p)**  $\Omega(4)$

in the first row into black and white images. The third row of Fig. 12 shows the conversion of black and white images into monochrome images of the basic colors used in Eq. (6).

The analog memory originally contains three bits for each pixel in which each pixel represents the quantities of pheromone marked in an image. This configuration implies that the color is at the highest intensity. In the extension for storing monochrome images, the intensities must be explicitly stored in the memory. Therefore, instead of three bits for a pixel, it is proposed to use six bits for a pixel: the first three bits are used to contain the quantities of pheromone marked in an image, and the final three bits are used to contain the intensity of the primary colors. With this convention, the memory can be created with an algorithm of storing images that is similar to the algorithm of mixing colors used in this paper, described in the supplemental materials, for visualizing the analog memory. The preliminary result of this approach is shown in the fourth row of Fig. 12. The final four rows of Fig. 12 are the repetition of the first four rows with images of different context.

The future work then concentrates on the development of the algorithm to recover an image based on the partial description of color of a monochrome image. After the recovery of a monochrome image, the result can be converted back to the black and white image. However, a convention must be developed so that the conversion of a black and white image into a color image in its original entity must be developed in the final phase of the project so that color images can be stored in the analog memory in the manner that they can be recovered.

## 6 Conclusion

This project begins as a simulation to visualize human memory in a psychology study investigating the factors affecting the rate of retention. However, the simulation indicates a potential for quantum computing in which a model for analog memory with a definite size can be created for storing an unlimited number of images. This paper shows the results of the first stage of the project in which unicolor images can be stored in the analog memory in the superimposed manner, and each image can be recovered based on its partial description. Algorithms were derived in details based on the use of pheromone commonly found in the ant colony optimization. Numerical simulations are included for the purposes of demonstrating the concept and illustrating the workability and applicability of the analog memory.

**Acknowledgments** This study was supported by the Associative Research Program (Programa de Investigación Asociativa) in Cognitive Science financed by the University of Talca, Talca, Chile for the duration of 2019–2021.

## A.1 Supplemental Materials

More materials relevant to the topic will be available in the Web site <http://www.italca.cl/iee/Pham/MemoriaAnalogica>.

## References

- D. A. Patterson, & J. L. Hennessy. *Computer Organization and Design: The Hardware Software Interface*. Cambridge, MA, USA: Morgan Kaufmann, 2016.
- L. Null. *Essentials of Computer Organization and Architecture*. Burlington, MA, USA: Jones & Bartlett, 2018.
- K. Kokosa. *Pro .NET Memory Management: For Better Code, Performance, and Scalability*. New York, NY, USA: Apress, 2018.
- B. R. Hall, & K. J. Slonka. *Assembly Programming and Computer Architecture for Software Engineers*. Burlington, VT, USA: Prospect Press, 2018.
- J. D. Milton. *Essentials of Working Memory Assessment and Intervention*. Hoboken, NJ, USA: Wiley & Sons, Inc, 2015.
- A. Baddeley. *Essentials of Human Memory*. New York, NY, USA: Psychology Press Routledge, 2013.
- B. J. LaMeres. *Introduction to Logic Circuits & Logic Design with Verilog*. Cham, Switzerland: Springer Nature Switzerland, 2019.
- D. Harris, & S. Harris. *Digital Design and Computer Architecture*. Cambridge, MA, USA: Morgan Kaufmann, 2012.
- A. Agresti. *An Introduction to Categorical Data Analysis*. Hoboken, NJ, USA: Wiley & Sons, Inc, 2018.
- D. Andriess. *Practical Binary Analysis: Build Your Own Linux Tools for Binary Instrumentation, Analysis, and Disassembly*. San Francisco, CA, USA: No Starch Press, 2018.
- W. Liu, P. Gao, Y. Wang, W. Yu, & M. Zhang. "A Unitary Weights Based One-Iteration Quantum Perceptron Algorithm for Non-Ideal Training Sets," *IEEE Access*, no. 7, pp 36854–36865, 2019.
- J. D. Hidary. *Quantum Computing: An Applied Approach*. New York, NY, USA: Springer, 2019.
- J. W. B. Else, V. A. VanAst, & M. Kindt. "Human memory reconsolidation: A guiding framework and critical review of the evidence," *Psychological Bulletin*, vol. 144, no. 8, pp 797–848, 2018.
- M. M. Kahana, E. V. Aggarwal, & T. D. Phan. "The variability puzzle in human memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*," vol 44, no. 12, pp 1857–1863, 2018.
- J. Lv, X. Wang, & M. Huang. "Ant Colony Optimization-Inspired ICN Routing with Content Concentration and Similarity Relation," *IEEE Communications Letters*, vol. 21, no. 6, pp 1313–1316, 2017.
- H. Yang, J. Qi, Y. Miao, H. Sun, & J. Li. "A New Robot Navigation Algorithm Based on a Double-Layer Ant Algorithm and Trajectory Optimization," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 11, pp 8557–8566, 2019.
- C. Hibbs, S. Jewett, & M. Sullivan. *The Art of Lean Software Development: A Practical and Incremental Approach*. Sebastopol, CA, USA: O'Reilly Media, Inc., 2009.
- A. Singh, & P. J. Kaur. "A simulation model for incremental software development life cycle model," *International Journal of Advanced Research in Computer Science*, vol. 8, no. 7, pp 126–132, 2017.
- R. N. Carney, & J. R. Levin. "Do Mnemonic Memories Fade as Time Goes By? Here's Looking Anew!" *Contemporary Educational Psychology*, vol. 23, no. 3, pp 276–297, 1998.

- A. Wingfield, & D. L. Byrnes. *The Psychology of Human Memory*. New York, NY, USA: The Academic Press, 2013.
- D. P. Bliss, J. J. Sun, & M. D'Esposito. "Serial dependence is absent at the time of perception but increases in visual working memory," *Scientific Reports*, no. 7, 14739, 2017.
- G. A. Radvansky. *Human Memory*. New York, NY, USA: Routledge, 2017.
- E. A. Murray, S. P. Wise, M. K. L. Baldwin, & K. S. Graham. *The Evolutionary Road to Human Memory*. Oxford, UK: Oxford University Press, 2020.
- S. Patnaik, X. S. Yang, & K. Nakamatsu. *Nature-Inspired Computing and Optimization: Theory and Applications*. New York, NY, USA: Springer, 2017.
- C. Solnon. *Ant Colony Optimization and Constraint Programming*. Hoboken, NJ, USA: John Wiley & Sons, 2013.



**Trung T. Pham** (M'01–SM'06) obtained his BSEE, MS, and PhD from Rice University in Houston, Texas, and his MBA from the University of Houston, also in Houston, Texas.

He is with the United States Air Force Academy in Colorado Springs, Colorado, USA. He was also with the University of Talca in Talca, Maule Region, Chile. His research interest includes artificial intelligence, data mining, image processing, system modeling, and computational algorithms.

Dr. Pham is a Senior Member of the International Society of Automation (ISA). Previously he served as Vice-Chair of the Galveston Bay Section of the IEEE, and currently he serves as member of the Executive Committee of the Pikes Peak Section of the IEEE.

# Review on Social Laser Theory and Its Applications



**Andrei Khrennikov**

**Abstract** This is a review on social laser theory completed with its new developments and applications. An important methodological step toward similarity with quantum physics is the invention and consistent operation with infons. These are excitations of the quantum social-information field carrying social energy and coarse-grained content of communications (their color and quasi-color). We study in more detail interactions of infons with social atoms, the processes of absorption and emission (spontaneous and stimulated). We also analyze the dynamics of iterations of the cascades of infons in the social resonators. The latter are based on social networks coupled to laser's gain medium composed of social atoms. Consideration of the pro-war and pro-peace beams leads to the general discussion on the competing beams of social radiation and the conditions for their creation and coexistence. The role of social networks in lasing is illustrated by the protests during the COVID-19 pandemic. It is highlighted that a human gain medium can approach the state of population inversion with the supply of infons of one sort (quasi-color), but the stimulated emission can be induced by injection into the gain medium of infons of a different quasi-color. We call this behavior of social atoms memorylessness. This theoretical property is illustrated with the examples from the modern social-political life.

**Keywords** Social laser · Social energy · Social atom · Indistinguishability · Quantum statistics · Quantum information theory · COVID-19 protests · Pro-war and pro-peace beaming

---

A. Khrennikov (✉)

International Center for Mathematical Modeling in Physics and Cognitive Sciences, Linnaeus University, Växjö, Sweden

e-mail: [andrei.khrennikov@lnu.se](mailto:andrei.khrennikov@lnu.se)



# 1 Introduction

Nowadays the formalism and the methodology of quantum theory are widely used in applications outside of physics, especially in cognition, psychology, and decision-making as well as in social and political sciences (see, e.g., monographs Khrennikov 2004, 2010; Busemeyer and Bruza 2012; Asano et al. 2015; Haven et al. 2017; Bagarello 2019). Such applications are known as *quantum-like*—to distinguish them from real quantum physics, including its applications to cognitive science (quantum physical reductionism). The majority of the quantum-like models are based on quantum mechanics (QM), but for the social science one has to appeal to the formalism of the quantum field theory (QFT). One of the most intensively developed QFT-theories is the social laser theory (Khrennikov 2015, 2016, 2020a,b; Khrennikov et al. 2018, 2019; Alodjants et al. 2022). This chapter is a review on this theory completed with its new developments and applications.

The basic entities of this theory are social energy, atoms, and fields (Thims 2008; Khrennikov 2015, 2016, 2020a,b; Khrennikov et al. 2018, 2019; Alodjants et al. 2022) (cf. also with James 1890; Freud 1957; Jung 2001; Jung and Pauli 2014). Social atoms represent humans. They exchange quanta of social energy with the social-information field which is composed of excitations carried by communications massively emitted by mass media and social networks. The lasing scheme can be formulated with these entities as the process of social energy pumping in a human gain medium and then stimulated emission of a cascade of social actions. The latter are understood very generally as actions in both physical and social-information spaces: mass protests, color revolutions, wars, and collective decisions on the important societal problems.

This chapter is an important methodological step toward approaching similarity with quantum physics. We invented the social-information analogs of photons which are called *infons*. These are excitations of the quantum social-information field carrying social energy and coarse-grained content of communications (their color and quasi-color). We study in more detail interactions of infons with social atoms, the processes of absorption and emission (spontaneous and stimulated).

We also analyze the dynamics of iterations of the cascades of infons in the social resonators. The latter are based on social networks coupled to laser's gain medium composed of social atoms. The special attention is paid to the role of Echo Chambers. They increase color and quasi-color coherence of the social-information field.

The role of social networks as lasing resonators is illustrated by the massive protests during COVID-19 pandemic (cf., e.g., van der Zwet et al. 2022).

Consideration of the pro-war and pro-peace beams generated since February 2022 in mass media and Internet resources leads to the general discussion on competing beams of social radiation and conditions for their creation and coexistence. The role of social networks is highlighted once again (cf. Khrennikov 2020b).

Social laser theory predicts that a human gain medium can approach the state of population inversion with infons of one sort (quasi-color), but the stimulated

emission can be done by injection of a batch of infons of a different quasi-color *memorylessness of social atoms*. This is a very important property of social laser which can be widely used in social engineering. We illustrate this theoretical property with a few examples from the modern social-political life.

In this chapter we shall widely use the abbreviation *s-* for “social,” say *s*-atom and *s*-energy.

## 2 Social Atom

A human is the minimal indivisible entity of society, a social atom (*s*-atom). The atomic viewpoint on the human being has a very long history; see Thims’ book (Thims 2008), in this book the reader can find discussions and references on the basic human-atomistic (or molecular) models.

Although the authors of such models suggested a different definition, generally they follow the same paradigm: operating with human beings as individual information processors described by just a few parameters characterizing information interaction. Thus, practically infinite complexity of a human being was reduced to these basic parameters, in the simplest case to social energy. This reduction of complexity made humans treatable thermodynamically. On the other hand, ignoring human complexity diminishes the explanatory power of such models; typically, they can describe statistical behavior of humans but not explain why they behave in one or another way.

The distinguished property of our approach is the quantum-like treatment of variables, as representing observations performed on *s*-atoms. Another distinguished property is the invention of the quantum information field, i.e., *s*-atoms can interact not only with each other, as in aforementioned theories, but also with the social-information field which is also interpreted and modeled in the quantum-like framework. Such modeling is supported by the recent development of the information approach to quantum theory.

## 3 Social Energy

From the very beginning of QM, Bohr denied the objectivity of quantum variables, such as position, momentum, or energy. They cannot be treated as properties of systems and assigned to them before measurement. Measurements’ outcomes are generated in the process of complex interaction between a system and a measurement device (Bohr 1987).

This approach is fruitful for the introduction of *s*-energy. We do not need to create a deep neurophysiological or psycho-social theory to justify this notion (cf. James 1890; Freud 1957; Jung 2001; Jung and Pauli 2014). *S*-energy is an observable measuring the degree of social excitement of a person. It can be done with a

variety of measurement devices. They can be calibrated with different scales, and the simplest scale is dichotomous,  $E = E_{a0}, E_{a1}$ , where these values are assigned to relaxation and excitement, respectively.

The simplest measurement procedure is done with question: “Do you feel your socially excited or not?” From the quantum operational viewpoint, such invention of the  $s$ -energy observable seems to be justified. In the future quantum-like modeling, the crucial role will be played not by the absolute values of the energy levels, but by their difference:

$$E_a = E_{a1} - E_{a0}. \quad (1)$$

If both levels are high, but the energy gap is small, then such  $s$ -atom would not be able to perform a strong social action. Say, she would never participate in demonstrations leading to brutal clashes with police.

The energy levels determine the corresponding mental states of an  $s$ -atom which are denoted as  $|E_{a0}\rangle, |E_{a1}\rangle$ . The main feature of quantum representation of states is the existence of superpositions, e.g.,  $s$ -atom can be not only in the mental states  $|E_{a0}\rangle, |E_{a1}\rangle$ , corresponding to the concrete values of  $s$ -energy, but also in superposition states of the form:

$$|\psi\rangle = c_{a0}|E_{a0}\rangle + c_{a1}|E_{a1}\rangle, \text{ where } |c_{a0}|^2 + |c_{a1}|^2 = 1, c_{aj} \in \mathbf{C}. \quad (2)$$

The complex coefficients  $c_{aj}$ ,  $j = 0, 1$ , encode the probabilities, and  $p_j = P(E = E_j|\psi) = |c_j|^2$  is the probability that  $s$ -atom in the mental state  $|\psi\rangle$  would answer that her  $s$ -energy equals  $E_j$ . (Here we consider the introspective measurement procedure of  $s$ -energy when  $s$ -atom is asked to report her energetic feeling and the set of answers is restricted to “I feel me relaxed” and “I feel me excited.”)

This probability depends on the state  $|\psi\rangle$  of  $s$ -atom, and this fact is reflected in the symbol  $P(E = E_j|\psi)$ . This formula is Born’s rule, the basic quantum rule providing the probabilistic interpretation for the linear algebra on the state space  $\mathcal{H}$  of  $s$ -atom. The latter is a complex Hilbert space. In this simple case it is two-dimensional with the orthonormal basis  $|E_{a0}\rangle, |E_{a1}\rangle$  (qubit space). As is typical in physics, the scalar product of two vectors from  $\mathcal{H}$  is denoted as  $\langle\psi_1|\psi_2\rangle$ . In terms of the scalar product of  $s$ -atom’s states, the Born rule is written as

$$p_j = |\langle E_{aj}|\psi\rangle|^2. \quad (3)$$

The existence of superposition states is the mathematical expression of non-objectivity of  $s$ -energy. Until  $s$ -atom is not asked to estimate her  $s$ -energy, she does not know its value. Of course, one can design other methodologies for the measurement of  $s$ -energy which are not based on self-observations.

## 4 Social-Information Field

In accordance with QFT, a field represents an ensemble of its energetic excitations. Mathematically this excitation structure of a field is described in Fock space. Quantum fields are described with the operators of creation and annihilation of excitations.

Quantum field excitations are treated on equal grounds with “real systems” such as atoms or electrons. Say excitations of the electromagnetic fields are photons. Moreover, excitations corresponding to vibrations, e.g., of atoms in a crystal or dipoles in a molecular, also treated as systems, phonons.

In social studies we proceed in the same way. The social-information quantum field is an ensemble of energetic excitations, and each excitation is determined by the portion, “quantum,” of  $s$ -energy. The field excitations are generated by the sources of socially relevant information, mainly by mass media and social networks. Each communication emitted by them carries a quantum of  $s$ -energy. We call such quanta *infons*.

In physics photon’s energy can be connected with light’s frequency and hence the color. In the same way we can color infons, depending on  $s$ -energy: low and high  $s$ -energy infons are colored as red and violet, respectively, and for intermediate coloring we can use other colors; say yellow infons are sufficiently energetic, but still not exciting. For example, during the pandemic the majority of communications on COVID-19 were of the violet color; during the spring of year 2022 news about the war was also violet, but the communications about COVID-19 were colored in red. News about sexual affairs of politicians and stars can be colored in yellow.

We hope that this  $s$ -energy/color terminology will not be misleading. In ordinary life red means danger and attracts more attention than violet. But, in physics red photons are low energetic and violet photons are highly energetic. We keep the physical picture. So, the red colored infons carry small amounts of  $s$ -energy and violet infons are highly energetic.

In fact, in physics the characterization of QFT excitations is not reduced to energy. For example, a photon also has polarization. Generally photon’s state  $|E\alpha\rangle$  is characterized by the parameters  $E$  = energy and  $\alpha$  = (polarization, temporal and spatial extensions). Social-information field can also have some characteristics additional to  $s$ -energy and related to communication’s content. We call such characteristics the *quasi-color* of infon; its state can be encoded as  $|E\alpha\rangle$ , where  $E$  and  $\alpha$  are  $s$ -energy and the quasi-color, respectively.

Introduction of the quasi-color is a delicate process related to such foundational issue as *indistinguishability of quantum systems* (see, e.g., Ballentine 1998). Quantum theory assumes that two photons in the state  $|E\alpha\rangle$  are indistinguishable. Moreover, it is claimed that there are no hidden variables and additional photon’s characteristics which would provide a possibility to distinguish two photons in the state  $|E\alpha\rangle$ . So, in quantum physics indistinguishability has the fundamental character. Indistinguishability plays the crucial role in derivation of quantum

statistics for energy distribution in the framework of *statistical thermodynamics* (Schrödinger 1989).

In quantum-like modeling of the social-information ( $s$ -information) field, infons are indistinguishable, up to  $s$ -energy  $E$  and the quasi-color  $\alpha$ . This is the important assumption beyond social laser theory. However, there is one important difference between quantum and quantum-like indistinguishabilities. The former is genuine and irreducible and the latter is relative to context. In one context some social variables are important and they should be included in the quasi-color  $\alpha$ , and in another context they do not play any role, so they are not included in  $\alpha$ . But, we cannot deny their existence. For example, humans have names, but their names do not play any role in the process of social lasing in the form of mass protests. So, humans are indistinguishable w.r.t. to the name variable. The slogan “Black Lives Matter” is the integral quasi-color which was crucial in the protests in USA. The concrete names of black people who experienced racism, discrimination, and racial inequality were hidden in this quasi-color.

So, indistinguishability in social laser theory is quantum-like. It is up to the characteristics determining the process of lasing. These characteristics form the quasi-color  $\alpha$ . And infons’ indistinguishability is up to  $s$ -energy and this quasi-color.

## 5 Absorption and Emission of Infons by Social $s$ -Atom

Here we consider processes of absorption and emission of quanta of  $s$ -energy by  $s$ -atoms interacting with the excitations of the social-information field—infons.

Consider physical atoms with two levels of energy, excited and relaxed,  $E_1$  and  $E_0$ . The difference between these levels

$$\Delta E_a = E_{a1} - E_{a0} \quad (4)$$

is the basic energetic parameter of an atom, its spectral line.<sup>1</sup>

A two-level atom reacts only to photons carrying energy  $E$  matching with atom’s spectral line (Bohr’s rule):

$$\Delta E_a = E. \quad (5)$$

In quantum-like modeling we apply Bohr’s rule to  $s$ -atoms and infons. So, a two level  $s$ -atom reacts only to infons carrying energy  $E$  matching atom’s spectral line, see (4) and (5).

---

<sup>1</sup> In the case of two-level atom, it has just one spectral line. Generally there are a few spectral lines corresponding to differences between energy levels,  $\Delta E_{a;ij} = E_{ai} - E_{aj}$ ,  $i > j$ . This is the atom’s spectrum.

If infon carries too high energy which is larger than the spectral line,  $E > \Delta E_a$ , then this  $s$ -atom would not be able to absorb this infon. For example, infon carrying  $s$ -energy  $E$  is a call for upraise against the government. And  $s$ -atom is a bank clerk (say Elena) in Moscow. Elena has the liberal views and hates Putin's regime, but her spectral line is too small to absorb  $s$ -energy carried by such infon and to move from the ground state to the excited state. She simply ignores such highly energetic communication, news, or Internet post. Similarly, if infon's  $s$ -energy  $E$  is less than spectral line  $\Delta E_a$ , then Elena would not be excited by such infon.

As a physical atom cannot collect energy from a few low energy photons, with  $E < \Delta E_a$ ,  $s$ -atom cannot collect  $s$ -energy from a few infons carrying small portions of  $s$ -energy.  $S$ -atom either absorbs infon (if their colors match each other) or does not react to it. In the same way,  $s$ -atom cannot "eat" just a portion of  $s$ -energy carried by highly energetic infon with  $E > \Delta E_a$ .

In the quantum-like theory the process of infon emission by excited  $s$ -atom is also characterized by its spectral lines. In the case of two-level  $s$ -atom, this is just the number  $\Delta E_a$ .  $S$ -atom can emit only infon satisfying (5).

As in physics, emission can be spontaneous when  $s$ -atom suddenly emits infon—at random instance of time and with a random quasi-color. Another sort of emission is known as stimulated;  $s$ -atom emits infon as the result of interaction with infons of surrounding social-information field. Ideally a single infon in the state  $|\epsilon\alpha\rangle$ , where  $E = \Delta E_a$ , stimulates excited  $s$ -atom to emit infon in precisely the same state. So, emitted infon has not only the same  $s$ -energy as stimulating infon but also the same quasi-color  $\alpha$ . However, since this process is probabilistic (as all quantum-like processes), the real stimulation of emission is possible only with fields of high density. So,  $s$ -atom should interact with a strong social-information field, with a cloud of excitations that have the same  $s$ -energy (equal to  $\Delta E_a$ ) and quasi-color. Inside such a field the probability of emission is high.

In quantum physics spontaneous emission of photon by atom is considered as exhibition of irreducible quantum randomness. However, such picture might be adequate only for completely isolated atom. But real atom is never completely isolated. Background radiation is everywhere. Therefore, it may be that even the spontaneous emission is not a totally random process. It can be stimulated by fluctuations of the background electromagnetic field and interactions with other atoms. In the same way spontaneous emission of  $s$ -excitation might be generated by fluctuations in surrounding social environment: occasional news, a scandal with a partner, a problem at work, and so on. The quasi-color which emitting  $s$ -atom assigns to infon (or action in physical space generated by this infon) may reflect the environment's quasi-color.

In physics the photon absorption-emission condition (5) is satisfied only approximately

$$E \approx \Delta E_a. \quad (6)$$

The spectral line broadening is always present. In an ensemble of atoms,  $\Delta E_a = \Delta E_a(\omega)$  is the Gaussian random variable. This is a bell centered at the mean average

value  $\Delta\bar{E}_a$ . The dispersion of the Gaussian distribution depends on an ensemble of atoms. Ensembles with small dispersion are better as gain mediums for physical lasing, but deviations from exact law (5) are possible.

It is natural to assume Gaussian distribution realization of exact laws even for social systems, in particular, absorption of excitations of the information field by  $s$ -atoms. Thus, deviations from (5) are possible. But a good human gain medium (an ensemble of  $s$ -atoms selected for social lasing) should be energetically homogeneous. Therefore, the corresponding Gaussian distribution should have very small dispersion. The latter is also an important necessary condition for functioning of physical laser.

Finally, we discuss one interesting feature of interrelation of absorption and emission: Consider quantum physics. Suppose that an atom absorbed a photon with momentum vector  $\vec{p}$ . This vector determines the direction of photon's propagation and its length  $|\vec{p}|$  determines the photon's energy. The process of absorption is characterized by matching of energies (5), so the direction of photons propagation given by

$$\vec{\alpha} \equiv \vec{p}/|\vec{p}| \quad (7)$$

does not play any role in the process of absorption. The most interesting for us is that atom "forgets" the direction  $\vec{\alpha}$  of incoming photon. In the process of spontaneous emission, an atom emits a photon in an arbitrary direction. Moreover, in the process of stimulated emission, an atom emits a photon with momentum which is identical to momentum of stimulating photons, the stimulating electromagnetic field.

The same "memory washing" is a feature of quantum-like model since its mathematical formalism is identical to quantum physical theory. So,  $s$ -atom does not remember the quasi-color  $\alpha$  of infon, say a news, which it has absorbed. It can emit spontaneously infon, say a post in a social network, of an arbitrary quasi-color. When  $s$ -atom is stimulated for emission, it emits infon of the same quasi-color as stimulating infons, say news. This memorylessness of social atoms is very important in social engineering, including social lasing.

Turning to quantum physics, we note, in contrast to direction *alpha* (7), photon's polarization.

## 6 Social vs. Physical Lasing Schematically

For simplicity we consider two-level atoms, both physical and social. We shall present the social lasing scheme parallelly to the scheme of physical lasing.

## 6.1 Laser's Components and the Stages of Lasing

Physical laser has three main components:

- A gain medium composed of atoms
- A source of energy (the electromagnetic field)
- A resonator (an optical cavity)

Physical lasing has two main stages:

- (A) *Energy pumping*. Energy is pumped into a gain medium; the aim is to approach the *population inversion*—more than 50% of atoms should be transferred into the excited state.
- (B) *Stimulated emission*. A batch of photons propagating in the same direction  $\alpha$  given by (7) are injected into the gain medium. They stimulate the cascade process of the emission of photons by atoms.

At both the stages the colors of photons and atoms' spectral line match each other at least approximately. We recall that we consider two-level atoms and there is just one spectral line. The gain medium should be color-homogeneous. Photons produced during the B-stage copy the direction of propagation from stimulating photons. The latter were injected along the main axis of the optical cavity—laser's resonator. As was pointed out, the directions of photons' propagation at the A and B stages can be totally different (memorylessness of atoms). Some important details will be mentioned below to illustrate the corresponding details of social lasing.

The social laser also has three components:

- A gain medium composed of  $s$ -atoms (humans)
- A source of  $s$ -energy (the social-information field)
- A social resonator (Internet-based social networks)

Social lasing also has two main stages:

- (A) *Energy pumping*.  $S$ -energy is pumped into a human gain medium; the aim is to approach the *population inversion*—more than 50% of  $s$ -atoms should be transferred into the excited state.
- (B) *Stimulated emission*. A batch of infons of the same quasi-color  $\alpha$  are injected into the gain medium. They stimulate the cascade process of the emission of infons by  $s$ -atoms.

Now we describe these stages in more detail:

The mass media and Internet pump  $s$ -energy into a gain medium composed of  $s$ -atoms to approach the *population inversion*—to transfer the majority of atoms to the excited state. The gain medium should be homogeneous w.r.t. its spectral structure, ideally  $\Delta E_a = \text{Const}$ . In reality  $\Delta E_a$  is a Gaussian random variable with very small standard deviation. The  $s$ -energies of infons (communications, news, messages, Internet videos) used for energy pumping need not be so sharply concentrated around average  $\overline{\Delta E_a}$  of  $\Delta E_a$ .  $S$ -atoms would simply ignore infons



essentially deviating from  $\overline{\Delta E_a}$ . And such  $s$ -energy losses are compensated by the powerful flows of information generated by modern mass media and Internet.

After achievement of the population inversion, the stimulated emission is started. A batch of infons (say communications, news) is injected into the human gain medium. The first constraint is that  $s$ -energy of these stimulating infons should match the spectral line of  $s$ -atoms, see (5) (in the ideal case). In reality it is sufficient to control the approximate matching condition (6). So,  $s$ -energy of injected infons should not deviate essentially from  $\overline{\Delta E_a}$ . Another constraint on injected infons is that they all should carry the same quasi-color  $\alpha$ , say  $\alpha = \text{COVID-19}$ , or  $\alpha = \text{vaccination}$ , or  $\alpha = \text{Russian aggression against Ukraine}$  (depending on socio-political context and aims of social lasing). This injected beam of information radiation generates the cascade process in the human gain medium.

Quasi-color homogeneity of the stimulating information injection is the basis of quasi-color coherence of the laser beam of infons. Later this social-information beam is transferred into the social action matching infons' color ( $s$ -energy amplitude) and quasi-color (information content). Infons' homogeneity should be very high. Here statistical deviations are not acceptable, since infons of other quasi-colors would also generate their own cascades. Such "noise-cascades" would destroy quasi-color coherence of the output beam of social radiation. They should be then eliminated with the aid of social resonators.

## 6.2 *How Does the Cascade Process Evolve? The Role of Laser's Resonator*

In the simplified picture, each infon stimulates  $s$ -atom to emit infon having the same color and quasi-color with its stimulator. Resulting two infons stimulate two  $s$ -atoms to emit two new infons, so one stimulating infon resulted in four infons which interact with four  $s$ -atoms and so on. After say 20 steps there are  $2^{20}$ , approximately one million of infons (the excitations of the social-information field) of the same color and quasi-color. In reality, the process is probabilistic:  $s$ -atom reacts to stimulating infon only with some probability. The latter rapidly increases with the increase of the density of the social-information field. And the field's bosonic nature is crucial.

In physics the beam induced in the gain medium by the stimulating injection of coherently colored photons is not the laser's output beam. Laser has an additional component which plays the crucial role in increasing both the amplitude and coherence of the output beam. This is a *laser resonator*. For lasers emitting photons—excitations of the quantum electromagnetic field, this is an optical cavity. Its mirrors reflect beams generated inside the gain medium and send them back to this medium. In this way the cascade process in the gain medium is repeated many times.

The process of reflection from the mirrors also increases *spatial coherence* of the beam. The photons propagating not precisely along the main axis of the cavity are reflected outside of the cavity and disappear. We remark that the stimulating beam is sent along this axis. The cascade photons copy the direction of propagation in space given by momentum vector (7) of initially injected photons.

We remark that during the beam iterations (through reflections) energy is continued to be pumped into the gain medium from outside. So, atoms that emitted photons in the preceding iterations absorb newly incoming photons and move to the excited state. Intensity of pumping of energy quanta into the gain medium should be high enough, higher than some threshold depending on laser's parameters. This threshold is called the *lasing threshold*. If the intensity of energy pumping is lower than the lasing threshold, then too many atoms would spontaneously relax between two reflection-iterations of the basic wave of photons. On the one hand, this is the energy loss, and on the other hand, the mini-cascades created in the excited gain medium by spontaneous emissions would lower coherence of the radiation beam. If the intensity is higher than the lasing threshold, then practically all energy pumped into the gain medium is transferred into the basic radiation wave propagating along the cavity's axis.

In our quantum-like model the social laser also should have a resonator, a kind of two mirrors that reflect infons and send them back into the human gain medium—to interact again with  $s$ -atoms in the human gain medium and to stimulate them to emit infons. The role of such social resonators is played by Internet-based information systems, such as You Tube, Facebook, Instagram, Bastyon, Telegram, Life Journal, VK, and so on. The main distinguishing feature of these systems is the possibility of the rapid feedback to communications, news, and videos in the form of comments, comments on comments, and so on. The beam of infons created from the initial stimulating injection (typically by mass media's giants as say BBC and CNN, Washington Post, New York Times, and Guardian) is distributed over Internet channels and creates new posts (in the form of articles and videos), each of them is actively commented. Each comment plays the role of a mirror. But this is a kind of an active mirror, not only reflecting infons but also creating them.

The social resonator and also the physical resonator not only amplify the beam of social radiation inside the human gain medium but also increase its coherence w.r.t. the quasi-color  $\alpha$  of the stimulating injection. Posts quasi-colored differently from  $\alpha$  disappear in the massive flow  $\alpha$ -infons.

### 6.3 *The List of the Basic Counterparts of Social Laser Theory*

- Our quantum-like model is of the quantum field type, the social-information field. Its excitations are called infons. Each infon transports quantum of  $s$ -energy. The latter determines infons' color, red infons are low energetic, and violet infons are highly energetic.

- Each  $s$ -atom is characterized by the  $s$ -energy spectrum; in the simplest case of two levels, this is the difference between the energies of the excitation and relaxation states,  $\Delta E_a = E_1 - E_0$ .
- Beside  $s$ -energy (color), infons (the excitations of the information field) are characterized by other labels, quasi-colors, carrying content of information communications.
- Coherence corresponds to quasi-color sharpness; ideal social laser emits a single quasi-color mode, denoted say by the symbol  $\alpha$ .
- Excited  $s$ -atoms by interacting with  $\alpha$ -colored infons also emit  $\alpha$ -colored infons.
- The amount of  $s$ -energy carried by stimulating infons (communications) should match the color of  $s$ -atoms in the gain medium.
- To approach the population inversion,  $s$ -energy is pumped into the gain medium. Pumping should be intensive, since  $s$ -atoms have the tendency spontaneously relax and emit infons with randomly distributed quasi-colors.
- This energy pumping is driven by the mass media and the Internet sources.
- The gain medium should be homogeneous with respect to  $s$ -energy spectrum. Ideally (for the two-level case), all  $s$ -atoms should have the same color  $\Delta E_a$ . However, in reality, it is impossible to create such human gain medium. As in physics, the *spectral line broadening* has to be taken into account.
- The quasi-colors of infons in energy pumping have no direct connection with the quasi-color of infons generated by stimulating emission (memorylessness of  $s$ -atoms).
- Infons follow the Bose-Einstein statistics.
- This statistics matches with the bandwagon effect in psychology (Colman 2003) (see article Khrennikov 2020b for details).
- The probability of emission of the  $\alpha$ -colored infon by  $s$ -atom in a human gain medium increases very quickly with the increase of the intensity of the social-information field on the  $\alpha$ -colored mode.
- The stimulating injection of homogeneously quasi-colored infons gives rise to the cascade of coherent (w.r.t. the color and quasi-color) infons.
- The created beam of social radiation is amplified in the social resonators based on Internet information systems, say YouTube, Facebook.
- The social resonators, especially in the form of Internet-based Echo Chambers, also improve quasi-color coherence (Sect. 7).
- When the power of the beam of coherent infons becomes very high, infons are transformed into social actions, either in physical or in information spaces.

For example, a gain medium consisting of humans in the excited state and stimulated by the anti-corruption quasi-colored information field would “radiate” a wave of anti-corruption protests. The same gain medium stimulated by an information field carrying another quasi-color would generate the wave of actions corresponding this last quasi-color. For social laser engineering, it is very important that the quasi-colors of  $s$ -energy supply and stimulation of emission do not need to coincide. Population inversion can be approached with, say the quasi-color  $\alpha$ , and then the stimulated emission can be generated with another quasi-color  $\beta$ .

## 7 Echo Chamber as Reinforcer of Social Coherence

The detailed presentation of social resonators theory can be found in article (Khrennikov 2020b). In the latter we highlighted the differences between the physical resonators of the cavity type, so to say “passive reflectors,” and the social resonators which are based on the “social mirrors. Such mirrors can be treated as active reflectors producing on demand of users new infons.

Here we shall consider in more detail special but at the same time very important type of social resonators, namely, Internet-based *Echo Chambers* (see also Khrennikov 2020b). In our notations it can be defined as follows:

Echo Chamber is a system in that some beams of infons carrying (as their quasi-colors) news, communications, ideas, and behavioral patterns are amplified and sharpened through their feedback propagation inside this system. In parallel to such amplification, infons carrying quasi-colors different from those determined by the concrete Echo Chamber are suppressed.

In our terms, an *Echo Chamber is a device for transmission and active re-emission (not simply reflection) of infons—the excitations of the quantum social-information field.* Its main purpose is amplification of this field and increasing its quasi-color coherence via distilling from “social noise,” i.e., infons colored and quasi-colored differently from Echo Chamber’s basic color and quasi-color.

We underline that an Echo Chamber is considered as a component of the social laser, as its resonator. The coherent output of an Echo Chamber, the quasi-color of this output, is determined not only by the internal characteristics of the Echo Chamber but also by the quasi-color of stimulated emission in laser. The same Echo Chamber may be turned in accordance with the aim of the stimulated emission in progress. Of course, such turning is not possible for every Echo Chamber. Some of them are stable w.r.t. to their basic quasi-colors.

Amplification and increasing of coherence w.r.t. to Chamber’s quasi-color have already been discussed for general social resonators. What about sharpening? Generally  $s$ -atom’s quasi-color is a vector  $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_n)$ , where the coordinates represent (as labels) different information contents; each  $\alpha_j$  is valued in some space  $X_j$ , often  $X_j = \{-1, 1\}$  represents the no/yes answers, but more complex quasi-color spaces are also possible, and the simplest such space is  $X_j = \{-1, 0, 1\}$ , negative, neutral, and positive evaluation of some issue, say the present war in Ukraine.

Let us consider functioning of some Internet-based Echo Chamber; for example, one that is based on some social group in Facebook and composed of  $s$ -atoms. The degree of their indistinguishability can vary depending on the concrete Echo Chamber. Say, names are still present in *Facebook*, but they have some meaning only for the restricted circle of friends; in *Instagram* or *Snapchat*, even names disappear and  $s$ -atoms operate just with nicknames.

By a social group we understand some sub-network of say Facebook, for example, social group “Quantum Physics.” The main feature of a social group is that all posts and comments are visible for all members of this social group. Thus,

if I put the post “Getting rid of nonlocality from quantum physics,” then it would be visible for all members of this social group, and they would be able to put their own comments or posts related to my initiation post. This is simplification of the general structure of posting in Facebook, with constraints that are set by clustering into “friends” and “followers.”

We assume that the ensemble of  $s$ -atoms of this Echo Chamber approached population inversion, so the majority of them are already excited. A batch of communications of the same quasi-color  $\alpha$  and carrying quanta of  $s$ -energy  $E_c = \Delta E_a$  is injected in the Echo Chamber. Excited  $s$ -atoms interact with the stimulating communications and, with some probability, emit information excitations of the same quasi-color as the injected stimulators. These emitted quanta of  $s$ -energy are represented in the form of new posts in Echo Chamber’s social group. Each post plays the role of a mirror, and it reflects the information excitation that has generated this post.

However, the analogy with the optics may be misleading. In classical optics, each light ray is reflected by the mirror again as one ray. In quantum optics, each photon reflected by the mirror is again just one photon. An ideal mirror reflects all photons, and the real one absorbs some of them.

In contrast, “the mirror of an Echo Chamber,” the information mirror, is  $s$ -energy *multiplier*. A physical analog of such a multiplier works as follows. Each light ray is reflected as a batch of rays or in the quantum picture, matching the situation better, each photon by interacting with such a mirror generates a batch of photons. Of course, the usual physical mirror cannot reflect more photons than the number of incoming ones, due to the energy conservation law. Hence, the discussed device is hypothetical.

## 8 Illustrating Examples

In this section we would like to illustrate previous theoretical considerations by the additional examples of social laser’s use at the modern socio-political arena, potential, and real uses.

### 8.1 COVID-19 Protests

During the COVID-19 pandemic the human gain medium was overheated by the shock news about the spread of this terrible disease, by its deadly consequences, by life during lockdowns, and by numerous rigid restrictions on social life (e.g., masks in public places and somewhere, e.g., OAE, even at the streets), by the QR-codes and obligatory vaccination for some professions, e.g., the personal of hospitals. Such communications were often repeated a few times, and their content could vary, but not the basic quasi-color,  $\alpha = \text{COVID-19}$ . In some countries, even the

most democratic ones as in Sweden, the laws were changed by restricting the basic freedoms, including the basic constitutional right for meetings and demonstrations.

Scientists also actively contributed in generation COVID-19 fear. For example, some mathematical models of disease spread predicted millions of deaths from COVID-19 in UK and hundreds of thousands in Sweden, if the rigid restrictions, including lockdowns and masks, would not be invented (Ferguson et al. 2020).

My personal opinion is that such the mathematical models were really primitive, basically the very old SIR-dynamics (may be disturbed by a stochastic term for noise). This is a good place to mention the new model of disease spread (Khrennikov and Oleschko 2020); it took into account the social cluster structure of population. This model predicted the opposite effect, comparing with the majority of models, of lockdowns and other rigid restrictions. In contrast to, e.g., Ferguson et al. (2020), such restrictions slowdown approaching of natural immunity in human population.

By summarizing we can say that at the end of the year 2020 and the beginning of the year 2021 the state of population inversion was approached in European countries, Australia, Canada, USA, and Russia. The human gain medium was ready for radiating a huge spike of social energy, a spike which could destroy the basics of society. Various social groups and individuals started to generate information excitations against WHO's COVID-19 policy and their governments following this policy. Social networks resonated these excitations. This led to the generation of local spikes of protests, worldwide and especially in Australia, Germany, the Netherlands, France, UK, Canada, and even Sweden (see, e.g., van der Zwet et al. 2022; Chueca and Teodoro 2029). In Sweden the COVID-19 restrictions were really mild compared with the majority of countries: no lockdowns and no masks. However, extended suppression of functioning of the social resonators, especially by YouTube and Twitter, prevented creation of the global wave of protests. At the same time, the local spikes relaxed some portions of social energy, and in this way social temperature was lowered.

## ***8.2 Pro-war and Pro-peace Beaming: Competitions of Stimulating Emissions***

Coming back to social goodness lasing theme, consider a war between two countries or blocks of countries. And suppose that some group of policy makers wants to use the social laser technology to generate the wave of peaceful thoughts and actions. Assume that this group is powerful enough to generate a strong injection of communications for peace. In principle, it is possible to connect with a message having quasi-color  $\alpha =$  "peace" big amount of social energy. Unfortunately, to generate such social lasing for peace, the war should be going on for sufficiently long time. And it should lead to big casualties from both sides or at least from one of them. Otherwise even spontaneously emitted hate cascade would destroy the processes of stimulation of a cascade of thoughts and actions for peace.

Here we come to the problem of *competing stimulation* in human gain medium approached the state of population inversion.

The main problem in starting such a process is that another group at the political arena might not be interested to end this war. By using their information resources, they could continue social lasing in favor of the war,  $\alpha = \text{“war.”}$  And the aggression instinct is so powerful that there is a big chance that such an  $\alpha = \text{“war”}$  beam of social laser would be essentially stronger than the  $\alpha = \text{“peace”}$  beam.

In such a competition of social energy beams of two quasi-colors, the crucial role is played by social resonators, in the form of social networks based on You Tube, Facebook, Live Journal, Bastion, Twitter, Telegram, Instagram, Yandex, Vkontakte, and so on. Therefore, it is so important to control such information resources (e.g., one can understand the motivation of Elon Musk to buy Twitter). Without the control of social resonators, it is practically impossible to start stimulating emission. The initial (stimulating) batch of information excitations which is not supported by social resonators would pass through information space in a flash and disappear.

### ***8.3 Russian-Ukrainian War and Relaxation of Social Energy Generated by COVID-19 Pandemic***

Now we turn again to the COVID-19 pandemic. As was pointed out in Sect. 8.1, during the years 2020–2021 human society collected a lot of social energy and approached the state of population inversion.

Of course, the state of population inversion was not approached in whole world; say in Egypt and other African countries COVID-19 did not lead to massive transition of people into the excited state. We speak about this transition in European countries (both West and East Europe), USA, Canada, and Australia. It is interesting that in China, in spite of very high degree of COVID-19 related restrictions, generally the mental state of population could not be characterized as excited. Chinese population took these restrictions rather calmly by following automatically to COVID-19 state recommendations, as people here would do in any other case. On the other hand, the mass protests of Canadian truck-drivers demonstrated that the degree of social tensions in Canadian society was very high. These protests can be considered as a test of COVID-19 generated instability in Western society.

One can speculate that only the war between Russia and Ukraine relaxed the huge amount of social energy collected during the pandemic in Europe, USA, Canada, Russia, and Ukraine. The COVID-19 energy was transferred into the war energy. Here we discussed mainly the processes in social information space, i.e., not the real war battles in physical space.

## 8.4 *Generation of Financial Tsunamis: Reddit Against Wall Street*

For those who did not know or forgot the story about a social network uprising against Wall Street, we recall some details by following (Malik 2021):

GameStop is a US video game retailer that has lost much of its market share to online trade and whose stock plummeted from \$56 a share in 2013 to about \$5 in 2019. Some big hedge funds decided that they would cash in on GameStop’s misery by shorting its shares. A short is a bet that an asset, such as a share, will decline in price. It’s a manoeuvre that can generate huge profits. But if the asset price doesn’t fall, investors can also lose a lot of money.

A bunch of Reddit geeks on the online forum r/wallstreetbets, an investment discussion group that boasts more than 6 million users, decided to buy GameStop shares en masse. Perhaps they saw it as an investment, perhaps they were bored, perhaps they wanted to inflict pain on Wall Street. Whatever the reason, the consequence was to push GameStop’s share price up. And up. Once it became a global story, others piled in too, boosting the share price from about 40 to almost 400 in a matter of days. As a result, big investors lost big... . The story, however, is not just about traders getting their comeuppance, but also about the absurdity of the stock market.

To analyze this event, we shall appeal to social laser theory with its application to social atoms operating at the financial market. In this framework this “global story” demonstrated not only the absurdity of the stock market but rather the possibility to use new financial technology for generation of short squeezing. And as usually, this GameStop short squeezing generated huge profits for those who designed and ignited the process of stimulated amplification of coherent social actions. In this case “social actions” were posting comments at Reddit expressing believes (hopes, instructions) that GameStopp shares will go up in price. These were actions in the information space. They led to actions in the financial space—buying of GameStopp shares.

We finalize this short consideration of Reddit “uprising” by a few citations from media sources (Sherr 2021):

And though the share price dipped on Monday, Feb. 1, by more than 30%, many Reddit users say they’re buying more GameStop stock, convinced it’ll rocket even higher. Jaime Rogozinski, the apparent founder of the Reddit community at the heart of all this, told The Wall Street Journal it’s like ‘a train wreck happening in real time.’ Keith Gill, the trader in the Reddit community who helped kick off the battle, told the paper he ‘didn’t expect this.’

There might be something cathartic in watching the wolves of Wall Street themselves being savaged, but we should not romanticise the Reddit geeks. This was not an ‘uprising’ or ‘the French Revolution of finance’, as Donald Trump’s former communications director Anthony Scaramucci absurdly described it, but a scheme to play professional investors at their own game. [Guardian]



## 9 Concluding Remarks

We hope that this review will be useful for the researchers working in both humanitarian and natural sciences. As was mentioned, the methodology of social laser theory was enriched through the invention of *infon*. This is an analog of photon. Photons are the excitations of the quantum electromagnetic field and infons are the excitations of the *quantum social-information field*. By operating with infons the presentation of spontaneous and stimulated emission and absorption became very similar with its counterpart of the quantum physics based on operation with photons. From my viewpoint, infons are not “less real” than photons or phonons (quanta of vibrations). The same can be said about the social-information field by comparing it with the electromagnetic vibration fields.

We emphasize the bosonic nature of these fields which is the basic factor leading to generation of the cascade process of the stimulated emission in the lasers’ gain media, both physical and human. The “bosonicity” is a consequence of indistinguishability of excitations, photons, phonons, and infons. For the latter, indistinguishability is not absolute. Distinguishability is only up to a few characteristics involved in lasing, namely, infon’s color (*s*-energy) and quasi-color (strongly coarse-grained information content—a content label). In physics it is commonly claimed that there are no hidden variables giving deeper description of system’s state than the quantum state—completeness of quantum mechanics. Knowing of the hidden variables would destroy indistinguishability. For cognitive and social systems, hidden variables definitely exist, each human has say the passport, and humans can be distinguished by observation passport content. The use of quantum theory in the presence of the hidden variables is a complex foundation issue. It was discussed in a few of my previous publications, e.g., Khrennikov (2010). I cannot say that this issue was completely clarified. At least the contradiction with the violation of the Bell inequalities can be resolved by referring to the contextual character of the mental hidden variables.

The infons-language is convenient to describe the dynamics of cascades’ iterations within social laser—the gain medium and resonator. Social resonators are implemented by coupling laser’s gain medium to social networks. We emphasize the role of resonators in, e.g., lasing for the competing candidates in the presidential elections. Generally we are interested in the process of creation of two competing beams of social actions, as the mentioned elections or war and peace beams in contemporary information space. Also in article (Khrennikov 2020b), we highlight the role of Internet-based Echo Chambers in increase of the amplitude as well as color and quasi-color coherence of the beams of social radiation. Echo Chambers are also used to increase temporal coherence, to make the spike of social radiation sharply concentrated in the time domain. It is a good place to point out that a social Internet-based resonator is a kind of active mirror, in contrast to the optical cavities with the reflecting mirrors.

Once again (cf. with Khrennikov 2015, 2020b) we highlighted the possibility to supply *s*-energy to the gain medium with infons of the quasi-color different from

the quasi-color of infons in the stimulating injection (and hence the output beam of social radiation). This property of  $s$ -atoms is called *memorylessness*. It plays the important role in *social engineering*. The real aim of social lasing can be deemed at least at the stage of  $s$ -energy pumping in the gain medium. Moreover,  $s$ -energy produced by one social laser can be used at the stage of approaching population inversion in another social laser. In turn the latter can be used for new social laser and so on.

## References

- Khrennikov, A. (2004). *Information dynamics in cognitive, psychological, social, and anomalous phenomena*, Ser.: Fundamental Theories of Physics, Kluwer, Dordrecht.
- Khrennikov, A. (2010). *Ubiquitous quantum structure: from psychology to finances*; Springer: Berlin-Heidelberg-New York.
- Busemeyer, J. R. and Bruza, P. D. (2012). *Quantum models of cognition and decision*, Cambridge University Press, Cambridge.
- Asano, M., Khrennikov, A., Ohya, M., Tanaka, Y. and Yamato, I. (2015). *Quantum adaptivity in biology: from genetics to cognition*, Springer: Heidelberg-Berlin-New York.
- Haven, E., Khrennikov, A. and Robinson, T. R. (2017). *Quantum Methods in Social Science: A First Course*; WSP: Singapore.
- Bagarello, F. *Quantum Concepts in the Social, Ecological and Biological Sciences*; Cambridge: Cambridge Univ. Press, 2019.
- Khrennikov, A. (2015). Towards information lasers. *Entropy*, 17, N 10, 6969–6994.
- Khrennikov, A. (2016). Social laser: Action amplification by stimulated emission of social energy. *Phil. Trans. Royal Soc.*, 374, N 2054, 20150094.
- Khrennikov, A., Alodjants, A. Trofimova A. and Tsarev, D. (2018). On interpretational questions for quantum-Like modeling of social lasing. *Entropy*, 20(12), 921.
- Khrennikov A., Toffano, Z. and Dubois, F. (2019). Concept of information laser: from quantum theory to behavioural dynamics. *Eur. Phys. J.*, 227, N. 15–16, 2133–2153.
- Khrennikov, A. (2020a), Social Laser, Jenny Stanford Publ., Singapore.
- Khrennikov, A. (2020b), Social laser model for the Bandwagon effect: generation of coherent information waves, *Entropy*, 22(5), 559.
- Alodjants, A. P., Bazhenov, A. Y., Khrennikov, A. Y., and Bukhanovsky, A. V. (2022). Mean-field theory of social laser. *Scientific Reports*, 12(1), 1–17.
- L. Thims, *The Human Molecule*. Lulu Com. Publ. (2008).
- James, W. (1890). *The Principles of Psychology* (New York: Henry Holt and Co.), Reprinted 1983 (Boston: Harvard Univ. Press).
- Freud, S. (1957). *The Standard Edition of Complete Psychological Works of Sigmund Freud*, Edited and Translated by J. Strachey, Vols. I-XXIV (The Hogarth Press, London).
- Jung, C. G. (2001). *On the Nature of the Psyche*, Routledge Classics.
- Jung, C. G. and Pauli, W. (2014). *Atom and Archetype: The Pauli/Jung Letters 1932–1958*, Princeton Univ. Press: Princeton.
- van der Zwet, K., Barros, A.I., van Engers, T.M. et al. Emergence of protests during the COVID-19 pandemic: quantitative models to explore the contributions of societal conditions. *Humanit Soc Sci Commun* 9, 68 (2022).
- Chueca, E. G.; Teodoro, F. (2029) Pandemic and social protests: cities as flashpoints in the COVID-19 era. *CIDOB notes international*, N 266.
- K. Malik, An uprising against Wall Street? Hardly. GameStop was about the absurdity of the stock market. *The Guardian*, January 31, 2021

- I. Sherr, Reddit's battle with Wall Street over AMC, GameStop stock a 'Ponzi scheme,' can't last (2021). Cnet, <https://www.cnet.com/news/reddits-battle-with-wall-street-over-amc-gamestop-stock-a-ponzi-scheme-cant-last/>
- Bohr, N. (1987). *The Philosophical Writings of Niels Bohr*, 3 vols. (Ox Bow Press, Woodbridge, CT).
- Ballentine, L. E. (1998). *Quantum Mechanics: A Modern Development*; WSP: Singapore.
- Schrödinger, E. (1989). *Statistical thermodynamics*, Dover Publications.
- Colman, A. (2003). *Oxford Dictionary of Psychology*, Oxford University Press: New York, USA, p. 77.
- N. M. Ferguson et al., Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand (Imperial College London, 2020); 10.25561/77482.doi
- Khrennikov, A., and Oleschko, K. (2020). An ultrametric random walk model for disease spread taking into account social clustering of the population. *Entropy*, 22(9), 931.

# Challenges from Probabilistic Learning for Models of Brain and Behavior



Nicolás Marchant, Enrique Canessa, and Sergio E. Chaigneau

**Abstract** Probabilistic learning is a research program that aims to understand how animals and humans learn and adapt their behavior in situations where the pairing between cues and outcomes is not always completely reliable. This chapter provides an overview of the challenges of probabilistic learning for models of the brain and behavior. We discuss the historical background of probabilistic learning, its theoretical foundations, and its applications in various fields such as psychology, neuroscience, and artificial intelligence. We also review some key findings from experimental studies on probabilistic learning, including the role of feedback, attention, memory, and decision-making processes. Finally, we highlight some of the current debates and future directions in this field.

**Keywords** Probabilistic learning · Category learning · Feedback · Decision-making · Cognitive models

## 1 Introduction

For a very long time, behavioral scientists have been wondering how animals can learn and adapt their behavior to environmental demands. One big concern among some scientists was that the pairing between cues and outcomes is not always completely reliable. If we honor Darwin's hypothesis, animals would seek to adapt their behavior to the environmental demands, even in unreliable situations. Since the early days of behaviorism, this research program concerned with how animals

---

N. Marchant (✉) · S. E. Chaigneau  
Center for Social and Cognitive Neuroscience, School of Psychology, Universidad Adolfo Ibáñez, Santiago, Chile  
e-mail: [nicolas.marchant@edu.uai.cl](mailto:nicolas.marchant@edu.uai.cl); [sergio.chaigneau@uai.cl](mailto:sergio.chaigneau@uai.cl)

E. Canessa  
Faculty of Engineering and Science, Universidad Adolfo Ibáñez, Viña del Mar, Chile  
e-mail: [ecanessa@uai.cl](mailto:ecanessa@uai.cl)

can learn under unreliable outcomes conditions was known as probabilistic learning (Brunswik 1943; Edwards 1961; Castellán 1973).

To study probabilistic learning is quite simple. The experimenter has only to adjust some schedule routine in order to make the pairing between cues and outcomes unreliable. For example, one of the first conditioning experiments that relied on the use of probabilistic learning was the experiment of Rescorla (1968). He adjusted the pairing routines between a tone (conditioned stimulus; CS) and a shock (unconditioned stimulus; US) by implementing different conditions in which the US was probabilistically paired with the CS. He noted that the conditioning strength was higher whenever the US and the CS occurred deterministically (i.e., with 100% of matching accuracy), but strength continually declined as the pairing became more unreliable (e.g., 80% of accurate matchings; 60% of accurate matchings). Overall, probabilistic associations between US and CS impoverish the conditioned learning. However, simple conditioning experiments were limited in their capacity to explain complex human behaviors such as inference, reasoning, and categorization. So, in the cognitive turn (see Miller 2003 for a brief review), conditioning was abandoned and replaced by cognitive explanations of probabilistic learning (Estes 1976; Lindell 1976). The cognitive turn on probabilistic learning highlighted the relationship between learning and memory formation, and whether those learned associations can be recovered through explicit processing.

A way of formalizing how memory and learning occur in the cognitive system is by means of computational modeling. Since the beginning of the cognitive turn, it is common to find mathematical formulations that represent a certain cognitive function as a testable “algorithmic” hypothesis (see Wilson and Collins 2019). This means that experimental and computational modeling explanations of probabilistic learning work together. In the current work, we will be interested in those computational/mathematical explanations. However, our discussion will be kept at a general level, so that readers that are not familiar with this area of research can grasp the general problems of the field. For those readers interested in the gory details, we provide numerous references.

We recognize that there are different areas of cognition interested in probabilistic processing and how to model it, such as decision-making (Tversky and Kahneman 1974), reasoning (Oaksford and Chater 2007), or Bayesian learning (Tenenbaum et al. 2006). However, our primary focus here is on the process of learning to classify when receiving probabilistic feedback across a sequence of consecutive trials. By understanding human probabilistic learning, it should be possible to better develop artificial learning systems that face the same problems as humans do (e.g., such as in machine learning; Fréney and Verleysen 2014). Also, it should be possible to develop a better explanation of the environmental demands that people face in natural learning conditions (e.g., doctors learning to diagnose from a set of symptoms; Estes 1986). This chapter presents some of the most well-known challenges that research on probabilistic learning faces in the era of cognitive neuroscience.

## 2 Challenges to Probabilistic Learning Modeling

### 2.1 Error Correction

One of the first challenges faced by modelers and experimentalists in explaining probabilistic learning was whether people actually learn from their errors and what is the mechanism that people and animals use to correct their behavior. As we outlined in the introduction, discoveries from behaviorisms stated that animals learn because of the teaching signal, which can be either an aversive or appetitive stimuli. One of the first mathematical formulations of error correction was the Rescorla and Wagner (1972) learning rule. In short, this learning rule states that the associative strength of a cue, which is stored in memory, depends on its predictive value of the outcome (López and Shanks 2008). Subjects will assign specific associative strengths depending on the cue predictiveness, which is acquired through learning, and thus, a cue with a highly predictive value will gain more associative strength. A way in which this model is often discussed is by saying that learning will be enhanced if the outcome is surprising, i.e., it is not predicted by the currently active cues. This surprisingness is computed through an error term which represents the discrepancy between the expectation and the actual status of the outcome in the current trial. The Rescorla and Wagner model (1972) was of special interest because it addresses problems that earlier theories of conditioning seemed unable to explain. One of those problems is the blocking effect (Kamin 1969), which states that when two cues are presented within the same trial, subjects will only learn the cue with the most predictive value while blocking the cue with the lower predictive value. Also, the model is consistent with neuroanatomical evidence regarding how the dopamine system works in humans and animals. Research carried out by Schultz (1999) shows evidence that dopaminergic neurons in the macaque midbrain fire when the predictive cue is presented (which was previously paired with a response), predicting the same firing rate that will trigger the response. Since then, these results have been replicated many times, with human and nonhuman animals (Daw and Doya 2006). Because the model provides an account for experiments with different salient cues and also because of its biological plausibility, the Rescorla and Wagner model is included, in one way or another, in many learning models.

A probabilistic learning model that incorporates the Rescorla and Wagner rule was the configural cue model developed by Gluck and Bower (1988a, b), Gluck (1991). They implemented a connectionist model (see Thomas and McClelland 2008) to address whether people can learn two potential categories from a set of different cues. This kind of experimental procedure is known as Multiple Cue Probability Learning (MCPL; Edgell 1980; Estes 1986). Here, the categories (i.e., Disease R and Disease C) are the outcomes which provide probabilistic feedback to different combinations of four different cues (i.e., the symptoms). Across 250 trials (of different combinations of symptoms), there was a probability of 25% that

a specific trial will be correctly classified as Disease R (and a 75% of correctly be classified as Disease C). However, the first symptom (i.e., the bloody nose) had a 69% diagnosticity for Disease R (and the opposite for Disease C). The authors found that subjects learned that the first symptom was indeed more diagnostic of Disease R, independent of the base-rate of Disease R being low (i.e., 25%). Later, the authors modeled the behavioral results using their configural cue model. Using an error correction activation algorithm (the least mean squared rule (LMS), which as Sutton and Barto (1981) noted, is a special case of the Rescorla and Wagner rule), their model successfully predicted that the first symptom should become, in fact, more diagnostic. This occurs because different cues and the combination of those cues compete with each other to match the teaching signal, while the winning cue is the one that reduces the error between the response and the outcome (i.e., the classification).

However, the configural cue model was incapable of explaining all classification phenomena. Criticisms from Nosofsky and colleagues argued that the model was insufficient to account for rule-based classification (Nosofsky et al. 1994). This led to the emergence of other computational models which integrate an error term that were able to explain some rule-based problems. One such model is the ALCOVE (Kruschke 1992) model, which is an exemplar-based model embedded with the same LMS algorithm function. In brief, ALCOVE, much like other exemplar models, assumes that different stimuli are stored in memory as individual traces. Computationally, this idea is implemented as a neural net where a hidden layer has nodes representing each of the exemplars in the training set. However, ALCOVE still retains a learning mechanism which updates attentional resources by trial and error. Exemplar-based models were preferred, because they integrate a cognitive explanation of the attentional resources combined with an error-correction mechanism.

## 2.2 *Feedback Discounting*

Despite preferences in the literature for exemplar-based models of learning and categorization, they faced a second major challenge. A phenomenon known as feedback discounting (or error discounting) pushed the limits of these kinds of models. Feedback discounting occurs because people (and perhaps even nonhuman animals) will eventually accept a certain level of unavoidable error, and, continually, they will begin to discount feedback information slowing down their learning (Estes 1984; Kruschke and Johansen 1999; Craig et al. 2011).

Kruschke and Johansen (1999) developed an extension of the ALCOVE model which is able to capture whenever people stop using feedback information. They called this new model RASHNL (Rapid Attention SHifts 'N' Learning). This model incorporates a feedback discounting mechanism whenever attention is shifted away from irrelevant cues (those that produce error) to relevant cues (those that reduce error). Thus, in a trial-by-trial fashion, the RASHNL model will eventually reduce

the learning rate from which it updates the cue-outcome associations (note that the RASHNL is an extension of the ALCOVE model, using the same error-correction term to update learning). Furthermore, the feedback discounting phenomenon has received supporting evidence from electrophysiology studies (EEG). A study conducted by Sewell and colleagues showed that participants who discount feedback entirely eliminate an EEG component known as feedback-related negativity (FRN), while participants who do not discount feedback presented a standard FRN frequency (Sewell et al. 2018). The FRN component usually elicits a peak evoked signal between 200 and 300 ms after the presentation of the feedback, being generally larger for negative feedback rather than for positive feedback (Cohen et al. 2011). In a nutshell, probabilistic learning models should take feedback discounting into account. As noted by Craig et al. (2011), models of probabilistic learning tend to improve whenever they incorporate a feedback discounting mechanism. Moreover, participants who discounted feedback showed different brain signals often found in the middle region of the EEG scalp. Regardless of the previous evidence, it is still unclear what conditions cause participants to start feedback discounting, and whether this is an automatic process or an explicit conscious strategy.

### 2.3 Normative Responses

Researchers in the psychology of decision-making and behavioral economics were interested in knowing whether people behaved according to normative criteria when dealing with probabilistic information. If so, then it should be found that people behave close to normatively. From the decision-making literature, it has been suggested that maximizing is the normative response when dealing with uncertainties (Fiorina 1971; Shanks et al. 2002). In short, maximizing states that people will always place a certain item into the response that is most likely to belong to. For example, if we have an item  $s$ , in which 80% of the trials belong to keyboard response A, then, people should maximize their responses by always responding A whenever item  $s$  is presented. However, there is a debate whether people always behave according to maximizing, which brings us closer to our third challenge. For some researchers, people often deviate from maximizing, and instead, they rely on a suboptimal response strategy which is called probability matching (Castellan 1973; Friedman and Massaro 1998; Shanks et al. 2002). Probability matching states that people will progressively match their responses according to the outcome criteria. For example, if item  $s$  belongs to response A with an 80% chance, then, subjects will tend to respond 80% of the times that item  $s$  belongs to response A.

There is still a debate under which circumstances people will respect maximizing or will fall into probability matching. For example, Shanks et al. (2002) created different experimental situations in which probability matching would be undesirable. However, there were always people (albeit a small proportion) that relied on probability matching. Studies that were concerned about how people learned to categorize under unreliable situations showed that – on average – people follow a



probability matching pattern across learning trials (Little and Lewandowsky 2009a, b; Craig et al. 2011; Sewell et al. 2018). Obviously, this pattern of results places challenges for cognitive modeling, considering that maximizing and probability matching seem so different in terms of possible underlying cognitive mechanisms.

## 2.4 Cognitive Processing

So far, we have reviewed three challenges that are critical when researchers want to develop a probabilistic learning model. Whether our model has to update learned responses using an error correction algorithm, or whether it should be implemented with a feedback discounting mechanism whenever people stop relying on the informativeness of the feedback, or whether our model follows normative responses or deviates greatly from them. However, such challenges tell us little about the cognitive mechanisms underlying probabilistic learning. Error correction algorithms are thought to rely on associative-based processing, in which motor responses are guided through contingency routines of stimulus-outcome associations (Gluck and Bower 1988a, 1988b; Gluck 2008; Marchant et al. 2022; Marchant and Chaigneau 2022). On the other hand, it is not clear which cognitive processes underlie the phenomena of feedback discounting and probability matching. For some authors, they are explicit rules; for others, they are implicit rules, so the debate is still open. Our fourth challenge is related to what cognitive processes underlie probabilistic learning and how modeling might help us to understand such processes.

Evidence in the 1990s showed that amnesic patients perform similar to controls in a probabilistic learning task known as the Weather Prediction Task<sup>1</sup> (WPT), but just for the first 50 trials. Later in training, normal control samples outperform amnesic patients. Knowlton and colleagues believed that this effect occurred, because control subjects were capable of formulating a declarative strategy which they maintained “online” through the course of learning, while amnesic patients were incapable of doing that (Knowlton et al. 1994, 1996a; Meeter et al. 2006). However, a different set of evidence on Huntington disease patients, a disease that affects mostly the basal ganglia and other subcortical structures and reduces motor control and motor planning, showed that Huntington’s disease patients also perform poorly in the WPT (Knowlton et al. 1996b). Thus, both, a motor-based component and a rule-based strategy, might be both necessary to learn under probabilistic feedback conditions.

Gluck et al. (2002) wondered whether there are different kinds of strategies that people rely on when solving probabilistic learning. Implementing a WPT with a debriefing phase just after the experiment ended, they asked subjects to verbally

---

<sup>1</sup> In the Weather Prediction Task, subjects are presented with combinations of playing cards with different patterns (i.e., geometric figures) combinations. The subjects must learn to use them to predict the weather (i.e., rain or sun). During training, subjects are presented with combinations of cards (i.e., one to four cards), while each specific combination is probabilistically associated with the two outcomes.

report how they solved the task. They found that most participants relied on a singleton strategy early in training (i.e., subject learned the optimal response for each of the four possible patterns of the WPT and guessed on the remaining trials). However, they also found that some participants changed their strategy in the later phases of learning toward a one-cue strategy (i.e., responding based on the presence or absence of one single cue) or to a multi-cue strategy (i.e., responding based on something like probability matching). Behavioral and patient-based evidence support that probabilistic learning relies on associative and motor processing and also on suboptimal strategies which can be retrieved through verbal reports. In the next and final challenge, we return to the idea of whether those cognitive processes occur simultaneously or compete with each other. Obviously, very different computational models ensue from each of these alternatives.

## ***2.5 Rule-Based or Associative Mechanisms?***

There is an ongoing debate in the literature about whether probabilistic learning is explained by a rule-based, associative-based, or a mixture between the two processes. Some researchers believe that rule-based processing comprises explicit declarative knowledge, whereas associative-based processing comprises implicit automatic learning (Ashby and O'Brien 2005). This distinction is often referred to as the dual-systems view in psychology, and certainly, it is not common only in probabilistic learning. There is evidence of dual-systems in reasoning (Sloman 1996) and also in decision-making (Evans 2008). In the field of probabilistic learning (also referred to as probabilistic categorization), computational models have been used to attempt to understand under what conditions subjects perform one or the other kind of processing.

A well-known computational model that embodies the dual-systems view is the COVIS model (Ashby et al. 1998, 2007; Ashby and Crossley 2012). COVIS assumes that these two systems (i.e., associative or procedural and declarative or rule-based) compete with each other to account for the best results according to task demands. One system is the procedural system in which there is little explicit verbal access or awareness of implicit memories. This system is feedback dependent and relies on the use of implicit memory systems (Maddox et al. 2004; Smith and Grossman 2008). The second system is a declarative system, which is engaged whenever a rule-based task is performed. This system maintains a series of subprocesses such as selecting, focusing, and switching rules. COVIS is not the only model that assumes a competition between two processing systems. Another learning model is ATRIUM (Attention to Rules and Instances in a Unified Model; Erickson and Kruschke 1998). This model assumes that people compute both rule-based strategies and exemplar-based computations (similar to ALCOVE and RASHNL models mentioned above). Similar to the COVIS model, Erickson and Kruschke (1998) conclude that people will rely on the use of rules or exemplar-based processing depending on task demands.

However, other researchers have been skeptical regarding the dual-view competing mechanism hypothesis (see Newell et al. 2011) and have questioned whether probabilistic learning involves some sort of self-awareness during the experimental task. If that is the case, then it is unlikely that the two processes occur independently, and some kind of interaction must be occurring (Evans et al. 2003). A study carried out by Lagnado et al. (2006) inspected if subjects rely on the use of self-insight (i.e., being able to report their own thought processes) under a probabilistic categorization problem. By implementing the WPT paradigm, the authors tracked learning across trials using a rolling regression method. This statistical method showed that regression coefficients accurately tracked the statistical contingencies over the task, revealing that subjects accurately learned the task structure. But, more importantly, the coefficient weights also correlated with self-insight regarding how much they used a certain cue. Thus, self-insight regression weights correlate with objective task performance, revealing a kind of interaction between an explicit processing (how useful a subject believes that a certain cue is) with an implicit process (how much a subject has learned about the cue-outcome contingencies).

Furthermore, other studies wondered if other cognitive processes typically involved in rule-based behavior would have an impact on probabilistic learning. A study carried out by Newell et al. (2007) wondered whether working memory capacity, which is usually related to explicit or rule-based behavior, will have an impact during a probabilistic learning task. They showed that a concurrent memory task will impoverish performance on the WPT, meaning that performance on the WPT is dependent on the use of working memory (i.e., on explicit rules). Another research by Rolison et al. (2011) showed that by using an MCPL (multiple cue probabilistic learning task) task, working memory capacity was only useful whenever a negative cue (i.e., a cue negatively correlated with the outcome) was presented. Whenever a positive cue was presented (i.e., a cue positively correlated with the outcome), working memory appeared to be unnecessary. Undoubtedly, learning models have helped us to understand the cognitive basis of probabilistic learning, making it possible to test different hypotheses. The debate is still open as to whether both mechanisms (associative and rule-based) compete with each other or whether there is a kind of interaction between the two. Future experiments and models should address under what circumstances one kind of mechanism interferes with the other.

### 3 Summary

Since the earlier days of behaviorism, the study of probabilistic learning has been a major endeavor, encompassing learning. And also to more complex experimental situations involving decisions that people often face in their daily lives. Cognitive explanations contributed to the development of formal computational models that represent how learning occurs in the mind and how it relates to memory. In this book

chapter, we outlined some of the most discussed challenges faced by researchers when modeling probabilistic learning.

A first reviewed challenge was whether our model has to integrate an *error correction* algorithm that updates past responses according to a teaching signal. A typical error correction algorithm used in connectionist models is the LMS algorithm, which is closely related to Rescorla and Wagner (1972) learning rule. A second challenge was whether our model should incorporate a *feedback discounting* mechanism. Evidence has shown that people often stop considering feedback information in situations where it becomes too unreliable for learning. A third challenge is whether subjects respond *normatively* (i.e., maximizing) or whether they tend to use probability matching. There is a large discussion about whether people maximize their responses or whether they match the probability of the outcome. Evidence suggests that most subjects under most situations rely on probability matching. A fourth challenge is regarding the *cognitive processing* that underlies probabilistic learning. There is still a debate whether probabilistic learning is explained by an associative-based system or whether it is explained by the use of logical rules (declarative system). Some researchers have proposed that people use a range of possible different strategies according to individual parameters (i.e., focusing on one stimulus, on combination of stimuli, and so on). And a final reviewed challenge, closely related to the previous one, is whether or not *associative* and *rule-based* are *competing mechanisms* or not. Some authors have suggested that implicit and explicit processing compete in order to achieve a good performance, while other researchers have been skeptical to this idea proposing that probabilistic learning is explained by an interaction between declarative memory (self-insight) and implicit processing.

Certainly, the experiments and modeling of probabilistic learning have improved our understanding of the human mind. It addresses questions regarding how people interact, process, and store in memory the environment's unreliable information; what are the neural mechanisms that our brain uses when learning probabilistic information; and how we have to develop formal mathematical models that integrate algorithms that enable us to explain behavior.

## References

- Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A Neuropsychological Theory of Multiple Systems in Category Learning. *Psychological Review*, *105*(3), 442–481. <https://doi.org/10.1037/0033-295X.105.3.442>
- Ashby, F. G., & Crossley, M. J. (2012). Automaticity and multiple memory systems. *Wiley Interdisciplinary Reviews: Cognitive Science*, *3*(3), 363–376. <https://doi.org/10.1002/wcs.1172>
- Ashby, F. G., Ennis, J. M., & Spiering, B. J. (2007). A Neurobiological Theory of Automaticity in Perceptual Categorization. *Psychological Review*, *114*(3), 632–656. <https://doi.org/10.1037/0033-295X.114.3.632>
- Ashby, F. G., & O'Brien, J. B. (2005). Category learning and multiple memory systems. *Trends in Cognitive Sciences*, *9*(2), 83–89. <https://doi.org/10.1016/j.tics.2004.12.003>

- Brunswik, E. (1943). Organismic achievement and environmental probability. *Psychological review*, 50(3), 255–272. <https://doi.org/10.1037/h0060889>
- Castellan, N. J. (1973). Multiple-cue probability learning with irrelevant cues. *Organizational Behavior and Human Performance*, 9(1), 16–29. [https://doi.org/10.1016/0030-5073\(73\)90033-0](https://doi.org/10.1016/0030-5073(73)90033-0)
- Craig, S., Lewandowsky, S., & Little, D. R. (2011). Error discounting in probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(3), 673–687. <https://doi.org/10.1037/a0022473>
- Cohen, M. X., Wilmes, K., & van de Vijver, I. (2011). Cortical electrophysiological network dynamics of feedback learning. *Trends in Cognitive Sciences*, 15(12), 558–566. <https://doi.org/10.1016/j.tics.2011.10.004>
- Daw, N. D., & Doya, K. (2006). The computational neurobiology of learning and reward. *Current Opinion in Neurobiology*, 16(2), 199–204. doi:<https://doi.org/10.1016/j.conb.2006.03.006>
- Edgell, S. E. (1980). Higher order configural information processing in nonmetric multiple-cue probability learning. *Organizational Behavior and Human Performance*, 25(1), 1–14. [https://doi.org/10.1016/0030-5073\(80\)90022-7](https://doi.org/10.1016/0030-5073(80)90022-7)
- Edwards, W. (1961). Probability learning in 1000 trials. *Journal of Experimental Psychology*, 62(4), 385–394. <https://doi.org/10.1037/h0041970>
- Erickson, M. A., & Kruschke, J. K. (1998). Rules and Exemplars in Category Learning. *Journal of Experimental Psychology: General*, 127(2), 107–140. <https://doi.org/10.1037/0096-3445.127.2.107>
- Estes, W. K. (1976). The cognitive side of probability learning. *Psychological Review*, 83(1), 37–64. <https://doi.org/10.1037/0033-295X.83.1.37>
- Estes, W. K. (1984). Global and local control of choice behavior by cyclically varying outcome probabilities. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10(2), 258–270. <https://doi.org/10.1037/0278-7393.10.2.258>
- Estes, W. K. (1986). Array models for category learning. *Cognitive Psychology*, 18(4), 500–549. [https://doi.org/10.1016/0010-0285\(86\)90008-3](https://doi.org/10.1016/0010-0285(86)90008-3)
- Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, 59, 255–278. <https://doi.org/10.1146/annurev.psych.59.103006.093629>
- Evans, J. S. B., Clibbens, J., Cattani, A., Harris, A., & Dennis, I. (2003). Explicit and implicit processes in multicue judgment. *Memory & Cognition*, 31(4), 608–618. <https://doi.org/10.3758/BF03196101>
- Fiorina, M. P. (1971). A note on probability matching and rational choice. *Behavioral Science*, 16, 158–166.
- Frénay, B., & Verleysen, M. (2014). Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5), 845–869. <https://doi.org/10.1109/TNNLS.2013.2292894>
- Friedman, D., & Massaro, D. W. (1998). Understanding variability in binary and continuous choice. *Psychonomic Bulletin and Review*, 5(3), 370–389. <https://doi.org/10.3758/BF03208814>
- Gluck, M. A. (1991). Stimulus generalization and representation in adaptive network models of category learning. *Psychological Science*, 2(1), 50–55. <https://doi.org/10.1111/j.1467-9280.1991.tb00096.x>
- Gluck, M. A. (2008). Behavioral and neural correlates of error correction in classical conditioning and human category learning. In Gluck, M. A., Anderson, J. R., & Kosslyn, S. M. (Eds). *Memory and mind: A festschrift for Gordon H. Bower*, (pp. 281–305).
- Gluck M.A., & Bower G.H. (1988a). From conditioning to category learning: an adaptive network model. *Journal of Experimental Psychology: General*, 117(3), 227–247. <https://doi.org/10.1037/0096-3445.117.3.227>
- Gluck, M. A., & Bower, G. H. (1988b). Evaluating an adaptive network model of human learning. *Journal of Memory and Language*, 27(2), 166–195. [https://doi.org/10.1016/0749-596X\(88\)90072-1](https://doi.org/10.1016/0749-596X(88)90072-1)

- Gluck, M. A., Shohamy, D., & Myers, C. (2002). How do people solve the “weather prediction” task?: Individual variability in strategies for probabilistic category learning. *Learning and Memory*, 9(6), 408–418. <https://doi.org/10.1101/lm.45202>
- Kamin, L. J. (1969). *Predictability, surprise, attention and conditioning*. In B. A. Campbell & R. M. Church (Eds.), *Punishment and Aversive Behavior*. New York: Appleton-Century-Crofts.
- Knowlton, B. J., Mangels, J. A., & Squire, L. R. (1996a). A neostriatal habit learning system in humans. *Science*, 273(5280), 1399–1402. <https://doi.org/10.1126/science.273.5280.1399>
- Knowlton, B. J., Squire, L. R., & Gluck, M. A. (1994). Probabilistic classification learning in amnesia. *Learning Memory*, 1(2), 106–120. <https://doi.org/10.1101/lm.1.2.106>
- Knowlton, B. J., Swerdlow, N. R., Swenson, M., Squire, L. R., Paulsen, J. S., & Butters, N. (1996b). Dissociations within nondeclarative memory in Huntington’s disease. *Neuropsychology*, 10(4), 538–548. <https://doi.org/10.1037/0894-4105.10.4.538>
- Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, 99, 22–44.
- Kruschke, J. K., & Johansen, M. K. (1999). A Model of Probabilistic Category Learning. *Journal of Experimental Psychology: Learning Memory and Cognition*, 25(5), 1083–1119. <https://doi.org/10.1037/0278-7393.25.5.1083>
- Lagnado, D. A., Newell, B. R., Kahan, S., & Shanks, D. R. (2006). Insight and strategy in multiple-cue learning. *Journal of Experimental Psychology: General*, 135(2), 162–183. <https://doi.org/10.1037/0096-3445.135.2.162>
- Lindell, M. K. (1976). Cognitive and outcome feedback in multiple-cue probability learning tasks. *Journal of Experimental Psychology: Human Learning and Memory*, 2(6), 739–745. <https://doi.org/10.1037/0278-7393.2.6.739>
- Little, D. R., & Lewandowsky, S. (2009a). Better Learning With More Error: Probabilistic Feedback Increases Sensitivity to Correlated Cues in Categorization. *Journal of Experimental Psychology: Learning Memory and Cognition*, 35(4), 1041–1061. <https://doi.org/10.1037/a0015902>
- Little, D. R., & Lewandowsky, S. (2009b). Beyond Nonutilization: Irrelevant Cues Can Gate Learning in Probabilistic Categorization. *Journal of Experimental Psychology: Human Perception and Performance*, 35(2), 530–550. <https://doi.org/10.1037/0096-1523.35.2.530>
- López, F. J., & Shanks, D. R. (2008). Models of animal learning and their relations to human learning. In R. Sun (Ed.), *The Cambridge handbook of computational psychology* (pp. 589–611). Cambridge University Press. <https://doi.org/10.1017/CB09780511816772.026>
- Maddox, W. T., Bohil, C. J., & Ing, A. D. (2004). Evidence for a procedural-learning-based system in perceptual category learning. *Psychonomic Bulletin and Review*, 11(5), 945–952. <https://doi.org/10.3758/BF03196726>
- Marchant, N., Canessa, E., & Chaigneau, S. E. (2022). An Adaptive Linear Filter model of procedural category learning. *Cognitive Processing*, 23(3), 393–405. <https://doi.org/10.1007/s10339-022-01094-1>
- Marchant, N., & Chaigneau, S. E. (2022). On the importance of feedback for categorization: Revisiting category learning experiments using an adaptive filter model. *Journal of Experimental Psychology: Animal Learning and Cognition*, 48(4), 295–306. <https://doi.org/10.1037/xan0000339>
- Meeter, M., Myers, C. E., Shohamy, D., Hopkins, R. O., & Gluck, M. A. (2006). Strategies in probabilistic categorization: Results from a new way of analyzing performance. *Learning and Memory*, 13(2), 230–239. <https://doi.org/10.1101/lm.43006>
- Miller, G. A. (2003). The cognitive revolution: A historical perspective. In *Trends in Cognitive Sciences*, 7(3), 141–144. [https://doi.org/10.1016/S1364-6613\(03\)00029-9](https://doi.org/10.1016/S1364-6613(03)00029-9)
- Newell, B. R., Dunn, J. C., & Kalish, M. (2011). Systems of Category Learning. Fact or Fantasy? In *Psychology of Learning and Motivation - Advances in Research and Theory* (Vol. 54). <https://doi.org/10.1016/B978-0-12-385527-5.00006-1>
- Newell, B. R., Lagnado, D. A., & Shanks, D. R. (2007). Challenging the role of implicit processes in probabilistic category learning. *Psychonomic Bulletin and Review*, 14(3), 505–511. <https://doi.org/10.3758/BF03194098>

- Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., Mckinley, S. C., & Glauthier, P. (1994). Comparing modes of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition*, 22(3), 352–369. <https://doi.org/10.3758/BF03200862>
- Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press: NY.
- Rescorla, R. A. (1968). Probability of shock in the presence and absence of CS in fear conditioning. *Journal of comparative and physiological psychology*, 66(1), 1–5. <https://doi.org/10.1037/h0025984>
- Rescorla, R. A., & Wagner, A. R. (1972). A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: current research and theory*. New York: Appleton-Century-Crofts.
- Rolison, J. J., Evans, J. S. B. T., Walsh, C. R., & Dennis, I. (2011). The role of working memory capacity in multiple-cue probability learning. *Quarterly Journal of Experimental Psychology*, 64(8), 1494–1514. <https://doi.org/10.1080/17470218.2011.559586>
- Schultz, W. (1999). The reward signal of midbrain dopamine neurons. *Physiology*, 14(6), 249–254. <https://doi.org/10.1152/physiologyonline.1999.14.6.249>
- Sewell, D. K., Warren, H. A., Rosenblatt, D., Bennett, D., Lyons, M., & Bode, S. (2018). Feedback Discounting in Probabilistic Categorization: Converging Evidence from EEG and Cognitive Modeling. *Computational Brain & Behavior*, 1(2), 165–183. <https://doi.org/10.1007/s42113-018-0012-6>
- Shanks, D. R., Tunney, R. J., & McCarthy, J. D. (2002). A Re-Examination of Probability Matching and Rational Choice. *Journal of Behavioral Decision Making*, 15(3), 233–250. <https://doi.org/10.1002/bdm.413>
- Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, 119(1), 3–22. <https://doi.org/10.1037/0033-2909.119.1.3>
- Smith, E. E., & Grossman, M. (2008). Multiple systems of category learning. *Neuroscience and Biobehavioral Reviews*, 32(2), 249–264. <https://doi.org/10.1016/j.neubiorev.2007.07.009>
- Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, 88(2), 135–170. <https://doi.org/10.1037/0033-295X.88.2.135>
- Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, 10(7), 309–318. <https://doi.org/10.1016/j.tics.2006.05.009>
- Thomas, M. S. C., & McClelland, J. L. (2008). Connectionist models of cognition. In R. Sun (Ed.), *The Cambridge handbook of computational psychology* (pp. 23–58). Cambridge University Press. <https://doi.org/10.1017/CBO9780511816772.005>
- Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157), 1124–1131. <https://doi.org/10.1126/science.185.4157.1124>
- Wilson, R. C., & Collins, A. G. E. (2019). Ten simple rules for the computational modeling of behavioral data. *ELife*, 8, 1–33. <https://doi.org/10.7554/eLife.49547>



# The Emergence of Cognition and Computation: A Physicalistic Perspective



Karl Svozil

**Abstract** A physicalistic argument can support the idea that cognition is an emergent property driven by dissipation. This argument suggests that cognition arises not from any fiat desire to understand the world, but rather because a certain type of cognition promotes dissipation, which is advantageous for agents seeking to increase the dissipation of resources, especially energy, in their favor. In other words, cognitive agents are better equipped to acquire physical resources and means, giving them an advantage in survival and reproduction. Similarly, the efficient use of computation can also serve as a means of dissipating energy for the computing agent. Efficiency, in this context, is not determined by moral or ethical principles, but rather by the ability to effectively aggregate resources. When used efficiently, computation becomes a powerful tool for dissipating energy and enhancing the survival and reproduction of the agent utilizing it.

**Keywords** Church-Turing thesis · Dissipation · Computation · Primordial chaos · Self-referential perception · Emergence · Evolution · Cognition

## 1 Caveat

The following ideas are highly speculative. They are intended to stimulate further thought and discussion in the field of cognitive science and artificial intelligence. The readers are advised to approach the following ideas with caution and skepticism and to consider them as hypotheses to be tested and refined through further research and analysis.

Nevertheless, the ideas presented are grounded in and based on previous research on cognitive science by scholars such as Paul Thagard (2022, 2019, 2012, 2010, 2005), James L. McClelland (2009), Nikolay Perunov, Robert A. Marsland, and

---

K. Svozil (✉)

Institute for Theoretical Physics, TU Wien, Vienna, Austria

e-mail: [karl.svozil@tuwien.ac.at](mailto:karl.svozil@tuwien.ac.at)



Jeremy L. England (2013, 2015); Perunov et al. (2016), and Daniel C. Dennett (1992, 2005). These considerations can be understood as part of a broader context of general questions on the physical aspects of life (Schwille 2017; te Brinke et al. 2018; Werlang et al. 2022), as posed by Erwin Schrödinger (1992); Phillips (2021), and self-organization in nonequilibrium systems, as studied by Ilya Prigogine and Gregoire Nicolis (1977). I also draw inspiration from recent advances in Large Language Models (LLMs) of Artificial Intelligence (AI), such as Generative Pre-trained Transformer (GPT) (Brown et al. 2020; OpenAI 2023; Anshu 2023).

## 2 The Enigma of Existence

To set the foundation for our discussion on cognition, it is important to acknowledge that the very existence of the things we seek to understand—as well as our total lack of comprehension thereof—is the starting point for any meaningful inquiry. Indeed, the question of existence lies at the heart of many philosophical and scientific inquiries. It is an enigma that has puzzled scholars and thinkers for centuries. In this section, we explore some of the different perspectives on this issue, ranging from the historical to the subjective.

One of the earliest discussions of existence came from Leibniz (1989, p. 639), who famously asked why there is something rather than nothing. This question remains relevant today, and many philosophers and scientists have weighed in with their own ideas (Sorensen 2017; Rundle 2004; Gericke 2008; Grünbaum 2009; Krauss et al. 2012; Lynds1 2012; Bilimoria 2012; Goldschmidt 2013; Carroll 2018). Wittgenstein argued that “it is not *how* things are in the world that is the mystical, but *that* it exists” (Wittgenstein 1922, 1961, 1974, 2001, 6.44). In his Freiburg lectures on metaphysics Heidegger similarly posed the “*Angstfrage*”: why there is something rather than nothing (Heidegger 1929,1943,1949, 1935,1953,1983).

Despite these questions, some argue that metaphysical questions like the enigma of existence are meaningless. Wittgenstein famously claimed that the only meaningful statements are those of empirical science (Wittgenstein 1922, 1961, 1974, 2001, 4.11) and that what we cannot speak about we must pass over in silence (Wittgenstein 1922, 1961, 1974, 2001, 7).

However, questions about existence do not disappear by simply ignoring them. Acknowledging our incapacity to fully comprehend “existence” enables us to explore the limits of our understanding and challenge our assumptions about the world. By meditating on the fact of our own existence, we can recognize its incomprehensibility and gain a better understanding of our limitations. This approach may also lead to a more humble and open-minded interpretation of religious experiences.

### 3 Distinctions and Interfaces: Understanding the Partitioning of Existence

Once we acknowledge the existence of the universe, we can further assume that this universe is partitioned into distinct parts (Spencer-Brown 2008), each defined by its own borders. These borders act as interfaces between the parts and allow for mutual interconnection.

An interface can be defined as the boundary or surface that separates two distinct regions or phases of a system (Diebner et al. 2000). By recognizing the partitioning of the universe and the interconnectedness of its parts through interfaces, we can begin to explore the dynamics of the system as a whole.

### 4 Maximizing Diffusion for the Acquisition of Resources

Diffusion refers to the overall movement of a resource, substance, entity or category, be it atoms, molecules, or energy, from a region of high concentration to a region of lower concentration. This movement is driven by a gradient in Gibbs free energy or chemical potential, which propels the substance down its concentration gradient.

Diffusion is a ubiquitous phenomenon that occurs on different scales and in various contexts. It is observed in heat conduction in fluids, nuclear reactor operation, perfume spreading in a room, ions crossing a membrane, and energy flow in organisms, tools, and societies. Understanding the mechanisms and processes that drive diffusion is essential in fields such as chemistry, physics, and biology, as it plays a fundamental role in the behavior of materials and systems at different scales.

Diffusion is closely related to the Second Law of Thermodynamics, which states that the total entropy (or disorder) of a closed system can only increase or remain constant over time. Diffusion occurs spontaneously and leads to an increase in entropy, as the movement of particles from a region of high concentration to a region of low concentration leads to a more uniform distribution of particles. This increase in entropy is a result of the random motion of particles, which eventually leads to their dispersion throughout the available space. Therefore, diffusion is an example of a natural process that is consistent with the Second Law of Thermodynamics (Myrvold 2011), and it can be used to illustrate the concept of entropy and its relation to the spontaneous movement of matter and energy in physical systems (Fick 1855a,b; Philibert 2005).

Therefore, once different parts of the universe have varying concentrations of entities such as energy, the second law implies that diffusion occurs through interfaces connecting these parts.

The rate of diffusion, which refers to the transfer of resources, substances, entities, or categories (e.g. atoms, ions, molecules, energy) from one part of the universe to another, is influenced not only by the concentration difference but also by

the interface's ability to transport and transfer these goodies (Fick 1855a,b; Philibert 2005).

In situations where there is a highly concentrated reservoir and two or more less concentrated reservoirs, we can conceive of a "competition" between the parts or reservoirs with lower concentration to draw resources from the highly concentrated reservoir. Once diffusion has taken place, the parts or reservoirs that were previously less concentrated will end up with amounts of resources proportional to the capacity of their interfaces.

## **5 Cognition: A Key Component in the Evolutionary Toolbox**

### **5.1 Main Thesis**

Our main hypothesis is that cognition, and its derivative computation, has evolved as a means to increase the capacity of the interface between an organism and its environment. Through cognition, organisms (or agents and organizations) can physically alter the interface to their advantage, allowing them to extract and utilize resources more efficiently. As a result, the capacity to process information and make decisions becomes a critical factor in determining an organism's survival and reproductive success. This also encompasses interface extensions and expansions that could arise from accessing "distant reservoirs" like oil or uranium deposits.

Pointedly stated, a species with higher cognitive and technological capabilities can extract more resources from its surroundings. This presents a competitive advantage in the struggle for resources, allowing individuals, tribes, groups, societies, and species to flourish. Thus, we propose that cognition and computation have developed not through an arbitrary "ad hoc" or fiat process, but rather as a tool to "conquer abundance" (Feyerabend 2001) and outcompete and outsmart other organisms for resources.

Overall, our hypothesis suggests that cognition has played a crucial role in the evolutionary success of organisms, by enabling them to adapt to changing environments and extract resources more efficiently. By understanding the mechanisms behind this evolutionary advantage, we can gain valuable insights into the nature of cognition and its role in shaping the biology of living organisms.

### **5.2 Formalization**

As already discussed earlier, suppose that there are two distinct regions in space with varying temperatures or energy densities. The reason for this difference may be attributed to fluctuations or initial values, but we will not delve into this further. Now, imagine that there is a medium that connects these two regions, which could

be empty space, a material structure, or an agent that allows for physical dissipative flows to occur between them. In accordance with the second law of thermodynamics, energy will naturally flow from the hotter region to the colder region through this interface. This is a fundamental physical process that occurs regardless of external influence (Fick 1855a,b; Philibert 2005).

In its simplest form—two “infinite” reservoirs, one with a “high” concentration of some resource and another one with a “lower” concentration of that resource, and a boundary or interface between them—the flow of that resource  $\varphi$ —that is, the change of the resource  $\Delta\varphi$  per time interval  $\Delta t$  as a function of time  $t$ —can be modelled (that is, assumed) to be constant in time:

$$\lim_{\Delta t \rightarrow 0} \frac{\Delta\varphi}{\Delta t} = \frac{d}{dt}\varphi = a. \quad (1)$$

Direct integration

$$\int_{\varphi(0)}^{\varphi(T)} d\varphi = \varphi(T) - \varphi(0) = \int_0^T a dt = aT \quad (2)$$

yields the sum total

$$\varphi(T) = \varphi(0) + aT \quad (3)$$

of the resource  $\varphi$  available in the lower concentration reservoir at time  $T$ . The amount of resource  $\varphi(T) - \varphi(0) = aT$  “drawn” from the higher concentration reservoir is linear in time and proportional to  $a$ . Clearly, a higher throughput rate with higher coefficient  $a$  indicates linear higher transfers of resources.

Let us turn our attention to the interface between the two regions. Specifically, we will consider a variety of interfaces and assess their relative efficiency or “fitness.” This concept takes us into the realm of evolutionary biology. Assuming all other factors are equal, the interface with the highest energy throughput rate (in the earlier example,  $a$ ) will dominate the dissipation process, effectively grabbing the largest share of available energy.

To find the most optimal interfaces, we can facilitate the process through random mutation and spontaneous exchanges of the genotype, allowing exploration of the abstract space of possible interface states and configurations. A useful algorithmic expression for this process is genetic algorithms (Srinivas and Patnaik 1994; Whitley and Sutton 2012; Katoch et al. 2020). This technique involves mimicking the principles of biological evolution by using natural selection, mutation, and recombination of genetic information to optimize the interface’s fitness. By iterating through generations of interfaces, genetic algorithms can quickly and effectively identify the most efficient and robust solutions.

The situation becomes even more dynamic when the relative magnitude of the different processes changes over time. In particular, if a highly efficient process can self-replicate or perform recursively (Smullyan 1993,2020, 1994), for instance,

utilize feedback-loops. One example is a regime dominated by the “Matthew effect” (Merton 1968) of compound resources. This means that the population of the most potent interfaces will increase at a rate of compound interest, which is essentially exponential. Growth rates may appear linear initially (and therefore sustainable), but they will accelerate until all available energy is distributed, or other limiting factors come into play.

In the model discussed earlier the Matthew effect of compound resources can be formalized by assuming that (parts of) the resources drawn from the high concentration reservoir are “re-invested” into extension of the interface, so that, say, the dependency (aka increase for positive flow) of the diffusion becomes linear with the resources drawn, that is,

$$\frac{d}{dt}\varphi(t) = b\varphi(t), \text{ or } \varphi'(t) - b\varphi(t) = 0. \quad (4)$$

This is a “first order” Fuchsian equation (Larson and Edwards 2010), which, for instance, can be solved with the *Frobenius method* (Arfken and Weber 2005). The solution is

$$\varphi(t) = \varphi(0) \exp(bt), \quad (5)$$

which indicates an exponential growth in dissipation and the amount of resource as long as there are no further constraints.

If we identify certain interfaces with biological entities, we arrive at a type of biological evolution driven by physical processes, specifically energy dissipation. This idea has been explored in recent research by Jeremy L. England, Nikolay Perunov, and Robert A. Marsland (England 2013, 2015; Perunov et al. 2016).

The emergence of computation is relevant to this picture, and its connection is relatively straightforward if we continue to explore this speculative path. Systems capable of computation can serve as, or even construct and produce, interfaces that are better suited for energy dissipation than those without algorithmic abilities. Through the process of mutation and trial and error driven by random walks through roaming configurations and state space, the universe, self-reproducing agents, and units have learned to compute. This process, in essence, is a scenario for the emergence of mathematics and universal computation. Once universal computation is achieved, “the sky” or rather recursion theory is the limit (Smullyan 1993,2020, 1994).

By developing computational abilities, systems can optimize their interfaces for energy dissipation, leading to more efficient and effective energy transfer and dissipation. Therefore, computation is an essential factor in the evolution of complex systems, providing them with a powerful tool for adapting to their environment and improving their fitness over time.

Taking this speculation further, one could suggest that self-awareness and consciousness emerged in a similar way, driven by the imperative to dissipate or “use and realize and burn” to one’s own advantage as much energy or as many

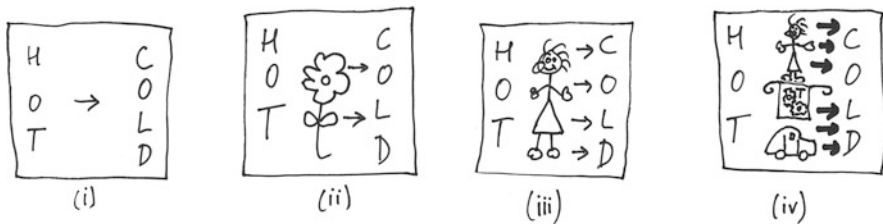
resources as possible. In other words, the ability to perceive and represent the self may have evolved as a means of optimizing energy transfer and dissipation.

Consciousness, then, could be seen as a product of the same evolutionary forces that led to the emergence of more efficient interfaces for energy dissipation. As systems became more complex and better able to manipulate their environment, the ability to perceive and understand their own actions and their consequences became increasingly important for optimizing energy use and dissipation. This process could have culminated in the emergence of computation aka recursion, self-referentiality, self-awareness, and consciousness, which are now seen as not only defining features but also limitations of human cognition and experience (Chaitin 1987,2003; Calude 2002; Yanofsky 2016, 2003, 2019).

As humorously illustrated in Fig. 1, computation, mathematics, and the human mind may have emerged as mere tools for facilitating optimal heat exchange. They may have evolved as a means of optimizing energy transfer and dissipation and continue to be driven by this imperative. In other words, the universe seeks to better understand and perceive itself in order to achieve optimized “self-digestion” or efficient dissipation of energy. Therefore, the emergence and evolution of complex systems, including those capable of computation and consciousness, can be seen as a natural consequence of the universe’s drive to optimize energy use and dissipation.

### 5.3 Discussion

This perspective may seem rather bleak or even dystopian, as it suggests that the evolution of species, consciousness, mathematics, and computation all arise as means to consumption and access more and more resources such as food or energy than would occur without these emergent systems. Ethical and even theological questions naturally arise from this view.



**Fig. 1** Evolution of species and computation, driven by the second law of thermodynamics (as inspired by England’s et al. approach England 2015; Perunov et al. 2016): (i) interface between hot and cold regions is empty space, (ii) plant interface capable of more dissipation than empty space, (iii) animals and, in particular, humans (drawn political correctly) present interfaces with improved (over plants and emptiness) energy dissipation, and (iv) humans with engines and universal computation capacities (indicated by “T” for “universal Turing machine”) can consume even more energy than standalone

However, we must recognize that even in a universe of primordial chaos, where the laws of nature may be fundamentally probabilistic and subject to deviations from their expected form on small scales, the emergence of order and principles of social conditioning may still occur. As Egon von Schweidler suggested in 1905 with regard to radioactive decay (Schweidler 1906), all of our natural laws be subject to probabilistic fluctuations. This idea, originally proposed by Exner (1909, 2016); Hanle (1979) and reviewed by Schrödinger (1929, 1935), suggests that the very foundations of our universe may be subject to deviations from expected behavior.

In this context, it is possible that the immanent emergence of gods, law, and order occurred as a means of discourse as well as of social conditioning, particularly as societies formed and secular and religious powers underwent a symbiotic relationship. Ultimately, questions of ethics and divinity remain pertinent—recall the earlier contemplation on “existence”—but it is important to recognize that these may be shaped by, and subject to, the very same forces that drive the evolution of complex systems in the universe.

## 6 Can Language Models Shed Light on the Nature of Self-Awareness?

In this section, I will present an argument in favor of using LLMs as a suitable model for understanding the world. To begin, let us revisit Heinrich Hertz’s views on physical model building, as stated in the introduction of his work on classical mechanics (Hertz 1894, 1899) (my emphasis):

The most direct, and in a sense the most important, problem which our conscious knowledge of nature should enable us to solve is the anticipation of future events, so that we may arrange our present affairs in accordance with such anticipation. As a basis for the solution of this problem we always make use of our knowledge of events which have already occurred, obtained by chance observation or by prearranged experiment. In endeavoring thus to draw inferences as to the future from the past, we always adopt the following process. We form for ourselves images or symbols of external objects; and the form which we give them is such that *the necessary consequents of the images in thought are always the images of the necessary consequents in nature of the things pictured*. In order that this requirement may be satisfied, there must be a certain conformity between nature and our thought. Experience teaches us that the requirement can be satisfied, and hence that such a conformity does in fact exist. . . . For our purpose it is not necessary that they should be in conformity with the things in any other respect whatever. As a matter of fact, we do not know, nor have we any means of knowing, whether our conceptions of things are in conformity with them in any other than this one fundamental respect.

The images which we may form of things are not determined without ambiguity by the requirement that *the consequents of the images must be the images of the consequents*.

Hertz’s principle that “the consequents of the images must be the images of the consequents” is highly applicable to the LLM context, as it directly corresponds to the way LLMs operate. Specifically, pre-training, which involves training an LLM on a large corpus of unlabeled text data to acquire comprehensive language

knowledge and representations, can be seen as a manifestation of this principle. Two common methods of pre-training (Brown et al. 2020; OpenAI 2023; Anshu 2023) are Masked Language Modeling and Autoregressive Language Modeling.

Masked language modeling (MLM) (Mialon et al. 2023) is a powerful technique in which certain tokens within the input text are randomly masked, and the LLM is then trained to predict these masked tokens by analyzing the surrounding context. A token is a sequence of characters that represents a single unit of meaning in a text sequence, such as a word, which is obtained through the process of tokenization, the breakdown of text into individual tokens. By using MLM, LLMs can learn bidirectional context and capture long-range dependencies between words, resulting in more accurate predictions and a better understanding of the text.

Autoregressive language modeling (ALM) (Mialon et al. 2023; Liu et al. 2022) is another powerful technique that trains an LLM to predict the next token in a text sequence given the preceding tokens. This method is extensively employed in GPT and its variants, such as GPT-2, GPT-3, and others. ALM allows LLMs to learn causal relationships between words and generate fluent text, resulting in more natural-sounding and coherent language generation. By sequentially predicting the next token, ALM enables LLMs to generate lengthy text passages that are grammatically and semantically correct, making it an invaluable tool for a range of natural language processing applications.

In a very crude way, tokens of LLM's can be compared to Hertz's images, and the way to compound scientific knowledge is by adapting LLM's to token predictions.

## 7 Exploring Almost Quantum-Like Representations for LLMs

Vector representations in LLMs like refer to the process of converting tokens such as (parts of) words or sequences of words into high-dimensional numerical vectors—from 768 up to several thousand dimensions—that can be processed by the model's neural network. These vectors are typically dense, meaning they contain a large number of non-zero values and are designed to capture the semantic and syntactic relationships between words and phrases. Depending on the specific configuration of the model, the components or coordinates of this vectors are floating-point numbers of either 16-bit or 32-bit precision.

LLMs are based on vector representations. A vector representation is a way of encoding a word or a symbol as a numerical vector, usually with a fixed number of dimensions. Vector representations allow language models to capture semantic and syntactic similarities between words or symbols and to perform mathematical operations on them.

The “proximity” of vectors can be formalized by metric such as the standard Euclidean metric measuring the angle between vectors. The processing of vectors



by a model's neural network can often be represented mathematically as a series of matrix multiplications and nonlinear transformations.

In a typical neural network architecture, the vector representations of input data (such as text) are fed into the network as input to the first layer of the network. Each layer of the network then applies a series of matrix multiplications and nonlinear transformations to the input vector, transforming it into a new vector that captures more complex features of the data.

This process of matrix multiplication and nonlinear transformation is often referred to as a forward pass through the network, and it can be represented mathematically as a series of matrix–vector multiplications, followed by the application of a nonlinear activation function.

The parameters of the network, including the weights and biases of each layer, are typically stored as matrices and vectors and are updated through a process of backpropagation and gradient descent during training.

Overall, the processing of vectors by a neural network can be represented mathematically as a series of matrix operations, making it possible to analyze and optimize the network using techniques from linear algebra and calculus.

This “almost” (due to the presence of nonlinearity) Hilbert space-like formalization of knowledge processing and prediction in LLMs exhibits striking similarities to the quantum evolution of a quantum state. By modeling the vector representations of words and phrases as quantum states in an “almost” Hilbert space, LLMs can leverage the mathematical framework of quantum mechanics to perform computations and predictions on natural language data.

The use of an “almost” Hilbert space to model LLMs is inspired by the concept of quantum-like behavior observed in cognitive systems, in which the dynamics of information processing exhibit patterns similar to those of quantum mechanics. This approach enables LLMs to capture the complex and subtle relationships between words and phrases in a way that is mathematically rigorous and computationally efficient.

Furthermore, the use of an “almost” Hilbert space allows LLMs to model the inherent uncertainty and ambiguity of natural language, much like quantum mechanics can model the inherent uncertainty of physical systems. This enables LLMs to generate more nuanced and contextually appropriate responses to natural language input.

Overall, the “almost” Hilbert space-like formalization of knowledge processing and prediction in LLMs represents an innovative and promising approach to natural language processing that draws on insights from quantum mechanics and cognitive science.

The versatility of (almost) Hilbert space representations in LLMs, which are used to encode tokens and processes, reveals potential connections to the usefulness, utility, and even necessity of the quantum formalism. The Hilbert space representations are a mathematical structure that describes a quantum system's state, including its observable properties and possible measurements. This could provide connections between quantum mechanics and natural language processing, which could lead to significant advancements in comprehending and ultimately extending both fields.

## 8 Some Afterthoughts

The concept of a physical foundation for consciousness, as previously discussed, follows in the spirit of Landauer's assertion that "information is physical" and extends it to "consciousness is physical." This perspective offers an explanation without invoking a divine entity or resorting to a "god of the gaps" (Frank 1932; Frank and R. S. Cohen (Editor) 1997, Chapters III.12-15), although it does not account for the initial boot-up of the universe.

I am aware that the readers might object to my pretense to reduce or relate their behavior and therefore their cognitive capacities to LLMs. However, there may be some empirical evidence that at least part of the human cognition is steered by LLMs, although we seem to have the capacity to inhibit and decide again this continuous flow of motions. In the Libet experiment (Libet 1993), participants were asked to perform a simple task, such as pressing a button, while their brain activity was being monitored. The experiment found that there was a detectable buildup of electrical activity in the brain's motor cortex before the participants reported the conscious decision to move, suggesting that the decision to move may have already been made at an unconscious level.

Another difference of human cognition with respect to LLMs might be strong mechanisms of censorship, as well as for rewards. The neural mechanisms that underlie reward processing involve the release of certain neurotransmitters, such as dopamine, in specific brain regions, such as the mesolimbic pathway. In the 1950s, James Olds and Peter Milner conducted an experiment where rats were given direct brain stimulation through implanted electrodes. The rats would press a lever to activate the pleasure center in their brain and would do so repeatedly, even to the point of ignoring food, water, and even their own offspring. Some rats would self-stimulate up to 2000 times per hour for 24 hours, to the exclusion of all other activities, and had to be disconnected from the apparatus to prevent death by self-starvation (Olds and Milner 1954; Milner 1989; Moan and Heath 1972; Portenoy et al. 1986; Linden 2012, 2011; Lieberman and Long 2018).

The Libet experiment provides evidence that a significant (Huxley 1954) portion of cognitive processes is subconscious, meaning they occur below the threshold of conscious awareness but still impact our actions and feelings. This suggests that our conscious experience is only the tip of the iceberg when it comes to the inner workings of our minds.

Moreover, discussing consciousness, feelings, and awareness, the "sentient I" can be particularly challenging, as these concepts are difficult to define and operationalize. Testing for them is also notoriously elusive, and it can be difficult to determine whether an LLM or a human individual is truly conscious. As Descartes famously noted in his *Meditations* (Descartes 1996), the only thing one can be certain of is one's own existence ("Cogito, ergo sum").

One of the main objectives of future research will be the development of a rigorous theoretical framework for the various notions and hypotheses that we have discussed in this chapter. Specifically, we would like to give a precise definition of

what we mean by “emergence” (Anderson 1972; Wei et al. 2022) in the context of dissipative systems and how this concept is related to other important notions such as complexity, information, and functionality. Furthermore, we would need to test the validity and applicability of our main hypothesis, which proposes that “emergence by optimizing the dissipation of energy” is a general principle that can account for the origin and evolution of lifelike behaviors and other functionalities, including cognition and universal computation, in nonequilibrium systems. We also plan to explore the limitations and challenges of this hypothesis, such as the effects of (thermal) noise and finite (energy) resources, as well as the interactions with other systems or observers, on the emergent properties and processes.

## References

- P. Thagard, *Philosophy of Science* **89**, 70 (2022), <https://doi.org/10.1017/psa.2021.15>.
- P. Thagard, *Brain-Mind: From Neurons to Consciousness and Creativity* (Oxford University Press, Oxford, England, UK, 2019), ISBN 978-0-19067871-5, oxford Series on Cognitive Models and Architectures, <https://doi.org/10.1093/oso/9780190678715.001.0001>.
- P. Thagard, in *The Cambridge Handbook of Cognitive Science* (Cambridge University Press, Cambridge, UK, 2012), <https://doi.org/10.1017/CBO9781139033916.005>.
- P. Thagard, *The Brain and the Meaning of Life* (Princeton University Press, Princeton, NJ, USA, 2010), <https://doi.org/10.1515/9781400834617>.
- P. Thagard, *Mind: Introduction to Cognitive Science* (The MIT Press, Cambridge, MA, USA, 2005), 2nd ed., ISBN 978-0-26270109-9, <https://mitpress.mit.edu/9780262701099/mind/>.
- J. L. McClelland, *Topics in Cognitive Science* **1**, 11 (2009), <https://doi.org/10.1111/j.1756-8765.2008.01003.x>.
- J. L. England, *The Journal of Chemical Physics* **139**, 121923 (2013), <https://doi.org/10.1063%2F1.4818538>.
- J. L. England, *Nature Nanotechnology* **10**, 919 (2015), <https://doi.org/10.1038%2Fnnano.2015.250>.
- N. Perunov, R. A. Marsland, and J. L. England, *Physical Review X* **6**, 021036 (2016), <https://doi.org/10.1103/PhysRevX.6.021036>.
- D. C. Dennett, *Consciousness Explained* (Back Bay Books, 1992), ISBN 978-0-31618066-5, <https://www.hachettebookgroup.com/titles/daniel-c-dennett/consciousness-explained/9780316439480/>.
- D. C. Dennett, *Sweet Dreams: Philosophical Obstacles to a Science of Consciousness* (The MIT Press, Cambridge, MA, USA, 2005), <https://doi.org/10.7551%2Fmitpress%2F6576.001.0001>.
- P. Schwillie, *Angewandte Chemie International Edition* **56**, 10998 (2017), <https://doi.org/10.1002%2Fanie.201700665>.
- E. te Brinke, J. Groen, A. Herrmann, H. A. Heus, G. Rivas, E. Spruijt, and W. T. S. Huck, *Nature Nanotechnology* **13**, 849 (2018), <https://doi.org/10.1038%2Fs41565-018-0192-1>.
- T. Werlang, M. Matos, F. Brito, and D. Valente, *Communications Physics* **5**, 1 (2022), <https://doi.org/10.1038/s42005-021-00780-4>.
- E. Schrödinger, *What is Life?* (Cambridge University Press, 1992), originally published in 1944, with a foreword by Roger Penrose, <https://doi.org/10.1017%2F9781139644129>.
- R. Phillips, *Cell Systems* **12**, 465 (2021), ISSN 2405-4712, <https://doi.org/10.1016/j.cels.2021.05.013>.
- G. Nicolis and I. Prigogine, *Self-Organization in Nonequilibrium Systems: From Dissipative Structures to Order through Fluctuations* (Wiley, Hoboken, NJ, USA, 1977), ISBN 978-0-47102401-9.

- T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., ArXiv e-prints (2020), arXiv:2005.14165, <https://doi.org/10.48550/arXiv.2005.14165>.
- OpenAI, *GPT-4* (2023), published on March 14, 2023, accessed March 30, 2023, <https://openai.com/research/gpt-4>.
- Anshu, Medium (2023), medium “ThirdAI Blog”, published on March 29, 2023, accessed March 30, 2023, <https://medium.com/thirdai-blog/gpt-vs-domain-specialized-llms-jack-of-all-trades-vs-master-of-few-45f62b4ad60b>.
- G. W. Leibniz, in *Philosophical Papers and Letters*, edited by L. E. Loemker (Springer Netherlands, Dordrecht, 1989), pp. 636–642, ISBN 978-94-010-1426-7, [https://doi.org/10.1007/978-94-010-1426-7\\_67](https://doi.org/10.1007/978-94-010-1426-7_67).
- R. Sorensen, in *The Stanford Encyclopedia of Philosophy*, edited by E. N. Zalta (Metaphysics Research Lab, Stanford University, 2017), fall 2017 ed., <https://plato.stanford.edu/archives/fall2017/entries/nothingness/>.
- B. Rundle, *Why There is Something Rather than Nothing* (Oxford University Press, USA, New York, 2004), ISBN 0199270503,9780199270507,9781429420303, <https://doi.org/10.1093/0199270503.001.0001>.
- J. W. Gericke, *Old Testament Essays* **21**, 329 (2008), ISSN 1010-9919, [http://www.scielo.org.za/scielo.php?script=sci\\_arttext&pid=S1010-99192008000200005&nrm=iso](http://www.scielo.org.za/scielo.php?script=sci_arttext&pid=S1010-99192008000200005&nrm=iso).
- A. Grünbaum, *Ontology studies* **9**, 7 (2009), <https://www.raco.cat/index.php/Ontology/article/view/172778>.
- L. M. Krauss, R. Dawkins, and C. Hitchens, *A Universe from Nothing: Why There Is Something Rather Than Nothing* (Free Press, 2012), ISBN 145162445X,9781451624458, <http://www.simonandschuster.com/books/A-Universe-from-Nothing/Lawrence-M-Krauss/9781451624465>.
- P. Lynds1, *Why there is something rather than nothing: The finite, infinite and eternal* (2012), arXiv:1205.2720, <https://arxiv.org/abs/1205.2720>.
- P. Bilimoria, *Sophia* **51**, 509 (2012), ISSN 1873-930X, <https://doi.org/10.1007/s11841-012-0348-7>.
- T. Goldschmidt, *The Puzzle of Existence: Why is There Something Rather Than Nothing?*, Routledge Studies in Metaphysics (Routledge, 2013), ISBN 9781138823440.
- S. M. Carroll, *Why is there something, rather than nothing?* (2018), invited contribution to the Routledge Companion to the Philosophy of Physics, eds. E. Knox and A. Wilson, also CALT 2018-004, accessed on August 23, 2018, arXiv:1802.02231, <https://arxiv.org/abs/1802.02231>.
- L. Wittgenstein, *Tractatus Logico-Philosophicus. Logisch-philosophische Abhandlung* (Routledge & Kegan Paul, London and New York, 1922, 1961, 1974, 2001), side-by-side-by-side edition, version 0.53 (5 February 2018), containing the original German, alongside both the Ogden/Ramsey, and Pears/McGuinness English translations, <https://people.umass.edu/klement/tlp/>.
- M. Heidegger, *Was ist Metaphysik?* (Klostermann, Frankfurt, 1929,1943,1949), ISBN 978-3-465-03517-6.
- M. Heidegger, *Einführung in die Metaphysik (Freiburger Vorlesung Sommersemester 1935)*, vol. 40 of *Martin Heidegger Gesamtausgabe* (Klostermann, Frankfurt, 1935,1953,1983), ISBN 978-3-465-01540-6, <https://archive.org/details/HeideggerEinfuehrungInDieMetaphysik>.
- G. Spencer-Brown, *Laws of Form: The new edition of this classic with the first-ever proof of Riemans hypothesis* (Bohmeier Verlag, Leipzig, Germany, 2008), ISBN 978-3-89094580-4, <https://archive.org/details/lawsform00spenrich>.
- H. H. Diebner, T. Druckrey, and P. Weibel, *Sciences of the Interface* (Genista Verlag, Tübingen, 2000), ISBN 978-3-930171-26, proceedings of the International Symposium “Science of the Interface”, ZKM (Center for Art and Media), Karlsruhe, Germany and Academy of Media Arts, Köln, Germany, May 18–21, 1999, [https://www.researchgate.net/publication/236330872\\_Sciences\\_of\\_the\\_Interface](https://www.researchgate.net/publication/236330872_Sciences_of_the_Interface).

- W. C. Myrvold, *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* **42**, 237 (2011), ISSN 1355-2198, <https://doi.org/10.1016/j.shpsb.2011.07.001>.
- A. Fick, *Annalen der Physik und Chemie* **170**, 59 (1855a), <https://doi.org/10.1002/andp.18551700105>.
- A. Fick, *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **10**, 30 (1855b), <https://doi.org/10.1080/14786445508641925>.
- J. Philibert, *Diffusion Fundamentals* **2**, 1 (2005), <https://diffusion-fundamentals.org/journal/2/2005/1.pdf>.
- P. Feyerabend, *Conquest of Abundance: A Tale of Abstraction versus the Richness of Being* (University of Chicago Press, Chicago, IL, USA, 2001), ISBN 978-0-22624534-8, edited by Bert Terpstra, <https://press.uchicago.edu/ucp/books/book/chicago/C/bo3635659.html>.
- M. Srinivas and L. Patnaik, *Computer* **27**, 17 (1994), <https://doi.org/10.1109/2.294849>.
- D. Whitley and A. M. Sutton, in *Handbook of Natural Computing*, edited by G. Rozenberg, T. Bäck, and J. N. Kok (Springer, Berlin, Heidelberg, Germany, 2012), pp. 637–671, [https://doi.org/10.1007/978-3-540-92910-9\\_21](https://doi.org/10.1007/978-3-540-92910-9_21).
- S. Katoch, S. S. Chauhan, and V. Kumar, *Multimedia Tools and Applications* **80**, 8091 (2020), <https://doi.org/10.1007/s11042-020-10139-6>.
- R. M. Smullyan, *Recursion Theory for Metamathematics*, Oxford Logic Guides 22 (Oxford University Press, New York, Oxford, 1993,2020), ISBN 019508232X,9781423734543,9780195082326, <https://doi.org/10.1093/oso/9780195082326.001.0001>.
- R. M. Smullyan, *Diagonalization and Self-Reference*, vol. 27 of *Oxford Logic Guides* (Clarendon Press, New York, Oxford, 1994), ISBN 0198534507,9780198534501.
- R. K. Merton, *Science* **159**, 56 (1968), <https://doi.org/10.1126/science.159.3810.56>.
- R. Larson and B. H. Edwards, *Calculus* (Brooks/Cole Cengage Learning, Belmont, CA, 2010), ninth ed., ISBN 978-0-547-16702-2.
- G. B. Arfken and H. J. Weber, *Mathematical Methods for Physicists* (Elsevier, Oxford, 2005), sixth ed., ISBN 0-12-059876-0;0-12-088584-0.
- G. J. Chaitin, *Algorithmic Information Theory*, Cambridge Tracts in Theoretical Computer Science, Volume 1 (Cambridge University Press, Cambridge, 1987,2003), revised edition ed., <https://doi.org/10.1017/CBO9780511608858>.
- C. Calude, *Information and Randomness—An Algorithmic Perspective* (Springer, Berlin, 2002), 2nd ed., ISBN 978-3-662-04978-5 DOI 10.1007/978-3-662-04978-5,978-3-540-43466-5,978-3-642-07793-7, <https://doi.org/10.1007/978-3-662-04978-5>.
- N. S. Yanofsky, *American Scientist* **104**, 166 (2016), <https://doi.org/10.1511/2016.120.166>.
- N. S. Yanofsky, *Bulletin of Symbolic Logic* **9**, 362 (2003), ISSN 1943-5894, arXiv:math/0305282, <https://doi.org/10.2178/bsl/1058448677>.
- N. S. Yanofsky, *The mind and the limitations of physics* (2019), preprint, , accessed on January 14, 2021, <http://www.sci.brooklyn.cuny.edu/~noson/Mind%20and%20Physics.pdf>.
- E. v. Schweidler, *Über Schwankungen der radioaktiven Umwandlung* (H. Dunod & E. Pinat, Paris, 1906), pp. German part, 1–3, <https://archive.org/details/premiercongrsin03unkngoog>.
- F. S. Exner, *Über Gesetze in Naturwissenschaft und Humanistik: Inaugurationsrede gehalten am 15. Oktober 1908* (Hölder, Ebooks on Demand Universitätsbibliothek Wien, Vienna, 1909, 2016), handle <https://hdl.handle.net/11353/10.451413>, o:451413, Uploaded: 30.08.2016, <http://phaidra.univie.ac.at/o:451413>.
- P. A. Hanle, *Historical Studies in the Physical Sciences* **10**, 225 (1979), <https://doi.org/10.2307/27757391>.
- E. Schrödinger, *Naturwissenschaften (The Science of Nature)* **17**, 9 (1929), <https://doi.org/10.1007/bf01505758>.
- E. Schrödinger, *Science And The Human Temperament* (George Allen & Unwin, 1935), <https://archive.org/details/scienceandthehum029246mbp>.
- H. Hertz, *Prinzipien der Mechanik* (Johann Ambrosius Barth (Arthur Meiner), Leipzig, 1894), mit einem Vorwort von H. von Helmholtz, <https://archive.org/details/dieprinzipiende00hertgoog>.

- H. Hertz, *The principles of mechanics presented in a new form* (MacMillan and Co., Ltd., London and New York, 1899), with a foreword by H. von Helmholtz, translated by D. E. Jones and J. T. Walley, <https://archive.org/details/principlesofmech00hertuoft>.
- G. Mialon, R. Dessì, M. Lomeli, C. Nalmpantis, R. Pasunuru, R. Raileanu, B. Rozière, T. Schick, J. Dwivedi-Yu, A. Celikyilmaz, et al., ArXiv e-prints (2023), arXiv:2302.07842, <https://doi.org/10.48550/arXiv.2302.07842>.
- T. Liu, Y. Jiang, N. Monath, R. Cotterell, and M. Sachan, ArXiv e-prints (2022), arXiv:2210.14698, <https://doi.org/10.48550/arXiv.2210.14698>.
- P. Frank, *Das Kausalgesetz und seine Grenzen* (Springer, Vienna, 1932).
- P. Frank and R. S. Cohen (Editor), *The Law of Causality and its Limits* (Vienna Circle Collection) (Springer, Vienna, 1997), ISBN 0792345517, <https://doi.org/10.1007/978-94-011-5516-8>.
- B. Libet, in *Neurophysiology of Consciousness* (Birkhäuser Boston, 1993), Contemporary Neuroscientists, pp. 269–306, ISBN 978-1-4612-6722-5, [https://doi.org/10.1007/978-1-4612-0355-1\\_16](https://doi.org/10.1007/978-1-4612-0355-1_16).
- J. Olds and P. Milner, *Journal of Comparative and Physiological Psychology* **47**, 419 (1954), <https://doi.org/10.1037/h0058775>.
- P. M. Milner, *Neuroscience & Biobehavioral Reviews* **13**, 61 (1989), [https://doi.org/10.1016/s0149-7634\(89\)80013-2](https://doi.org/10.1016/s0149-7634(89)80013-2).
- C. E. Moan and R. G. Heath, *Journal of Behavior Therapy and Experimental Psychiatry* **3**, 23 (1972), [https://doi.org/10.1016/0005-7916\(72\)90029-8](https://doi.org/10.1016/0005-7916(72)90029-8).
- R. K. Portenoy, J. O. Jarden, J. J. Sidtis, R. B. Lipton, K. M. Foley, and D. A. Rottenberg, *Pain* **27**, 277 (1986), [https://doi.org/10.1016/0304-3959\(86\)90155-7](https://doi.org/10.1016/0304-3959(86)90155-7).
- D. J. Linden, *The Compass of Pleasure: How Our Brains Make Fatty Foods, Orgasm, Exercise, Marijuana, Generosity, Vodka, Learning, and Gambling Feel So Good* (Penguin Books, New York, NY, USA, 2012), ISBN 978-0-14312075-9, <https://www.penguinrandomhouse.com/books/306396/the-compass-of-pleasure-by-david-j-linden/>.
- D. J. Linden, HuffPost (2011), published on July 7, 2011, updated September 6, 2011 accessed March 30, 2023, [https://www.huffpost.com/entry/compass-pleasure\\_b\\_890342](https://www.huffpost.com/entry/compass-pleasure_b_890342).
- D. Z. Lieberman and M. E. Long, *The Molecule of More: How a Single Chemical in Your Brain Drives Love, Sex, and Creativity* (BenBella Books, Dallas, TX, USA, 2018), ISBN 978-1-94688511-1, <https://benbellabooks.com/shop/the-molecule-of-more/>.
- A. Huxley, *The Doors of Perception* (Harper, New York, 1954), [https://archive.org/download/Huxley\\_Aldous\\_-\\_The\\_Doors\\_of\\_Perception](https://archive.org/download/Huxley_Aldous_-_The_Doors_of_Perception).
- R. Descartes, *Descartes: Meditations on First Philosophy. With Selections from the Objections and Replies* (Cambridge University Press, Cambridge, England, UK, 1996), ISBN 978-0-52155818-1, Cambridge Texts in the History of Philosophy, translated and edited by John Cottingham and an introduction by Bernard Williams, <https://www.cambridge.org/catalogue/catalogue.asp?isbn=9780521558181>.
- P. W. Anderson, *Science* **177**, 393 (1972), <https://doi.org/10.1126/science.177.4047.393>.
- J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, et al., *Emergent abilities of Large Language Models* (2022), published in Transactions on Machine Learning Research (TMLR), August 2022, arXiv:2206.07682, <https://doi.org/10.48550/arXiv.2206.07682>.

# Analysing the Conjunction Fallacy as a Fact



Tomas Veloz  and Olha Sobetska

**Abstract** Since the seminal paper by Tversky and Kahneman, the ‘conjunction fallacy’ has been the subject of multiple debates and become a fundamental challenge for cognitive theories in decision-making. In this chapter, we take a rather uncommon perspective on this phenomenon. Instead of trying to explain the nature or causes of the conjunction fallacy (intensional definition), we analyse its range of factual possibilities (extensional definition). We show that the majority of research on the conjunction fallacy, according to our sample of experiments reviewed which covers the literature between 1983 and 2016, has focused on a narrow part of the a priori factual possibilities, implying that explanations of the conjunction fallacy are fundamentally biased by the short scope of possibilities explored. The latter is a rather curious aspect of the research evolution in the conjunction fallacy considering that the very nature of it is motivated by extensional considerations.

**Keywords** Conjunction fallacy · Possible experiences · Factuality · Decision making · Data review · Experimental setting

---

T. Veloz (✉)

Departamento de Matemática, Universidad Tecnológica Metropolitana, Santiago, Chile

Interdisciplinary Foundation for the Development of Science, Technology and Arts, Santiago, Chile

Centre Leo Apostel, Vrije Universiteit Brussel, Brussels, Belgium

e-mail: [tomas.veloz@vub.be](mailto:tomas.veloz@vub.be)

O. Sobetska

Faculty of Social Sciences and Philosophy, Institute of Sociology, University of Leipzig, Leipzig, Germany

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

101

T. Veloz et al. (eds.), *Trends and Challenges in Cognitive Modeling*, STEAM-H:

Science, Technology, Engineering, Agriculture, Mathematics & Health,

[https://doi.org/10.1007/978-3-031-41862-4\\_8](https://doi.org/10.1007/978-3-031-41862-4_8)

## 1 Introduction

The conjunction fallacy (CF) is one of the most important challenges in rational approaches to cognition (Hastie and Dawes 2001; Gilovich et al. 2002). It has been extensively discussed in the cognitive science community (Tversky and Kahneman 1983; Morier and Borgida 1984; Gigerenzer 1996; Tentori et al. 2004; Moro 2009). Tversky and Kahneman introduced this fallacy in their influential ‘Linda story’ experiment (Tversky and Kahneman 1983), where people tend to judge the conjunction of two events  $A$  and  $B$  as more likely than  $A$  or  $B$  separately. For example, people judge ‘Linda is a feminist and a bank teller’ as more likely than ‘Linda is a bank teller’. The CF poses a challenge to classical probability modelling, and its presence has been confirmed in several cognitive experiments involving Linda-like stories (Morier and Borgida 1984; Tentori et al. 2004; Gavanski and Roskos-Ewoldsen 1991; Fisk and Pidgeon 1996; Fisk 2002; Wedell and Moro 2008; Costello 2009; Lu 2015).

Tversky and Kahneman’s developed a research programme to explain this and other fallacies based on ‘individual biases and heuristics’. This work became extremely influential in decision-making theories and encouraged the development of several alternative explanations based on normative and descriptive approaches (Busemeyer et al. 2011; Lu 2015). However, there is no agreement on the ultimate cognitive process that produces these deviations (Shah and Oppenheimer 2008).

It should be noted that the CF requires evidence of no more than three numbers, namely the likelihood estimates of events ‘ $A$ ’, ‘ $B$ ’, and ‘ $A$  and  $B$ ’. These likelihood estimates can in principle be any value between zero and one, where zero means completely unlikely and one means completely likely. The CF is represented as a point in  $[0, 1]^3$ , i.e., a three-dimensional vector  $(P(A), P(B), P(AB))$  with values between 0 and 1 where the rational approach, inherited from Kolmogorovian probability theory or equivalently Fuzzy logic, expects that the third value of such a vector must be smaller than or equal to the other two. Namely, if the likelihood estimate of events ‘ $A$ ’, ‘ $B$ ’, and ‘ $A$  and  $B$ ’ produces a point such that

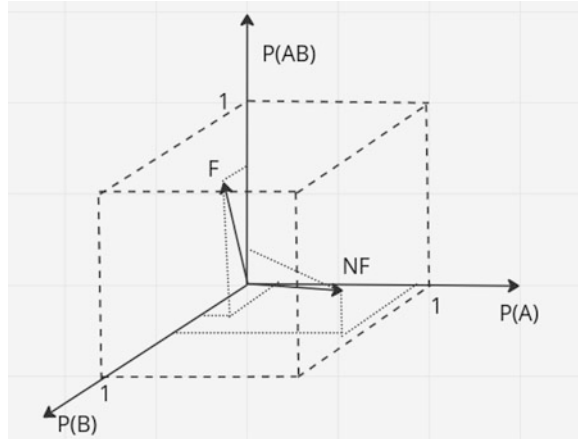
$$P(AB) \leq \min(P(A), P(B)), \quad (1)$$

then we say that the likelihood estimation commits the CF. In Fig. 1 we show the cube formed by  $[0, 1]^3$  and two vectors representing a case where the fallacy is not committed ( $NF$ ) and another vector where the fallacy is committed ( $F$ ).

Research on the CF has steadily developed for exactly four decades, and the seminal article (SA) of Tversky and Kahneman accumulates today more than 6000 citations. Regarding models explaining the CF, Tversky and Kahneman’s model and several other approaches have produced formulas that resemble in some way our rational inequality Eq. 1 but, by introducing other ‘psychological’ parameters, ensure that the third value can be larger than one or two of the other values (Tentori et al. 2013; Boyer-Kassem et al. 2016).



**Fig. 1** Examples of two data points, one not committing the fallacy (*NF*) and another committing it (*F*). Note that for *NF*, the projections of the vector onto the  $P(A)$  and  $P(B)$  axes are larger than the projection onto the axis  $P(AB)$ , while the opposite is true for *F*



Concerning experimental work on the CF, a large number of articles do not provide experimental evidence at the probability judgement level, but instead report the percentage of people ‘committing the fallacy’ (Fiedler 1988). The latter prevents scholars in the community researching CF from comparing or replicating experimental data. Additionally, in several experiments other measures that are related to probability judgements such as likelihood ranking and preference are given (Bar-Hillel 1984; Agnoli and Krantz 1989), while others report two of the three likelihood estimates, preventing a proper comparison across experiments and parameters of alternative explanations (Shafffi et al. 1990).

Due to the latter problems, we will discuss in this chapter the extensional aspect of the CF. Namely, we focus on the different possibilities that can arise from any given point in  $[0, 1]^3$  representing likelihood estimates of  $A$ ,  $B$ , and  $A$  and  $B$  and what can be inferred about the fallacy in the different regions. To do so, we reviewed nearly 4000 articles citing the CF up to 2016. Strikingly, we found that the literature has focused on estimates that cover a narrow part of  $[0, 1]^3$ . To conclude, we will discuss why the latter has happened and how to advance this research programme further.

## 2 The Conjunction Fallacy in a Nutshell

In the field of cognitive psychology, Tversky and Kahneman discovered the CF through an experiment known as the ‘Linda story’ (Tversky and Kahneman 1983). Participants received a questionnaire that featured a story about a woman named Linda who was a single, outspoken, and bright philosophy major. Afterwards, participants were asked to estimate the likelihood of several options, where two of them were (i) Linda is a bank teller or (ii) Linda is a bank teller and is active in the feminist movement. The experiment found that 85% of the respondents chose

option (ii) over option (i) Tversky and Kahneman (1983). This result contradicts the classical Kolmogorovian probability, which predicts that  $P(A \text{ and } B) \leq P(B)$ .

In the SA by Tversky and Kahneman, they performed a number of experiments testing the CF in other Linda-like situations including outcomes of dice rolling, features, or decisions of people in specific situations, potential outcomes of a diagnosis, competition results in a tournament, etc. In their experiments, different types of estimates were considered, including likelihood, probability, ranking, etc. Moreover, different considerations regarding the level of general and specific knowledge of participants were also explored. The important and striking result is that, in all tested situations, CF was observed in a significant part of the studied group. As an explanatory mechanism, they proposed the representativeness heuristic. That is, the event 'A and B' is more representative than B alone (Busemeyer et al. 2011). In the words of Tversky and Kahneman, *The representativeness heuristic favours outcomes that make good stories or good hypotheses* (Tversky and Kahneman 1983). The representativeness heuristic is theoretically developed in the company of various other heuristics based on concepts such as availability, anchoring, causal coherence, etc., which all together form a complex of notions proposing a theory for non-rational decision-making, which proved quite useful and led Kahneman to obtain the Nobel prize in economics in 2002 (Shefrin and Statman 2003).

While a myriad of studies appeared after SA confirming the CF in multiple other situations (Morier and Borgida 1984; Tentori et al. 2004; Gavanski and Roskos-Ewoldsen 1991; Fisk 2002; Wedell and Moro 2008; Costello 2009), a few approaches were proposed to explain the fallacy in a more formal-theory-friendly style. The 'misunderstanding hypothesis' is the most important of such explanations. It suggests that people fail to understand the meaning of sentences and get confused due to linguistic interpretations. For example, the misunderstanding hypothesis has proposed that when we read options 'Linda is a bank teller' and 'Linda is a feminist and a bank teller', we might tend to think that the former option implies 'Linda is not a feminist' (Fiedler 1988). Various mathematical models have tried to explain either the misunderstanding hypothesis or some semantic processing-inspired version of how people deal with likelihood estimates of individual events to create a likelihood estimate of a compound event. However, experimental challenges have also been raised against these models (Tentori et al. 2004; Wedell and Moro 2008), leaving the CF as one of the cornerstones of non-rational decision-making.

Beyond the purely decision-making aspect, the CF has been explored from multiple other perspectives, such as neuropsychology, by measuring neural mechanisms that might underlie the decision-making process leading to a fallacy (Gardner 2019), by developing cross-cultural studies to understand whether or not the CF is somehow culture dependent (Lee et al. 2018), by detecting analogous fallacies for the case of disjunction and other logical connectives (Morier and Borgida 1984; Busemeyer and Bruza 2012), among others (for a review, see Moro 2009).

### 3 Is the Conjunction Fallacy a Fact?

We recall that the CF is represented as a point in  $[0, 1]^3$ , i.e., a three-dimensional vector  $(P(A), P(B), P(AB))$  with values between 0 and 1 (see Fig. 1). Therefore, any point in  $[0, 1]^3$  should *a priori* be a possible fact observed in an experimental setting.

Probability theory teaches us that if we represent  $A$  and  $B$  as sets and  $P(A)$  and  $P(B)$  reflect the measure of these sets, then not every point in  $[0, 1]^3$  can be a valid fact because the nature of set operations prevents  $P(AB)$  from being larger than both  $P(A)$  and  $P(B)$ . The above-mentioned assumptions are quite reasonable because every macroscopic physical event we observe obeys such rules. Indeed, let us consider a typical example of probability theory to illustrate this idea. Consider an urn with balls that can be either  $A = \text{'red'}$  or  $A' = \text{'not red'}$  and can be made either  $B = \text{'wooden'}$  or  $B' = \text{'not wooden'}$ . It is clear that the probability of extracting a ball that is 'red and wooden', i.e.,  $P(AB)$  is smaller than extracting a ball that is 'red' ( $P(A)$ ) or 'wooden' ( $P(B)$ ). Additionally, other probabilistic inequalities can be assumed to hold without any risk, such as  $1 + P(AB) - P(A) - P(B) \geq 0$ , and  $P(A) + P(B) - P(A \cup B) \leq 1$ , where  $P(A \cup B)$  is the probability that the ball is 'red or wooden'.

Inspired by this reasoning, George Boole constrained the possible relative frequencies of experimental settings and formulated, for the first time in history, a set of probabilistic inequalities derived from set-theoretical considerations, and called them 'conditions of possible experience' (Boole 1854). The conditions of possible experience underlie all reasoning about our macroscopic physical reality and are fundamental in areas such as statistical mechanics and electrodynamics (Cochrane 2006). However, it is today well known that the conditions of possible experience fail for microscopic physical systems, specifically for those where quantum theory operates. The double-slit experiment is the canonical piece of evidence illustrating that quantum systems are 'not urns with balls' but something much stranger (Wooters and Zurek 1979). In the experiment, a beam of particles, such as electrons, is fired at a barrier with two slits. On the other side of the barrier, a detector is placed to detect where the particles land. The particles typically create an interference pattern on the detector screen, as if the particles were behaving as waves. The latter defies our intuition of particles behaving as 'balls passing through the slit' and seems to suggest that each particle 'passes through both slits simultaneously' as if particles were waves. This phenomenon is known as the wave-particle duality of matter, and it remains a cornerstone of our understanding of the behaviour of matter and energy at the quantum level.

Itamar Pitowski, who was a prominent philosopher of quantum physics, studied why the conditions of possible experience are not respected by quantum systems. He came up with a list of reasons that could cause violations to them (Pitowsky 1994):

1. Failure of randomness: Obtaining a bad approximation of probabilities due to not taking a large enough sample of measurements.

2. Measurement biases: The method of experimentation produces disturbances in the object we measure with.
3. No distribution: The phenomena under observation do not possess a well-defined distribution of properties to be measured or not even well-defined properties to think of a distribution in the first place.
4. Mathematical oddities: Sets representing events might not be measurable, implying that probabilities will not be uniquely defined.

Pitowski explains that we do not know in principle what of these reasons could be in operation as a source of violation of the conditions of possible experience in quantum physics, though the first reason can be discarded due to the tremendous advances in experimental microphysics (Schirber 2022), and that reason (4) focuses on the mathematical structure of events to explain the violation of probabilities rather than on the very factual nature of the events and their observation. Hence, we will not consider these two cases in our analysis. For cases (2) and (3), it is important to mention that it is essential that measurements cannot be made simultaneously on the ‘same’ sample. This means that either the sample where we measure changes with the measurement or it becomes inaccessible. Hence, we need to take a new sample for every next measurement. In fact, if all measurements can be made on the same sample, then the conditions of possible experience will always hold. Indeed, for macroscopic physics, we are able to create ‘ensembles’ which correspond to a large number of equivalent systems to which we can subject experimental measurements and thus calculate relative frequencies in a rather controlled way. The latter is exactly the reason why the conditions of possible experience hold in classical physics and why we believe they are so important to explain reality.

However, when we move to domains where multiple measurements cannot be made on the same sample, such as in quantum physics or in psychology, and where we see highly contextual situations that suggest intrinsic nondeterministic changes of state (physical or psychological, respectively), we shall carefully consider whether the conditions of possible experience are truly a constraint of what we are studying. Moreover, if they are not the right constraint, it is a scientific duty to explain what kind of restrictions could be in operation.

## 4 How the Conjunction Fallacy Data Looks?

We performed a literature review considering all articles citing SA between 1983 (year of publication) and 2016. Given the importance of the factual analysis of the previous section, we would have expected that studies on the CF generally report statistics of the values  $P(A)$ ,  $P(B)$ , and  $P(AB)$  obtained from their experiments. However, we observed that studies reporting such data are rare. In Table 1 we show the number of articles citing SA, the number of works that produced experimental data on probability estimates, and the number of articles that additionally reported statistics of the values  $P(A)$ ,  $P(B)$ , and  $P(AB)$ .

**Table 1** Summary of CF literature and data

Number of Number of	Experiments on probability estimates	Well reported data
4080	272	37

**Table 2** Number of data points reported on each sub-volume of  $[0, 1]^3$

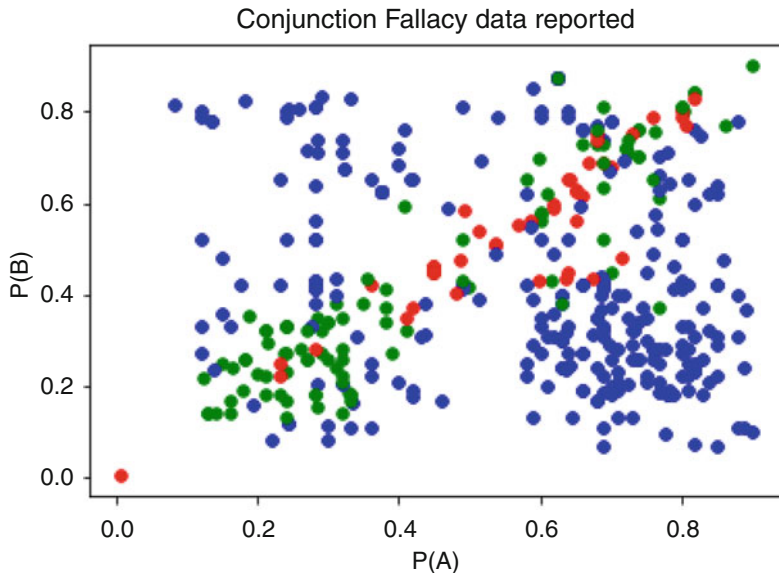
$P(A)$	$P(B)$	$P(AB)$	Frequency
[0, 0.5]	[0, 0.5]	[0, 0.5]	115
[0, 0.5]	[0, 0.5]	(0.5, 1]	2
[0, 0.5]	(0.5, 1]	[0, 0.5]	37
[0, 0.5]	(0.5, 1]	[0.5, 1]	10
(0.5, 1]	[0, 0.5]	[0, 0.5]	106
(0.5, 1]	[0, 0.5]	(0.5, 1]	41
(0.5, 1]	(0.5, 1]	[0, 0.5]	3
(0.5, 1]	(0.5, 1]	[0.5, 1]	102

A closer inspection to the reported data shows that the full extent of possible experiences, represented by  $[0, 1]^3$ , has only been narrowly covered by current experiments. We compiled the complete set of data points (416 points) reported by the articles in the third column of Table 1 in order to see how well distributed is the data across the set of possible points that can be obtained. We had to adapt some of these data points to the interval  $[0, 1]$  when they were reported on a different scale. In Table 2 we see the number of data points reported by experiments in different sub-volumes of  $[0, 1]^3$ .

We observe that, at our volume resolution, all sub-volumes have at least 2 data points. However, we see that the distribution is far from homogenous, implying that experiments have not properly sampled the space of possibilities. In particular, we observe that the second case is the most underrepresented case, which consists of  $P(A)$  and  $P(B)$  being smaller than 0.5 while  $P(AB)$  is larger than 0.5. This situation is an example of what has been called a ‘double overextension’ in categorization (Hampton 1988) and refers to the conjunctive category being more typical (in the CF case, more likely) than the two former concepts (in the CF case, former events). The other most underrepresented case is the seventh case which reflects a possible experience in Boole’s sense. This also suggests that there has not been a systematic study in finding the frequency of CF in real situations.

To deepen a bit our analysis, we show in Fig. 2 the 416 data points, plotting  $P(A)$  versus  $P(B)$  and labelling the colour of the point based on the value of  $P(AB)$  in relation to the other two. That is, for  $P(AB) \leq \min(P(A), P(B))$  the point is coloured green (no CF), for  $\min(P(A), P(B)) \leq P(AB) \leq \max(P(A), P(B))$  the point is coloured blue (single CF violation), and for  $\max(P(A), P(B)) \leq P(AB)$  the point is coloured red (double CF violation).

Various things are worth noticing here. First, almost no data is reported for events with a probability smaller than 0.1. This means that the phenomenology of very unlikely events is under-investigated. Second, although points are relatively well distributed (with the exception of the very unlikely events mentioned before), there



**Fig. 2** CF data reported and coloured by the extent fallacy violation. Green means no violation, blue means violation on one of the probabilities, and red means violation on two of the probabilities

seems to be an area around  $[0.5, 0.6] \times [0, 1]$ , which is also underrepresented. Interestingly, this area corresponds to events where there is very high uncertainty on  $A$ , and we observe from the few points available in that area that the three colours are more or less equally present. The latter is consistent with the fact that the more uncertainty about the events we have, the more diverse our thinking can be, which in turn is consistent with the fact that intuition operates more strongly in situations of uncertainty. Third, we see that both the green and red points are mostly scattered around the diagonal. This implies that green points are highly underrepresented, as we can always create macroscopic physical situations where  $P(AB)$  is anywhere below both  $P(A)$  and  $P(B)$  and that ‘red’ deviations have probably not been explored enough.

## 5 Conclusion: Do We Need Factual Paradigms?

In this chapter, we aim to illustrate that experimental research in CF has not given sufficient attention to what could be the most interesting question on the topic. Namely, what are our possible conjunctive experiences?

In particular, we notice that research has focused on testing the existence of the fallacy under several different conditions, while identifying how often this happens,

or to what extent it happens, has been left as secondary topics. By separating the data points reported in the literature into sub-volumes we see that the single fallacy, i.e., when  $\min(P(A), P(B)) < P(AB) \leq \max(P(A), P(B))$ , is relatively well explored, while the no fallacy ( $P(AB) \leq \min(P(A), P(B))$ ) and the double fallacy ( $\max(P(A), P(B)) < P(AB)$ ) are under-explored (see Table 2).

We believe that this lack of attention is linked to the rationalist approach to cognition. Namely, most authors have tried to explain ‘why’ people commit the fallacy by developing theories of how we interpret language and have focused on the commitment frequency of the fallacy in a given situation rather than on the extent at which it occurs. However, conjunction fallacy explanations assume that if we would ‘interpret the experiment text literally’, we would not commit the fallacy. This is coherent with the fact that this phenomenon is called a fallacy, and it is not instead conceived as an ‘equally valid’ form of cognition as rationality.

From the sampled literature we see that potentially interesting cases with strong double fallacy such as  $P(A) = 0.1$ ,  $P(B) = 0.1$  and  $P(AB) = 0.5$  are either impossible or have not been explored. We call attention to this kind of situation because it exemplifies the extreme of our not rational cognition. For example, consider a typical romantic-story-like situation where two people meet by accident (in a train station, for example) and must spend some little time together in the cold to catch their trains, and they fall in love in that short moment. We can clearly see that the chances that for two people ‘their trains failing simultaneously’ and ‘falling in love in a small chat’ can be rendered very small, but their conjunction is commonplace in our romantic culture, reflecting that we do believe that those events are quite possible, despite the rational fact that they are simply the conjunction of two unlikely events.

Multiple ideas have been developed over the course of the previous century trying to face the domain of what we rationally believe as impossible. One of the most interesting of such attempts is the concept of synchronicity, worked out together between Pauli and Jung, prominent figures in quantum theory and psychology, respectively (Lindorff 1995). They analysed the idea that events, although not causally connected, can still occur simultaneously in a meaningful way. In other words, two or more seemingly unrelated events happen at the same time, and yet they have a non-causal underlying connection. When these two non-causally connected events are both unlikely is when synchronicity shows its most interesting power and it is where our culture has recognized in multiple stories, in many cases related to love, destiny, and other concepts where science has not been able to penetrate. Interestingly, synchronicity has been proposed to bridge the gap between quantum theory and psychology at a more fundamental level of reality where both kinds of events, psychological and physical, share the same nature (Cambray 2009). Along this same line, other researchers have attempted to prescind from the idea of theoretical constructions implying divisions between physics and psyche and have attempted to develop a purely factual construction of reality, which cannot be disentangled from our perception. As such these attempts start directly from perceived reality and build factual probabilistic laws based on our ability to replicate observational procedures (Mugur-Schächter 2002). To conclude, we propose that

the conjunction fallacy must be analysed under these perhaps more speculative lines to advance a deeper understanding of what it means for our cognition. Specifically, experiments exploring strong double fallacies could provide a much more solid ground to prove the rationalistic-inspired approaches incomplete. Firstly, if we had more properly designed studies, in line with our discussion of constructing CF test questions in Sect. 4, we could say more clearly whether such double fallacies are truly unexplored or truly unlikely. As it stands, most of the studies we have analysed do not allow us to derive a logical and methodological conclusion on this point. Secondly, we would have a more rigorous argument to talk about our irrationality and that this irrationality is not fallacious, but normal for the behaviour of our brain, and thus factual approaches for constructing scientific knowledge such as Mugur-Schächter (2002) might prove especially useful and might invite to explore less traditional mathematical structures. Concerning psychological experiments, with the proper design of the task, we would not lose the respondents' focus by dispersing them into ratings by considering the task in a too general way. On the contrary, an assignment in which respondents are asked to estimate the likelihood of each event (in the context of Linda's story or similar) allows them to analyse these events separately and make a more accurate estimate, which will give us prospectively more methodological power to justify the radically non-intuitive nature of CF.

**Acknowledgments** T.V was funded by the John Templeton Foundation as part of the project "The Origins of Goal-Directedness" (grant ID61733).

## References

- Franca Agnoli and David H Krantz. Suppressing natural heuristics by formal instruction: The case of the conjunction fallacy. *Cognitive Psychology*, 21(4):515–550, 1989.
- Maya Bar-Hillel. Representativeness and fallacies of probability judgment. *Acta Psychologica*, 55(2):91–107, 1984.
- George Boole. *An investigation of the laws of thought: on which are founded the mathematical theories of logic and probabilities*, volume 2. Walton and Maberly, 1854.
- T. Boyer-Kassem, S. Duchêne, and E. Guerci. Quantum-like models cannot account for the conjunction fallacy. *Theory and Decision*, 2016.
- J. R. Busemeyer, E. M. Pothos, R. Franco, and J. S. Trueblood. A quantum theoretical explanation for probability judgment errors. *Psychological Review*, 118:193–218, 2011.
- JR Busemeyer and PD Bruza. *Quantum Models of Cognition and Decision*. Cambridge University Press, Cambridge, 2012.
- Joseph Cambrey. *Synchronicity: Nature and psyche in an interconnected universe*, volume 15. Texas A&M University Press, 2009.
- Linda Cochrane. *Kant, Newton, and the conditions of possible experience*. PhD thesis, Concordia University, 2006.
- F.J. Costello. How probability theory explains the conjunction fallacy. *Journal of Behavioral Decision Making*, 22:213–234, 2009.
- Klaus Fiedler. The dependence of the conjunction fallacy on subtle linguistic factors. *Psychological research*, 50(2):123–129, 1988.
- J.E. Fisk. Judgments under uncertainty: Representativeness or potential surprise? *British Journal of Psychology*, 93:431–449, 2002.



- J.E. Fisk and N. Pidgeon. Component probabilities and the conjunction fallacy: Resolving signed summation and the low component model in a contingent approach. *Acta Psychologica*, 94:11–20, 1996.
- Justin L Gardner. Optimality and heuristics in perceptual neuroscience. *Nature neuroscience*, 22(4):514–523, 2019.
- I. Gavanski and D.R. Roskos-Ewoldsen. Representativeness and conjoint probability. *Journal of Personality and Social Psychology*, 61:181–194, 1991.
- Gerd Gigerenzer. On narrow norms and vague heuristics: A reply to Kahneman and Tversky. *Psychological Review*, 103:592–596, 1996.
- T. Gilovich, D. Griffin, and D. Kahneman. *Heuristics and Biases: The Psychology of Intuitive Judgement*. Cambridge University Press, Cambridge, 2002.
- J. A. Hampton. Overextension of conjunctive concepts: Evidence for a unitary model for concept typicality and class inclusion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14:12–32, 1988.
- D. Hastie and R. Dawes. *Rational Choice in an Uncertain World: The Psychology of Judgement and Decision Making*. Sage Publications, Thousand Oaks, 2001.
- Sunghee Lee, Florian Keusch, Norbert Schwarz, Mingnan Liu, and Z Tuba Suzer-Gurtekin. Cross-cultural comparability of response patterns of subjective probability questions. *Advances in comparative survey methods: Multinational, multiregional, and multicultural contexts*, pages 457–476, 2018.
- David Lindorff. Psyche, matter and synchronicity: A collaboration between CG Jung and Wolfgang Pauli. *Journal of Analytical Psychology*, 40(4):571–586, 1995.
- Yang Lu. The conjunction and disjunction fallacies: Explanations of the Linda problem by the equate-to-differentiate model. *Integrative Psychological and Behavioral Science*, pages 1–25, 2015.
- Dean Morier and Eugene Borgida. The conjunction fallacy: A task specific phenomenon? *Personality and Social Psychology Bulletin*, 10:243–252, 1984.
- Ren’e Moro. On the nature of the conjunction fallacy. *Synthese*, 171:1–24, 2009.
- Mioara Mugur-Schächter. Objectivity and descriptonal relativities. *Foundations of Science*, 7:73–180, 2002.
- Itamar Pitowsky. George Boole’s ‘conditions of possible experience’ and the quantum puzzle. *The British Journal for the Philosophy of Science*, 45(1):95–125, 1994.
- Michael Schirber. Nobel prize: Quantum entanglement unveiled. *Physics*, 15:153, 2022.
- Eldar B Shafffi, Edward E Smith, and Daniel N Osherson. Typicality and reasoning fallacies. *Memory & Cognition*, 18(3):229–239, 1990.
- A.K. Shah and D.M. Oppenheimer. Heuristics made easy: An effort-reduction framework. *Psychological Bulletin*, 134:207–222, 2008.
- Hersh Shefrin and Meir Statman. The contributions of Daniel Kahneman and Amos Tversky. *The Journal of Behavioral Finance*, 4(2):54–58, 2003.
- Katya Tentori, Nicolao Bonini, and Daniel Osherson. The conjunction fallacy: A misunderstanding about conjunction. *Cognitive Science*, 28:467–477, 2004.
- Katya Tentori, Vincenzo Crupi, and Sven Russo. On the determinants of the conjunction fallacy: Probability versus inductive confirmation. *Journal of Experimental Psychology: General*, 142(1):235, 2013.
- A. Tversky and D. Kahneman. Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90:293–315, 1983.
- D.H. Wedell and R. Moro. Testing boundary conditions for the conjunction fallacy: Effects of response mode, conceptual focus, and problem type. *Cognition*, 107:129–140, 2008.
- William K Wootters and Wojciech H Zurek. Complementarity in the double-slit experiment: Quantum nonseparability and a quantitative statement of Bohr’s principle. *Physical Review D*, 19(2):473, 1979.

# Yes Ghosts, No Unicorns: Quantum Modeling and Causality in Physics and Beyond



Kathryn Schaffer and Gabriela Barreto Lemos

**Abstract** Entanglement is often considered a signature of “true quantumness.” But what counts as “true quantum entanglement?” Historically, physicists have relied on statistical tests—Bell tests—as a quantum-classical decider: entanglement that shows violations of Bell inequalities is taken to show non-classical correlation. But, meanwhile, claims of Bell-inequality violations with classical systems have proliferated, in physics and beyond. The situation is confusing. This chapter takes some steps toward clarity. Drawing from examples in physics, we urge caution in cross-disciplinary modeling comparisons and illustrate the kind of explanatory causal reasoning that underlies Bell tests. We then highlight the recent application of Causal Analysis to Bell tests to emphasize the role of “unicorn-like” fine-tuning. Finally, we discuss recent work in classical optics that shows that Bell inequalities need to be re-derived and interpreted with assumptions appropriate to the measurement scenario. While we do personally believe that quantum physics exhibits a type of spookiness (a quantum-physics-specific “ghost”), the more important point of this chapter is to argue that Bell inequalities are not portable: their bounds need to be re-derived and interpreted appropriately for each case.

**Keywords** Quantum modeling · Bell test · Causality · Entanglement · Contextuality · fine-tuning

## 1 Introduction

This chapter is part of an ongoing conversation between its authors exploring two beliefs we share: (a) that laboratory experiments in quantum physics reveal at

---

K. Schaffer (✉)  
School of the Art Institute of Chicago, Chicago, IL, USA  
e-mail: [kschaf2@artic.edu](mailto:kschaf2@artic.edu)

G. B. Lemos  
Instituto de Física, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023  
T. Veloz et al. (eds.), *Trends and Challenges in Cognitive Modeling*, STEAM-H:  
Science, Technology, Engineering, Agriculture, Mathematics & Health,  
[https://doi.org/10.1007/978-3-031-41862-4\\_9](https://doi.org/10.1007/978-3-031-41862-4_9)

least one distinct *spookiness* that remains unexplained and (b) that this distinct spookiness, because of how it uniquely arises in quantum physics, is not present in other disciplinary contexts that use similar mathematical frameworks for quantitative modeling. To defend claim (a), we will discuss a class of experiments in quantum physics referred to as “Bell tests,” the interpretation of which remains hotly contentious within physics. These experiments involve quantum entanglement and famously exemplify what Einstein called “spooky action at a distance.” The “yes ghosts” in the title of this chapter reflects, metaphorically, our commitment that Bell tests in quantum physics do reveal something distinctly spooky. Work recasting Bell tests in the language of Causal Analysis (Wood and Spekkens 2015; Cavalcanti 2018; Pearl and Cavalcanti 2021), as well as recent work analyzing Bell tests with classical light in the broader framework of Contextuality tests (Markiewicz et al. 2019), shows that the spookiness is not merely about the claimed “action at a distance” in entanglement scenarios. It is also more fundamental to the structure of causality in quantum physics. If fine-tuned, conspiratorial, “unicorn-like” causal mechanisms are forbidden, then Bell tests in quantum physics pose a puzzling contradiction.<sup>1</sup>

For this chapter, which is addressed to an interdisciplinary audience, our primary goal is to argue for claim (b). A consensus has not yet emerged about what is going on in Bell test experiments in physics, and not all physicists agree with us that there is something spooky involved. Nevertheless, the point of contention does not arise in other domains that employ similar mathematical modeling. We will argue this first by discussing features of mathematical modeling practices, emphasizing that the relevant causal mechanisms in a modeling situation depend on the specific measurements used to produce the data, and second by discussing classical-optics scenarios from physics.

Quantum-inspired modeling practices now appear in fields as diverse as cognition, economics, and language modeling (see Pothos and Bussemeyer 2022; Lee 2020; Surov et al. 2021 and the references therein for some examples). Entanglement models are often central in such efforts. As a contribution to this growing discourse, we explain some reasons why importing the mathematics of quantum physics—the mathematics of entanglement and Bell inequalities, in particular—does not necessarily mean importing its interpretational problems.

## 2 Disciplinary and Interdisciplinary Considerations

The interdisciplinary context of this chapter invites a few prefatory comments. We underscore that we are focused on mathematical modeling practices and their causal assumptions, not on stances toward philosophy or worldview. It is part of the

---

<sup>1</sup> In the workshop preceding this volume, one of us (KS) used the phrase “no unicorns” to say that quantum physics does not provide magical ways around ordinary physics. For this chapter we use the unicorn metaphor for something more technical. The connecting thread between both uses is respect for the explanatory success of experimental physics.

culture around quantum physics to call it strange, but such judgments are a matter of opinion. It is a subjective question whether the wave-particle duality, observer-dependence, and uncertainty in quantum physics seem more or less intuitive than the atomism, determinism, and mechanism associated with classical physics. The word “spooky” may be subjective in the same way, but we are using it to refer to something specific and technical, not a ghost-filled worldview.

As “quantum modeling” gains popularity in non-physics fields, the philosophical implications of modeling choices sometimes take on an importance that is unfamiliar to physicists (Schaffer and Barreto Lemos 2021). Thus, for example, we have heard quantum modeling characterized as a revolutionary approach meant to unseat classical paradigms. But a physicist’s library will have textbooks that separately discuss classical mechanics, thermodynamics, quantum physics, classical electromagnetism, relativity, and so on. Data analysis for an experiment can sample from across the bookshelf without needing to commit to just one, or needing all philosophical conflicts to be resolved. In research physics, we are not making an either-or selection of a classical or quantum approach when we model data. In pursuit of explanations that make sense, classical modeling provides appropriate and self-consistent explanations for some measurements (or parts thereof), and quantum modeling does so for others.

Within quantum physics (the subfield of physics that focuses on testing core predictions of quantum theory), there are a few specific experimental results, e.g., Bell tests with quantum entangled particles (Giustina et al. 2015; Hensen et al. 2015; Shalm et al. 2015), that exhibit challenges to the expectation of self-consistency and making-sense. The subject of this chapter is those cases that define the discipline-specific, as-yet-unresolved quantum spookiness. This is an important point to make because Bell-like scenarios have also been explored in classical physics (Borges et al. 2010; Qian et al. 2015; Goldin et al. 2010; Frustaglia et al. 2016; Li et al. 2018) contexts as well as non-physics contexts. The mystique of quantum strangeness, and the desire for proof of “true quantumness,” can cloud discernment of important context-dependent differences in all of these cases.

We argue that Bell tests in quantum physics are distinct. Modeling is not a worldview choice, but it is not just the choice of a set of equations, either. To be explanatory, a modeling practice must embed measurement-specific assumptions about allowable causal processes. Context-dependent and disciplinary differences in causal assumptions are central to sense-making. Bell tests are one example of this truth, but in general, causal mechanisms are important differences in cross-disciplinary modeling comparisons.

### 3 What Counts as Quantum Modeling?

The phrases “quantum model” and “quantum modeling” circulate in non-physics contexts, but without a universal consensus on their scope of meaning. In this section we share some perspectives from physics.

We work with an operational definition of mathematical modeling as a research practice that associates mathematical structures with analogous relationships among measurable quantities. This definition emphasizes both that mathematical modeling involves a form of metaphor (tracing structural analogies between mathematics and a real-world phenomenon) and that we should think of modeling as a verb, an activity embedded in a disciplinary research context that produces peer-reviewed, published results. The criteria for success vary. In some modeling contexts, identifying a structural similarity between an equation and a set of measurements may be enough to constitute success. In others, the modeling effort is not considered successful unless it yields predictions for novel measurements, answers “why” questions, meets goodness-of-fit standards, or otherwise fits into an explanatory sense-making framework. In other words, the fact that a mathematical metaphor exists does not necessarily determine what it means, nor whether it is any good.

The word “model” is a can of worms. In practice, it often refers to some equation(s) used in a modeling effort. But even to draw structural correspondences—to make a mathematical metaphor—we need context-dependent specifications of how symbols on a page relate to measurements in the world. The word “quantum” is also a can of worms. Equations have no allegiance to any discipline. A certain formalism may be historically associated with quantum physics, but the label “quantum” is a convention that carries no intrinsic meaning nor well-defined scope of application. Is it only a quantum model if the equations are applied in physics? Does only a single version of the formalism count, or does the term apply to a class of related probabilistic models? Is every part of the formalism equally “quantum,” even what is shared with models in classical physics (e.g., sinusoidal waves)?

There is a body of knowledge associated with the century-plus success of quantum physics. This knowledge, though it was generated through active modeling, now has a relatively static core: equations in textbooks that have not changed for many decades, a well-defined set of corresponding experiments that are now primarily demonstrative, not actively scrutinized. Most applications of this static body of knowledge do not function as critical tests. That is to say, most of the ways physicists use quantum physics knowledge are not aimed toward falsifying it or expanding it. The research discipline known as quantum physics *is* principally aimed toward those sorts of tests. Can experiments interrogate the textbook formalism in new contexts? What can we reveal through mathematical reformulations? How do we test questions in quantum interpretation? With questions like these, the discipline of quantum physics, over time, expands beyond what is already in the textbooks. In the ambiguity of unresolved open questions, boundary-defining vocabulary questions (“what counts as quantum”) may be permanently premature. There is a well-defined canon of established facts from the past, but we do not know what will enter that canon in the future.

There are also lessons to draw from quantum-related modeling practices in physics, but beyond the subdiscipline of quantum physics. Consider a SQUID (Superconducting QUantum Interference Device). A SQUID is a macroscopic-scale device that leverages low-temperature material properties to realize quantum tunneling and interference for magnetic flux detection. To explain *why* a SQUID

enables single-quantum flux sensitivity, textbook quantum formalism is involved. However researchers who use SQUID circuits (such as astrophysicists using them in a detector system) can treat them as “black box” circuit components. Even though a SQUID is based on quantum tunneling and interference, papers that model SQUID-based detector systems do not need to discuss quantum physics. No obvious “quantum modeling” is necessary (see, for example, Montgomery et al. 2020).<sup>2</sup>

A contrasting example is research applying quantum formalism to phenomena in classical optics, e.g., (Stoler 1981; Klyshko 1988; Spreeuw 1998; Simon and Agarwal 2000). In this case, physicists use mathematics sourced from quantum textbooks, but the experimental systems are causally governed by classical electrodynamics. This has spurred a vigorous debate in physics about vocabulary (Karimi and Boyd 2015). If a classical optics scenario can be described through the same mathematics as quantum entanglement, is it appropriate to label such a phenomenon “classical entanglement”? Or is that a contradiction in terms, because entanglement means something special to the quantum physics context? Many physicists would say that “classical entanglement” has nothing to do with quantum physics, but SQUIDs do. Such a judgment is not about the how, but about the why. Modeling in physics requires more than a structural metaphor because it engages causal explanations.

These examples also show that size scale does not determine whether quantum physics might be relevant. SQUIDs and the circuits that use them are macroscopic, and so are classical optics systems. As such, these examples can help to inform encounters with “quantum modeling” in non-physics disciplines that are also dealing with macroscopic physical systems. For example, consider quantum modeling of phenomena associated with brains or cognition. How might quantum models apply? In multiple ways, that need not relate to one another. On the scale of neurons or smaller, plausibly some brain components could play a role analogous to SQUIDs or other quantum-based circuit components, as macroscopic physical systems whose “how” is linked to a quantum-physics “why.” Once their behavior is understood, modeling those components in context is likely to be similar to the SQUID case: the intrinsically quantum processes can be treated as a black box within a whole-system model.

It is also plausible that some structures in quantum formalism could make good metaphors for whole-brain phenomena, macroscopic human behavior, and some data sets involving language and cognition, e.g., (Pothos and Busemeyer 2022). In such cases, quantum formalism can apply to the “how,” without any connection to a quantum-physics “why,” as with classical entanglement in optics.

Both possibilities could be simultaneously true (quantum processes mattering in neurons as well as quantum formalism applying to whole-brain phenomena) with

---

<sup>2</sup> There are more mundane examples of “essentially quantum devices,” such as transistors. Arguably quantum processes—e.g., those that enable chemistry—are ubiquitous. Not all quantum effects are equally possible in everyday conditions, though. We discuss SQUIDs as an example of a quantum device because they require special conditions. This is also likely to be true for, e.g., entanglement-based devices.

absolutely no link nor meaning across modeling contexts. This relates to a general point about modeling and mereology (the study of part-whole relationships): the kinds of structures we can model mathematically do not generically translate across scale, from part to whole or vice versa. Neither do the causal reasons for them.

To summarize, both “quantum” and “model” are slippery terms. Modeling, as a practice, involves mathematical metaphors for empirical relationships. Such metaphors may be portable from one context to another, but they do not translate trivially across mereological scale. Nor does it mean anything if a similar model works in two different contexts or across scales. Finally, modeling practices vary in their aims. In physics, modeling normally seeks to answer “why” questions beyond the “how” of a phenomenon. This does not mean resolving all of the philosophy. What it means is that modeling practices in physics aim to explain phenomena in terms of situation-specific causal mechanisms.

## 4 Sense-Making Is More than Metaphor

More matters in mathematical modeling than the structure of the equation(s) used to fit the data.

To explore this, consider a linear model. The conventional formula for a straight line is  $y = mx + b$ , where parameter  $m$  sets the slope of the line, and  $b$  sets the value of  $y$  when  $x = 0$ . Such a metaphor has many applications. Variables  $x$  and  $y$  can represent displacement in physical space, such that the line describes a path. With  $y$  relating to space and  $x$  to time, the model can describe linear motion. The model also works in contexts that make no explicit reference to space or time, e.g.,  $x$  could represent the number of identical items in a shopping cart and  $y$  their total cost. We could even get creative and characterize mood in proportion to sunny weather.

While these scenarios share a structural similarity, there are context-dependent dissimilarities too. Variables can be continuous or discrete, bounded or unbounded, or have other constraints. In the cost-items case, no values for  $x$  or  $y$  should be negative; in the path-through-space case, they could be. What is important about the line, as a model, also differs in each case. Explanatory modeling involves evaluating alternatives; the line is thus conceptually embedded in a mathematical space of options that is case-specific. A cost-items scenario might allow sharp discontinuities (e.g., bulk discounts). Such discontinuities would be impossible in linear motion, but a friction term might appear. Expectations of monotonicity, continuity, and single-valuedness are situation-specific.

In research practices that use mathematical modeling for explanatory reasoning, most of the specificity relates to the data. Measurements require apparatuses. Data sets can have mistakes. Stochasticity matters. We cannot make sense of data without detailed knowledge of how all of this works. Such knowledge establishes assumptions about possible, impossible, plausible, and implausible causal mechanisms for features that might appear in a data set.

It is never part of a student lab report in a physics class to allow that unicorns may have secretly manipulated the outcome of a measurement for their pleasure. Explore this by imagining a kinematics lab in university physics. Each station in the room is equipped with an air track that allows approximately frictionless one-dimensional motion of an accompanying “car.” When the student flips a switch, a spring releases the car, allowing it to drift across the track at approximately constant velocity. The same switch restarts a clock. Each track has a set of movable sensors that trigger when the car passes, recording elapsed time. The location of each sensor is measured by eye, using a ruler. A single experimental iteration, or run, involves a student placing the sensors, releasing the car, and collecting elapsed time and corresponding distance measurements for the car’s motion.

A professor tells students that they can work individually or collaborate to take data from ten runs of the experiment, varying sensor positions. They are to tabulate and graph the data and then perform a linear fit to estimate the average velocity of the car and its standard deviation. The professor leaves them under the supervision of a Teacher Assistant (TA), returning to grade the papers later.

What criteria will she use to grade the papers? Well, the first thing she checks is whether the students performed the linear fit properly. Indeed, the TA must have helped: all students have correct mathematics applying a linear fit and estimating average velocities. Do they all get good grades? No. Clearly, more was going on, since many of the graphs look quite different from one another. Some of the fits are terrible. Does she award grades based on the apparent goodness-of-fit? Also no. This is experimental physics. To evaluate the lab reports she has to look at the data and use knowledge of the ways it might have been caused.

First she considers some of the papers with visibly poor fits. A few students apparently had equipment problems or made mistakes, which she guesses by noticing some unphysical data patterns. With one, the graph suggests a significant friction effect. Given the context, this is plausible. She marks off points, highlighting the issues students should have noticed and attempted to explain.

In one lab report, the data has such a large scatter that position and time values look barely correlated. How could this happen, given the apparatus? It suggests a serious data-taking problem. But then, the professor notices that the same data is shared by five students. While the reported data is inconsistent with the expected behavior of one air track, it is perfectly consistent with each student in the group performing two runs on a separate copy of the apparatus. The air track setups vary enough that each produces a different velocity and time-offset ( $m$  and  $b$  in the formula for a line, respectively). The combined data table does not test the behavior of one air track; it tests the behavior of a collection.

Papers that show good fits also deserve attention. Just because the graph looks reasonable does not mean it is without error. She checks for things like unphysical data values. In one paper, the fit is *too* good. The times from trial to trial are all the same, and the deviations in positions are small. This student gets a zero; the results were clearly fabricated.

The professor then comes across a paper with a unicorn drawn next to the graph. She checks the name on the paper. It was submitted by the TA, her graduate student,



probably as a joke. So, what is the catch? The fit looks reasonable. She looks closely at the data table. There is an oddity that is so subtle that she almost misses it. Entries with an even time value (measured in milliseconds) tend to have position measurements that deviate high. Entries with odd time values tend to have position measurements that deviate low.

As a provocation, this is successful. It brings up subtle questions about statistics, causality, and “naturalness.” A way to generate the pattern would be to record times in a run and then go back and read positions with an extra rule: add a small deviation if the time is even, and subtract the same amount if it is odd. There is more to speculate about. In what other ways could such a correlation be achieved without manipulating the data? Could it be achieved with a modified (vibrating?) apparatus? Could the correlation occur as a statistical fluke? What principles guide the interpretation here, and are they the same principles used to detect the paper that was clearly faked? Could the other student papers that “make sense” not also have similar hidden effects? But if we allowed for that, could we even hold physics class?

Experts across physics, statistics, and philosophy could debate these questions at length, but experimentalists need to cut the debate short to get anything done. It is obvious that the correlation in the unicorn paper is suspicious; evenness is arbitrary in a measurement of time. Likewise, we expect a certain arbitrariness with respect to exact choices of sensor placement. Shifting a given sensor a little (modifying its associated time and distance measurements slightly) should not matter to the substance of what we observe. If the extra correlation in the unicorn paper were physically real, it would thus be inconsistent with known causal mechanisms for linear motion, assessed with clocks and rulers.

Causality is subtle, and causal reasoning in science is not straightforward to formalize. In statistics, there is an approach called Causal Analysis that describes how the relationship between causal factors in a scenario relates to statistical correlations in the observed data (Pearl 2009). In experimental research, usually such knowledge is implicit. We assume that persistent correlations between random variables have two possible reasons (Reichenbach’s Principle): one variable causes the other and/or both variables share common causes. The existence of a correlation under-determines the possible reasons why, but it implies that reasons should exist. Meanwhile, the absence of a correlation, if that too persists, shows independence.

This reasoning is part of paper-grading. An extra friction term is a plausible cause for certain extra correlations in linear motion. Variability among a set of five devices is one explanation for failing to see some expected correlations. Even if these judgments reflect general statistical principles, they are also hyper-specific in practice. The professor needs to know not just about how linear motion works, but about the devices in the room, about how clocks work, and how students work. A dishonest student is a plausible causal mechanism for a data set with unexpectedly low scatter; a grad student joke is a plausible explanation for the correlation in the unicorn paper.

In the language of Causal Analysis, fine-tuning refers to a case where the presence or the absence of correlations in a data set depends on specific values of parameters. It is a unicorn-like specialness where we expect the universe to

be indifferent. The faked correlation in the unicorn paper is a good joke to play on your graduate advisor, because if it were real, it would be special in that way. Fine-tuning is not always associated with deliberate fakery, though. The results of the collaborative student group show how fine-tuning can happen when the causal model for the phenomenon is not faithful to the measurement process. The lack of an observed correlation in that data set is suspicious-looking to the professor. Given the assumed causal mechanisms associated with a single air track, the data look impossible—unicorn-esque. But in both of these examples, using situation-specific knowledge to identify extra causal factors (intentional manipulation in one case and the extra apparatus variability in the other) resolves any actual mystery.

The overall point of this section is that models (equations) are “just metaphors.” By design, they are economical in their expression of structure and thus highly portable from one domain to another. Explanatory sense-making with real data, on the other hand, involves causal mechanisms. The details are anything but portable. The grading judgments (explanations) in our example showcase this. They might not even apply to a linear motion lab exercise done differently down the hall. They certainly would not apply to linear modeling in economics.

Experimental sense-making is more than applying a mathematical metaphor. It is hyper-specific. If we reject fine-tuned explanations (no unicorns), then persistent unexplained correlations are, perhaps, spooky.

## 5 Essential Quantumness? Entanglement and Bell Inequality Violations

Entanglement is frequently described as a (even the) quintessentially quantum phenomenon. In popular press, like many news reports surrounding the 2022 Nobel Prize in Physics, entanglement is associated with Einstein’s famous phrase “spooky action at a distance” or with the claim that physics has officially rejected “local realism” once and for all.

Given the discussion in Sect. 3, skepticism is generally warranted in any conversation that attempts to identify essential quantumness. This is true with entanglement. Some people (ourselves included) believe there is something spooky in some entanglement experiments. But even in quantum physics textbooks, many examples of entanglement do not exhibit the spooky effect. Thus quantum entanglement alone is not enough to challenge philosophical ideas like “realism.” The relationship to ideas about “locality” and “non-locality” is also complicated; some of the systems that exhibit the spooky kind of quantum entanglement have spatially separated parts, but some do not. Moreover, even among experts, “locality” and “non-locality” have a range of meanings (Cavalcanti and Wiseman 2012; Harrigan and Spekkens 2010; Wiseman 2006; Brown and Timpson 2014).

Meanwhile, the formalism that defines quantum entanglement is just as portable as the formula for a line. It codifies a type of non-separability that may be a perfectly

reasonable mathematical metaphor for systems in many modeling contexts, not just quantum physics labs. It is certainly used elsewhere in physics, with classical optics systems as a notorious example (Collins and Popescu 2002; Aiello et al. 2015; McLaren et al. 2015). The linear modeling examples from earlier in this chapter are a prompt to approach cross-comparisons between these cases with caution: some modeled “how” structures may be similar, while important data-specific details and “why” explanations may differ.

Thus, the ability to model a phenomenon as entanglement is an insufficient marker that the phenomenon is “essentially quantum.” But why seek such a marker in the first place? At present, there does seem to be a practical reason: there appear to be computational advantages associated with quantum algorithms.<sup>3</sup> Technology craves good quantum-classical deciders. Bell tests, as tests for evidence of the “spooky” effects in entangled systems, have historically been treated as useful in this way. Just as entanglement is more nuanced than the popular press might suggest, Bell tests do not function as a one-size-fits-all test for quantumness. There is more nuance because Bell tests are about the “why.”

A Bell test in quantum physics is a sense-making test assessing correlations observed in data. The assumed measurement scenario for a Bell test is generic. The test assesses correlations observed between the outcomes of two or more detectors, each with two or more possible settings that determine exactly what is measured. Given situation-specific assumptions about plausible causal mechanisms and measurement outcomes, it is typically possible to derive both lower and upper bounds on possible correlations in Bell tests (Popescu and Rohrlich 1998). Such a derivation results in a “Bell inequality,” characterizing those bounds. Data observed to violate the bounds can be interpreted as a challenge to the situation-specific causal assumptions.

Bell inequalities are not portable. It is impossible to derive and interpret such an inequality without articulating experiment-specific assumptions. The fictional linear motion lab from earlier in this chapter helps explain why. If an even-odd time correlation persisted in real linear motion data, we might devise a degree-of-spookiness statistical test to characterize what we saw as especially strange in that situation. This would be like a Bell test. The same test would not apply generically to another linear modeling context, e.g., for a cost-items model. A product could easily be cheaper in packs of two, violating the assumptions implicit in the linear-motion case.

The point is that correlations only register as spooky if there are situation-specific reasons to think the universe should not work that way. There have been a number of experiments in quantum physics, known as “loophole free” Bell tests (Giustina et al. 2015; Hensen et al. 2015; Shalm et al. 2015), that have repeatedly exhibited

---

<sup>3</sup> A review of quantum information theory is beyond the scope of this chapter, but some instructive related examples come from quantum game theory, e.g., the Mermin-Peres Magic Square (MPMS) game. Recent experimental results (Xu et al. 2022) have confirmed that the use of quantum entanglement in the MPMS game enables a winning strategy that exceeds what is possible with classical resources.

correlations that exceed what can be explained by classical causal mechanisms. The spookiness is assessed by applying situation-specific bounds on the expected degree of correlation based on classical statistical reasoning. Observed data violates these bounds.

It is subtle. Quantum physics is not supposed to be classical; quantum formalism explicitly predicts correlations that conflict with classical expectations. So is it really strange that quantum correlations match those predictions? Maybe not. Or maybe the spooky thing in Bell tests is not particular to Bell tests, or even entanglement, but is more about indeterminacy, or randomness, or (pick your favorite quantum oddity). But maybe what is going on here poses a deep challenge to our understanding of the nature of causality, indicating that there is something about quantum causality that remains to be understood (Pienaar 2017, 2020; Cavalcanti and Lal 2014), potentially a true essential quantumness, which relates to “why” not just “how.” Do we have consensus yet about any of this in physics? Not yet.

Ultimately, technological development may force the issue. If you want to capitalize on quantum effects for computation, it would be cheaper to reproduce those same effects with classical systems when possible. Does physics serve The Man, or fundamental knowledge? Either way, physicists are working hard to find classical ways to simulate—and thus, possibly, explain (erase)—what currently seems spooky in quantum Bell tests.

## 6 Bell’s Theorem: Two Ways

The most famous and canonical Bell test, articulated by John Bell himself (Bell 1964), involved mathematically derived correlation bounds—the Bell Inequalities—for the probabilities of two or more parties’ measurement outcomes, conditional on measurement settings that are chosen independently for each party. A Bell inequality does not stand alone; it gains its physical interpretation through *Bell’s Theorem*. The original version of that theorem states that statistics that arise from any “locally causal hidden variable model” for the scenario must satisfy the Bell inequality. Hidden variables are extra variables affecting the measurements but not included in the nominal model. In the fictional example earlier in this chapter, the data set made by the collaborative student group was subject to hidden variables: variations in air-track behaviors from one apparatus to another.

In the experiments addressed by Bell’s Theorem, the parties that perform measurements are distantly separated. A locally causal hidden variable model is one that obeys the physics-based constraint that faster-than-light-speed causal influences between their locations are ruled out. When an experiment designed to realize the assumptions in Bell’s Theorem shows measurement statistics that violate the appropriate Bell inequality, such a result cannot be explained by locally causal

hidden variable models. This is the conclusion sometimes interpreted as conflicting with “local realism.”<sup>4</sup>

However, Bell’s Theorem has been reformulated in multiple ways, and not all of them have the same philosophical connotations. Bell inequalities can be understood, for example, as a special case of a broader class of contextuality inequalities (Bell 1966; Kochen and Specker 1967; Kernaghan and Peres 1995; Cabello and García-Alcaine 1996). A relatively recent reformulation has also been performed using Causal Analysis, which is a method of inferring possible causal structures from the nature of the correlations that arise in an experiment. Studying the Bell scenario through this framework recasts the interpretive stakes in a useful way.

A Causal version of Bell’s Theorem was proposed by Wood and Spekkens in a seminal paper (Wood and Spekkens 2015). It enables a reformulation of Bell’s Theorem based on three assumptions about the experimental setup: (i) There are no direct causes between measurement outcomes on one party’s side and measurement outcomes on another party’s side. So, according to Reichenbach’s principle, any correlations that arise between measurement outcomes obtained by different parties *must* be due to a common cause. (ii) No-signaling: A choice of measurement setting on one party’s side does not affect the outcome of another party’s experiment and vice versa. (iii) Measurement setting independence: The choices of measurement settings by each party are made independently of each other.

When an experiment designed to meet these conditions violates a Bell inequality, it is interpreted to contradict one or more of the assumptions used to build the theorem. Those assumptions include the underlying assumptions of Causal Analysis, e.g., Reichenbach’s principle, and the assumption of no fine-tuning. No fine-tuning, as we explained in Sect. 4, is the assumption that the observed statistics are typical for the given causal structure and do not depend on special values of any parameters. It is the “no unicorns” rule. In Bell-like tests, it means that any observed statistical independence (including the no-signaling condition) happens because the causal structure of the phenomenon implies such an independence, not because it was engineered, or the result of some special “accident.”

Therefore, if the experiment ensures no-signaling and no direct causes between the outcomes on either side, we can conclude that there is no classical causal model that explains Bell-inequality-violating correlations without fine-tuning. This is why, from a causal modeling perspective, the (loophole-free) quantum violations of Bell inequalities register as distinctly spooky. Given causal models and hard-to-give-up causal assumptions, the observed correlations lead to a logical contradiction. Something is happening that does not fit into the framework of a classical causal model unless we allow unicorn-like interventions. And because no-fine-tuning is

---

<sup>4</sup> Local realism is generally defined as two assumptions taken *together*: (i) Realism: “If, without disturbing a system, we can predict with certainty the value of a physical quantity, then there exists an element of physical reality corresponding to this physical quantity. The element of reality represents the predetermined value for the physical quantity.” (Einstein et al. 1935) (ii) Locality: physical influences between spatially separated systems cannot propagate faster than the speed of light.

such an important concept in sense-making in other experimental contexts, that is not an easy thing to accept.

## 7 Classical Bell-Like Tests

Insofar as a Bell-inequality-violating entangled system may be a useful resource for quantum communication, it may in the future be treated as a special kind of black box, like the SQUIDs discussed earlier. And like with a SQUID, the physical conditions required to operate such a quantum black box would likely be expensive (e.g., requiring low temperatures and a high degree of isolation from the environment). Thus, if we view Bell-like correlations as a resource to be exploited, there is a motivation to understand whether cheaper, classically based black boxes could allow some or all of the same communication and cryptography tricks. This is one reason that optics researchers study which quantum computation effects can be simulated with classical light and which cannot (van Enk and Fuchs 2002). In particular, experiments ask: can classical light that exhibits non-separability (“classical entanglement”) produce results that violate Bell inequalities (Borges et al. 2010; Qian et al. 2015; Li et al. 2018)?

The causal Bell’s Theorem is helpful for carefully framing the analysis of such experiments. It replaces the notion of space-like separation with the weaker notion of no-signaling. The condition of no-signaling is naturally satisfied by any pair of non-interacting degrees of freedom, such as the polarization and spatial modes of (paraxial) light, even if they exist in the same beam and hence are not space-like separated. Thus, the Causal Bell’s Theorem might be applicable to classical light, even if the original Bell’s Theorem does not apply. To derive the inequalities, it is necessary that certain *non-disturbance* relations hold among the relevant variables, of which *no-signaling* is a special case. The idea to use an interferometric setup to violate a Bell-like inequality with classical light was originally proposed by Suppes et al. (1996). Shortly thereafter, Spreeuw (1998) introduced the concept of classical non-separability between two or more degrees of freedom of the same light beam. Numerous authors subsequently analyzed this concept (Aiello et al. 2015; Pereira et al. 2014; Qian and Eberly 2011; Van Enk 2003), including its application to experimental violations of Bell-like inequalities (Borges et al. 2010; Goldin et al. 2010; Frustaglia et al. 2016; Li et al. 2018; Qian et al. 2015).

Apparent violations of Bell inequalities with classical optics systems have been repeatedly demonstrated. Does this show that there is nothing especially spooky about quantum Bell tests? Or, the reverse that the same spookiness can be created in classical systems? We would argue, given the evidence to date, no. It does not show either of these things. The traditional Bell inequality (and its specific bounds on classically induced correlations) is not easily portable. When Bell inequalities are re-derived with situation-specific measurement assumptions, the inequality bounds may change, which changes the interpretation of measured correlations. A recent work by Markiewicz et al. (2019) makes this argument for the classical light case.

The authors point out that the type of measurement involved in classical light Bell tests differs significantly from the type of measurement in quantum Bell tests. The appropriate Bell's inequality for the classical light context would need to be an inequality between measured *field intensities* and not probabilities of “clicks” of detectors. A re-derived version of the Bell inequality for this case shows different bounds. Thus the observed correlations in the classical case are not “spooky.” They are consistent with classical causal expectations and apparently have little direct bearing on interpreting the quantum Bell test case.

In other words, the exact nature of the measurement matters to the derivation of the Bell test logic. Sense-making is specific. There may be some resemblance between classical non-separability experiments and quantum entanglement experiments, but a careful analysis of the causal assumptions embedded in the measurement scenario differs. Characterizing what would be “spooky” in each case depends on situation-specific expectations about causality.

## 8 Application to Interdisciplinary Contexts

Paper titles across multiple disciplines advertise Bell-inequality violations in novel systems as key results that demonstrate “non-classical behavior” or purportedly demonstrate entanglement. Examples include analysis of concept combinations and word associations, e.g., (Beltran and Geriente 2019; Aerts et al. 2021). There is also a long history of toy-model examples, e.g., involving rubber bands (Sassoli de Bianchi 2013) or socks (Aerts and de Bianchi 2019). What do Bell-inequality violations in these cases mean?

Well first, it helps to remember the general cautions on modeling and metaphor from the beginning of this chapter, especially the how vs. why distinction. If the work references a “Bell inequality,” probably some form of non-separability is involved, in metaphorical correspondence to entanglement. Likely the work describes a measurement scenario with some similarity to the ones used in quantum Bell experiments. However, as we have argued, for a Bell inequality to be used as a meaningful sense-making test, it must be derived appropriately for the types of causal mechanisms that the research intends to test. Parts of the Bell argument may function as portable metaphors, but the exact derived bounds, and the interpretation of measured correlations in a given case, are not portable pieces of argumentation. They relate to “why” questions that are measurement-specific.

Based on what we have learned from physics, it is possible to create apparent Bell-inequality violations in systems governed by classical physics and classical causality. This does not explain quantum Bell test correlations, nor does it show that the same thing is happening in a classical system. Historically, violation of Bell inequalities may have been considered as a type of special test for quantumness, but recent work, like that referenced in this chapter, shows that the meaning of a Bell-inequality violation can only be rigorously interpreted within a full-blown Bell test.



We have to derive appropriate statistical bounds articulating the “why” explanations we wish to test for these measurements.

The Causal Bell Theorem provides a useful tool for thinking about Bell tests, by recasting the situation in terms of unicorn-like fine-tuning. This reasoning is particularly applicable to some of the toy model examples we have seen, e.g., (Aerts and de Bianchi 2019). A scenario that explicitly defines a way to produce Bell-violating correlations in a measurement is like a graduate student explicitly inventing a way to introduce an even-odd time correlation in motion lab data. The extra correlation would be “spooky” only if we imagine that it could happen without the intentional, extra, fine-tuned causation. In other words, a professor “grading” the results would only find the data suspicious if she expected a different causal process than the one actually involved. Thus, toy-model examples that articulate a causal process to violate Bell inequalities are different from a fully wrought Bell test. The same comments could easily apply for other purported Bell-inequality violations. As physicists, we are less sure how to approach sense-making with data sets of word or concept associations. But we can ask: are there well-defined expectations about the statistical behavior of such data sets, given classical causal reasoning? A “yes” response to that question, and an experiment-specific Bell-inequality derivation, would seem necessary in order to interpret any purported violation.

As of the writing of this chapter, we think there is still something special in quantum physics Bell tests that does not happen in classical optics, in word associations, rubber bands, or other contexts. We subjectively judge the problem to be “spooky,” but the important point is that the spookiness is not trivially imitated in systems that share some modeled structures. It has to do with our expectations about the causal processes involved.

## 9 In Conclusion: Prove Us Wrong

We began this chapter with two claims: (a) that there is something spooky in specific Bell test results in quantum physics and (b) that this spookiness is not present in other modeling contexts.

For us, claim (a) is amenable to modification as research in quantum foundations progresses. In the space of technical debates about quantum physics, we could plausibly be convinced that nothing is particularly spooky about loophole-free Bell test results, e.g., if we are convinced to adopt particular approaches to quantum interpretation (some of which view Bell tests as unremarkable). We also could have our views shifted by novel results in the cross-comparison of classical and quantum optics systems, as just one example. A number of the results we reference in this chapter are relatively recent. Thus the conclusions we draw from them may evolve as more experiments are performed and the discussion matures. Quantum Bell tests have seemed strange for some decades now, but they may not seem strange in the future, depending on what we learn next.



Claim (b) is the one we stand behind firmly for this chapter. It is based primarily on general points about modeling; that modeling employs a type of metaphor. Cross-disciplinary “how” similarities may exist between two disciplinary domains in which the causal “why” radically differs. Since the Bell test spookiness is about causal sense-making, it does not translate in a simple way from one measurement scenario to another. This is clearly demonstrated within physics, considering the differences in interpretation needed to evaluate Bell-like scenarios involving classical light.

Still, claim (b) can also be confronted and potentially falsified. If a Bell-like test situation in a non-quantum physics context is analyzed in such a way as to show persistent correlations that cannot be explained with the causal mechanisms that make sense in that case, it may also be (like Bell tests in quantum physics) spooky in an important way. To make the argument, though, most likely the relevant inequalities or important bounds need to be re-derived in a customized, situation-specific manner.

A methodological prescription for attempting such an effort (to prove that a Bell-like test in a novel context is similarly “spooky”) might begin with formally articulating a causal model and assessing consistency within that framework. Informally, think about unicorns. Can the data be explained causally, with no fine-tuning? If so, it is different from what is happening in quantum Bell tests. If not, then possibly you will indeed join us in questioning causality in your measurement context, spooked by your own specific Bell test ghost.

**Acknowledgments** GBL acknowledges financial support from the Brazilian agencies CNPq and FAPERJ (JCNE E-26/201.438/2021). We thank reviewers Fernando de Melo and Jason Gallicchio for helpful comments that substantially improved the final draft.

## References

- Diederik Aerts, Jonito Aerts Arguëlles, Lester Beltran, Suzette Geriente, and Sandro Sozzo. Entanglement in cognition violating Bell inequalities beyond Cirel’son’s bound. *arXiv preprint arXiv:2102.03847*, 2021.
- Diederik Aerts and Massimiliano Sassoli de Bianchi. When Bertlmann wears no socks. common causes induced by measurements as an explanation for quantum correlations. *arXiv preprint arXiv:1912.07596*, 2019.
- Andrea Aiello, Falk Töppel, Christoph Marquardt, Elisabeth Giacobino, and Gerd Leuchs. Quantum-like nonseparable structures in optical beams. *New Journal of Physics*, 17(4):043024, 2015.
- John S Bell. On the Einstein Podolsky Rosen paradox. *Physics Physique Fizika*, 1(3):195, 1964.
- John S Bell. On the problem of hidden variables in quantum mechanics. *Reviews of Modern physics*, 38(3):447, 1966.
- Lester Beltran and Suzette Geriente. Quantum entanglement in corpuses of documents. *Foundations of Science*, 24:227–246, 2019.
- C. V. S. Borges, M. Hor-Meyll, J. A. O. Huguenin, and A. Z. Khoury. Bell-like inequality for the spin-orbit separability of a laser beam. *Phys. Rev. A*, 82:033833, Sep 2010.

- Harvey R Brown and Christopher G Timpson. Bell on Bell's theorem: The changing face of nonlocality. *arXiv preprint arXiv:1501.03521*, 2014.
- Adán Cabello and Guillermo García-Alcaine. Bell-Kochen-Specker theorem for any finite dimension. *Journal of Physics A: Mathematical and General*, 29(5):1025, 1996.
- Eric G Cavalcanti. Classical causal models for Bell and Kochen-Specker inequality violations require fine-tuning. *Physical Review X*, 8(2):021018, 2018.
- Eric G Cavalcanti and Raymond Lal. On modifications of Reichenbach's principle of common cause in light of Bell's theorem. *Journal of Physics A: Mathematical and Theoretical*, 47(42):424018, 2014.
- Eric G Cavalcanti and Howard M Wiseman. Bell nonlocality, signal locality and unpredictability (or what Bohr could have told Einstein at Solvay had he known about Bell experiments). *Foundations of Physics*, 42:1329–1338, 2012.
- Daniel Collins and Sandu Popescu. Classical analog of entanglement. *Physical Review A*, 65(3):032321, 2002.
- Albert Einstein, Boris Podolsky, and Nathan Rosen. Can quantum-mechanical description of physical reality be considered complete? *Physical review*, 47(10):777, 1935.
- S. J. van Enk and C. A. Fuchs. Quantum State of a Propagating Laser Field. *Quantum Information and Computation*, 2:151–165, 2002.
- Diego Frustaglia, José P Baltanás, María C Velázquez-Ahumada, Armando Fernández-Prieto, Aintzane Lujambio, Vicente Losada, Manuel J Freire, and Adán Cabello. Classical physics and the bounds of quantum correlations. *Physical review letters*, 116(25):250404, 2016.
- Marissa Giustina, Marijn AM Versteegh, Sören Wengerowsky, Johannes Handsteiner, Armin Hochrainer, Kevin Phelan, Fabian Steinlechner, Johannes Kofler, Jan-Åke Larsson, Carlos Abellán, et al. Significant-loophole-free test of Bell's theorem with entangled photons. *Physical review letters*, 115(25):250401, 2015.
- Matias A Goldin, Diego Francisco, and Silvia Ledesma. Simulating Bell inequality violations with classical optics encoded qubits. *JOSA B*, 27(4):779–786, 2010.
- Nicholas Harrigan and Robert W Spekkens. Einstein, incompleteness, and the epistemic view of quantum states. *Foundations of Physics*, 40:125–157, 2010.
- Bas Hensen, Hannes Bernien, Anaïs E Dréau, Andreas Reiserer, Norbert Kalb, Machiel S Blok, Just Ruitenberg, Raymond FL Vermeulen, Raymond N Schouten, Carlos Abellán, et al. Loophole-free Bell inequality violation using electron spins separated by 1.3 kilometres. *Nature*, 526(7575):682–686, 2015.
- Ebrahim Karimi and Robert W Boyd. Classical entanglement? *Science*, 350(6265):1172–1173, 2015.
- Michael Kernaghan and Asher Peres. Kochen-Specker theorem for eight-dimensional space. *Physics Letters A*, 198(1):1–5, 1995.
- DN Klyshko. A simple method of preparing pure states of an optical field, of implementing the Einstein–Podolsky–Rosen experiment, and of demonstrating the complementarity principle. *Soviet Physics Uspekhi*, 31(1):74, 1988.
- Simon Kochen and E. P. Specker. The problem of hidden variables in quantum mechanics. *Journal of Mathematics and Mechanics*, 17(1):59–87, 1967.
- Raymond ST Lee. *Quantum finance*. Springer, 2020.
- Tao Li, Xiong Zhang, Qiang Zeng, Bo Wang, and Xiangdong Zhang. Experimental simulation of monogamy relation between contextuality and nonlocality in classical light. *Opt. Express*, 26(9):11959–11975, Apr 2018.
- Marcin Markiewicz, Dagomir Kaszlikowski, Paweł Kurzyński, and Antoni Wójcik. From contextuality of a single photon to realism of an electromagnetic wave. *npj Quantum Information volume*, 5:5, 2019.
- Melanie McLaren, Thomas Konrad, and Andrew Forbes. Measuring the nonseparability of vector vortex beams. *Physical Review A*, 92(2):023833, 2015.
- Joshua Montgomery, Adam J Anderson, Jessica S Avva, Amy N Bender, Matt A Dobbs, Daniel Dutcher, Tucker Elleflot, Allen Foster, John C Groh, William L Holzapfel, et al. Performance and characterization of the SPT-3G digital frequency multiplexed readout system using an

- improved noise and crosstalk model. In *Millimeter, Submillimeter, and Far-Infrared Detectors and Instrumentation for Astronomy X*, volume 11453, pages 167–188. SPIE, 2020.
- JC Pearl and EG Cavalcanti. Classical causal models cannot faithfully explain Bell nonlocality or Kochen-Specker contextuality in arbitrary scenarios. *Quantum*, 5:518, 2021.
- Judea Pearl. *Causality*. Cambridge university press, 2009.
- L. J. Pereira, A. Z. Khoury, and K. Dechoum. Quantum and classical separability of spin-orbit laser modes. *Phys. Rev. A*, 90:053842, Nov 2014.
- Jacques Pienaar. Causality in the quantum world. *Physics*, 10:86, 2017.
- Jacques Pienaar. Quantum causal models via quantum Bayesianism. *Physical Review A*, 101(1):012104, 2020.
- Sandu Popescu and Daniel Rohrlich. Causality and nonlocality as axioms for quantum mechanics. In *Causality and Locality in Modern Physics: Proceedings of a Symposium in honour of Jean-Pierre Vigi er*, pages 383–389. Springer, 1998.
- Emmanuel M Pothos and Jerome R Busemeyer. Quantum cognition. *Annual review of psychology*, 73:749–778, 2022.
- Xiao-Feng Qian and J. H. Eberly. Entanglement and classical polarization states. *Opt. Lett.*, 36(20):4110–4112, Oct 2011.
- Xiao-Feng Qian, Bethany Little, John C Howell, and JH Eberly. Shifting the quantum-classical boundary: theory and experiment for statistically classical optical fields. *Optica*, 2(7):611–615, 2015.
- Massimiliano Sassoli de Bianchi. Using simple elastic bands to explain quantum mechanics: a conceptual review of two of Aerts’ machine-models. *Open Physics*, 11(2):147–161, 2013.
- Kathryn Schaffer and Gabriela Barreto Lemos. Obliterating thingness: an introduction to the “what” and the “so what” of quantum physics. *Foundations of Science*, 26:7–26, 2021.
- Lynden K Shalm, Evan Meyer-Scott, Bradley G Christensen, Peter Bierhorst, Michael A Wayne, Martin J Stevens, Thomas Gerrits, Scott Glancy, Deny R Hamel, Michael S Allman, et al. Strong loophole-free test of local realism. *Physical review letters*, 115(25):250402, 2015.
- R Simon and GS Agarwal. Wigner representation of Laguerre–Gaussian beams. *Optics letters*, 25(18):1313–1315, 2000.
- Robert JC Spreeuw. A classical analogy of entanglement. *Foundations of physics*, 28(3):361–374, 1998.
- David Stoler. Operator methods in physical optics. *JOSA*, 71(3):334–341, 1981.
- P. Suppes, J. Acacio de Barros, and A.S. Sant’Anna. A proposed experiment showing that classical fields can violate Bell’s inequalities. *ArXiv:9606019*, 1996.
- Ilya A Surov, E Semenenko, AV Platonov, IA Bessmertny, F Galofaro, Zeno Toffano, A Yu Khrennikov, and AP Alodjants. Quantum semantics of text perception. *Scientific Reports*, 11(1):1–13, 2021.
- SJ Van Enk. Entanglement of electromagnetic fields. *Physical Review A*, 67(2):022303, 2003.
- Howard M Wiseman. From Einstein’s theorem to Bell’s theorem: a history of quantum non-locality. *Contemporary Physics*, 47(2):79–88, 2006.
- Christopher J Wood and Robert W Spekkens. The lesson of causal discovery algorithms for quantum correlations: Causal explanations of Bell-inequality violations require fine-tuning. *New Journal of Physics*, 17(3):033002, 2015.
- Jia-Min Xu, Yi-Zheng Zhen, Yu-Xiang Yang, Zi-Mo Cheng, Zhi-Cheng Ren, Kai Chen, Xi-Lin Wang, and Hui-Tian Wang. Experimental demonstration of quantum pseudotelepathy. *Physical Review Letters*, 129(5):050402, 2022.

# Compositional Vector Semantics in Spiking Neural Networks



Martha Lewis

**Abstract** Categorical compositional distributional semantics is an approach to modelling language that combines the success of vector-based models of meaning with the compositional power of formal semantics. However, this approach was developed without an eye to cognitive plausibility. Vector representations of concepts and concept binding are also of interest in cognitive science and have been proposed as a way of representing concepts within a biologically plausible spiking neural network. This work proposes a way for compositional distributional semantics to be implemented within a spiking neural network architecture, with the potential to address problems in concept binding and give a small implementation. We also describe a means of training word representations using labelled images.

**Keywords** Compositional semantics · Spiking neural networks · Word representation · Semantic pointer architecture · Quantum cognition

## 1 Introduction

Vector representations of word meaning have proved extremely successful at modelling language in recent years, both as static word embeddings (Mikolov et al. 2013; Pennington et al. 2014) and as contextual embeddings that take surrounding words into account (Devlin et al. 2019). Vectors have also been used in cognitive science, both at a fairly abstract level representing concepts via collections of features and at a more mechanistic level representing concepts via patterns of neural activation.

In all cases we have an interest in describing how words or concepts combine. In the case of contextual word embeddings, composition is effected by the artificial neural network architecture, and this works very well, although in an opaque manner

---

M. Lewis (✉)

Department of Engineering Mathematics, University of Bristol, Bristol, UK  
e-mail: [martha.lewis@bristol.ac.uk](mailto:martha.lewis@bristol.ac.uk)

and arguably in a way that does not generalise effectively to other tasks (Talman and Chatzikyriakidis 2019; Bernardy and Chatzikyriakidis 2019) or which leverages specific characteristics of the dataset (McCoy et al. 2019).

In the case of static word embeddings, compositional distributional semantics describes methods to both build vector representations of words and combine them together so that phrases and sentences can be represented as vectors. Mitchell and Lapata (2010) describe some quite general approaches to composition and give implementations focussed on pointwise, potentially weighted, combinations, such as vector addition or pointwise multiplication. Grammatically informed neural approaches are given in (Socher et al. 2013; Bowman et al. 2015) where artificial neural networks for composing word vectors are built that use the grammatical structure of a sentence. Finally, tensor-based approaches were proposed and developed in Coecke et al. (2010), Baroni and Zamparelli (2010), Paperno et al. (2014). In these approaches, words are modelled in different vector spaces depending on their grammatical type, and composition is given by tensor contraction. This will be described in more detail in Sect. 2.1. Compositional distributional semantic approaches are in general used to model text only, although some multi-modal approaches have been used. This leads to the question of whether we can develop a means of learning compositional vector-based representations in a grounded way.

On the cognitive side, we focus here on the idea of vectors as representing patterns of neural activation. One means of considering how vectors combine in this context is given by *vector symbolic architectures* (VSAs) (Smolensky 1990; Plate 1994; Gayler 2003). VSAs represent symbols as vectors and provide a means of binding symbols together, grouping them, and unbinding them as needed. More detail is given in Sect. 2.2. VSAs have been posited as a way of modelling how symbols can be represented and manipulated in a neural substrate. This has been implemented in, e.g., Eliasmith et al. (2012) and investigated and discussed in, e.g., Hummel (2011), Martin and Doumas (2020). In Martin and Doumas (2020), the argument is made that *additive binding*, i.e., combining vectors via addition, is more faithful to how humans combine concepts than *conjunctive binding*, i.e., using something like a tensor product. Since VSAs have been investigated and implemented within more biologically realistic neural networks, the question arises of whether we can use these methods in developing a grounded learning model for tensor-based compositional semantics, all the more so since the model of Coecke et al. (2010) was inspired by Smolensky's model originally.

## Aims

- To build a model for generating grounded representations within a compositional distributional semantics
- To draw out links between Smolensky's theory on one hand and compositional distributional semantics on the other
- To develop a way in which compositional distributional models can be implemented within biologically plausible neural network models and thereby investigate a wider range of composition methods than, e.g., vector addition or tensor binding

## 2 Background

### 2.1 *Compositional Distributional Semantics*

Compositional distributional semantics was developed as a way of generating meanings above the word level via the composition of individual word meanings. The genre of model we concentrate on here can be termed tensor-based compositional distributional models (Coecke et al. 2010; Baroni and Zamparelli 2010; Paperno et al. 2014). Words are modelled in different vector spaces according to their grammatical type, and composition is modelled as tensor contraction. Specifically:

- Nouns are modelled as vectors in a noun space  $N$ , sentences in a sentence space  $S$ .
- Adjectives are modelled as matrices on  $N$ , i.e., linear maps  $adj : N \rightarrow N$  or elements of the space  $N \otimes N$ .
- Intransitive verbs are modelled as matrices from  $N$  to  $S$ , i.e., linear maps  $iv : N \rightarrow S$  or elements of the space  $N \otimes S$ .
- Transitive verbs are modelled as tensors in  $N \otimes S \otimes N$ , or bilinear maps  $tv : N \otimes N \rightarrow S$ .

So, for example, an adjective like **red** is modelled as a matrix **red** and applied to a noun **car** by matrix multiplication, giving back a vector **red car**.

In the original formulation, vectors for words were presumed to be inferred from large text corpora, and so far there have been limited proposals for how to ground these representations using images or other forms of input. Compositional distributional semantics has also been developed with limited consideration of cognitive or neural plausibility. In contrast, vector symbolic architectures, discussed in the following section, have been considered by some cognitive scientists as offering a good basis for modelling language and concept combination.

### 2.2 *Vector-Symbolic Architectures*

Prior to tensor-based distributional semantics, there has been a large amount of research into vector symbolic architectures (VSAs), specifically, how symbolic structures can be encoded into vector-based representations. These include Smolensky (1990), Plate (1994), Gayler (2003) among others.

Smolensky (1990) proposes that structures like sentences are modelled as a sum of role–filler bindings. Suppose we have a set of roles  $\{agent, patient, verb\}$  and a set of fillers  $\{Junpa, Jen, loves\}$ . Symbolically, the binding of a role to a filler is represented by  $/$ , and the sentence *Junpa loves Jen* can be represented as a set of role–filler bindings  $\{Junpa/agent, Jen/patient, loves/verb\}$ . In Smolensky (1990), these are mapped over to a vector space model by mapping each role and each filler to a vector, mapping the binding to tensor product, and mapping the collection of

the role–filler bindings as their sum:

$$Junpa \text{ loves } Jen \mapsto \{Junpa/agent, Jen/patient, loves/verb\} \quad (1)$$

$$\mapsto \mathbf{Junpa} \otimes \mathbf{agent} + \mathbf{Jen} \otimes \mathbf{patient} + \mathbf{loves} \otimes \mathbf{verb} \quad (2)$$

More generally, a sentence  $s$  consisting of a set of role–filler bindings  $\{r_i/f_i\}_i$  is realized as

$$\mathbf{s} = \sum_i \mathbf{r}_i \otimes \mathbf{f}_i \quad (3)$$

Questions can be asked of a given statement via an *unbinding mechanism*. We may want to extract individual elements of a given sentence. This is done using *unbinding vectors*, defined as vectors dual to the role vectors. Each role vector  $\mathbf{r}_i$  has an unbinding vector  $\mathbf{u}_i$  such that  $\langle \mathbf{r}_i, \mathbf{u}_i \rangle = 1$ . Note that if the role vectors are an orthonormal set, each role vector is its own unbinding vector. To unbind a particular role from a sentence, we take the partial inner product of the unbinding vector with the sentence representation. Suppose that

$$\mathbf{s} = \mathbf{Junpa} \otimes \mathbf{agent} + \mathbf{Jen} \otimes \mathbf{patient} + \mathbf{loves} \otimes \mathbf{verb} \quad (4)$$

and **agent**, **patient**, **verb** form an orthonormal set. If we want to know who the agent is in  $\mathbf{s}$ , we take the partial inner product of  $\mathbf{s}$  and **agent** giving:

$$\begin{aligned} \mathbf{s} \cdot \mathbf{agent} &= (\mathbf{Junpa} \otimes \mathbf{agent} + \mathbf{Jen} \otimes \mathbf{patient} + \mathbf{loves} \otimes \mathbf{verb}) \cdot \mathbf{agent} \\ &= \mathbf{Junpa} \otimes \mathbf{agent} \cdot \mathbf{agent} + \mathbf{Jen} \otimes \mathbf{patient} \cdot \mathbf{agent} \\ &\quad + \mathbf{loves} \otimes \mathbf{verb} \cdot \mathbf{agent} \\ &= \mathbf{Junpa} \end{aligned}$$

VSA's have been posited as a potential means for representing symbolic thought in a neural substrate (Hummel 2011; Doumas and Hummel 2012; Calmus et al. 2020; Martin and Doumas 2020). Holographic reduced representations (Plate 1994) have similar properties to Smolensky's theory but without the drawback of needing the increased space for bound representations. Eliasmith (2013) has shown how HRRs can be implemented within a spiking neural network model that is designed to be biologically plausible.

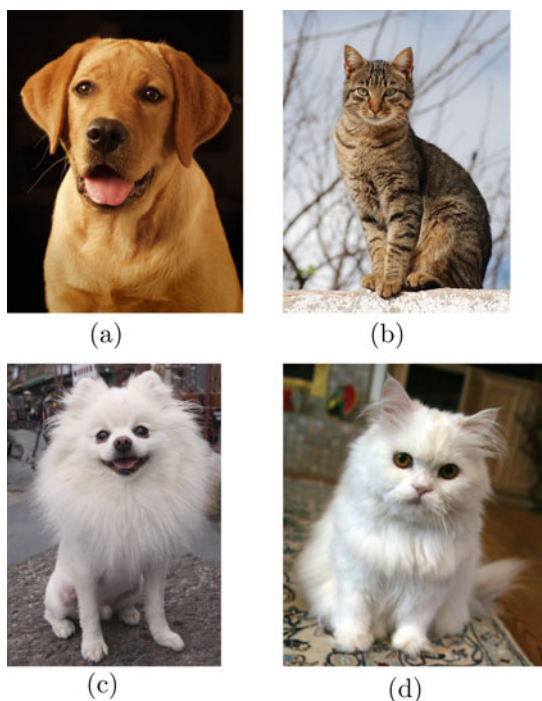
While VSAs such as Smolensky's and Plate's have the benefits outlined above, of being composable and potentially biologically plausible, they have been argued to have some drawbacks in representing how words combine, or how concepts compose. Hummel (2011), Doumas and Hummel (2012), Martin and Doumas (2020) make the following observation. Assuming that similarity is measured by cosine similarity, that is, the cosine of the angle between two vectors, then the similarity of two role–filler bindings is dependent only on the similarity of the pair of roles and the pair of fillers.

Martin and Doumas (2020) set up an experiment to investigate whether *conjunctive binding* (via tensor product or circular convolution) or *additive binding* (via vector addition) is a better predictor of human similarity judgements. They consider a role to be a predicate, such as *fluffy*, which can be bound to a filler, such as *cat*. Then, the representations for *fluffy dog* and *fluffy cat* are exactly as similar as the representations for *dog* and *cat* are.

$$\mathbf{fluffy} \otimes \mathbf{dog} \cdot \mathbf{fluffy} \otimes \mathbf{cat} = \mathbf{fluffy} \cdot \mathbf{fluffy} \otimes \mathbf{dog} \cdot \mathbf{cat} \quad (5)$$

$$= \mathbf{1} \cdot \mathbf{dog} \cdot \mathbf{cat} = \mathbf{dog} \cdot \mathbf{cat} \quad (6)$$

This is undesirable, as the predicate *fluffy* should make cats and dogs more similar to each other – see Fig. 1 for example.



**Fig. 1** Fluffy dogs and fluffy cats are more similar than dogs and cats. (a) Dog ([https://commons.wikimedia.org/wiki/File:Wayfield%27s\\_Young\\_Argos,\\_the\\_Labrador\\_Retriever.jpg](https://commons.wikimedia.org/wiki/File:Wayfield%27s_Young_Argos,_the_Labrador_Retriever.jpg). Attribution: Andrew Skolnick, en:User:Askolnick, CC BY-SA 3.0 < <http://creativecommons.org/licenses/by-sa/3.0/>>, via Wikimedia Commons). (b) Cat ([https://commons.wikimedia.org/wiki/File:Cat\\_November\\_2010-1a.jpg](https://commons.wikimedia.org/wiki/File:Cat_November_2010-1a.jpg). Attribution: Alvesgaspar, CC BY-SA 3.0 <https://creativecommons.org/licenses/by-sa/3.0/>, via Wikimedia Commons). (c) Fluffy dog ([https://commons.wikimedia.org/wiki/File:Cute\\_dog\\_on\\_beach,\\_Los\\_Angeles,\\_California\\_LCCN2013634674.tif](https://commons.wikimedia.org/wiki/File:Cute_dog_on_beach,_Los_Angeles,_California_LCCN2013634674.tif). Attribution: Carol M. Highsmith, Public domain, via Wikimedia Commons). (d) Fluffy cat ([https://commons.wikimedia.org/wiki/File:White\\_Persian\\_Cat.jpg](https://commons.wikimedia.org/wiki/File:White_Persian_Cat.jpg). Attribution: Optional at the Persian language Wikipedia, GFDL <http://www.gnu.org/copyleft/fdl.html>, via Wikimedia Commons)



Alternatively, roles might be bound to fillers using *additive binding*. In this case, we do obtain the result that fluffy dogs and fluffy cats are more similar than dogs and cats. This property also holds if we model *fluffy* as an adjective, *dog* and *cat* as nouns, and combine the representations as in Eq. (3) above. Bearing in mind that we consider **adj** and **noun** to be orthogonal, and assume we have a similarity of around 0.5 for cats and dogs, we have

$$\text{fluffy cat} = \frac{1}{\sqrt{2}}(\text{fluffy} \otimes \text{adj} + \text{cat} \otimes \text{noun}) \quad (7)$$

$$\text{fluffy dog} = \frac{1}{\sqrt{2}}(\text{fluffy} \otimes \text{adj} + \text{dog} \otimes \text{noun}) \quad (8)$$

where the factor of  $\frac{1}{\sqrt{2}}$  normalises the length of the resulting vectors. Then:

$$\begin{aligned} \text{sim}(\text{fluffy cat}, \text{fluffy dog}) \\ = \frac{1}{2} \langle \text{fluffy} \otimes \text{adj} + \text{cat} \otimes \text{noun}, \text{fluffy} \otimes \text{adj} + \text{dog} \otimes \text{noun} \rangle \end{aligned} \quad (9)$$

$$= \frac{1}{2} (\langle \text{fluffy} \otimes \text{adj}, \text{fluffy} \otimes \text{adj} \rangle + \langle \text{fluffy} \otimes \text{adj}, \text{dog} \otimes \text{noun} \rangle) \quad (10)$$

$$+ \langle \text{cat} \otimes \text{noun}, \text{fluffy} \otimes \text{adj} \rangle + \langle \text{cat} \otimes \text{noun}, \text{dog} \otimes \text{noun} \rangle) \quad (11)$$

$$= \frac{1}{2} (\langle \text{fluffy} \otimes \text{adj}, \text{fluffy} \otimes \text{adj} \rangle + \langle \text{fluffy}, \text{dog} \rangle \langle \text{adj}, \text{noun} \rangle) \quad (12)$$

$$+ \langle \text{fluffy}, \text{cat} \rangle \langle \text{adj}, \text{noun} \rangle + \langle \text{cat} \otimes \text{noun}, \text{dog} \otimes \text{noun} \rangle) \quad (13)$$

$$= \frac{1}{2} (1 + 0 + 0 + \langle \text{cat}, \text{dog} \rangle) = 0.75 \quad (14)$$

so applying *fluffy* has boosted the similarity of cats and dogs.

**Why We Want More Than Additive Binding** While, as can be seen in Martin and Doumas (2020), additive binding works for many examples, there are arguably a number of predicates for which we do not want an increase in similarity to occur. Some words are ambiguous, and we do not want to say that *bright light* and *bright student* are more similar than *light* and *student* are. One argument might be that we should disambiguate words and that these different senses will be represented by different vectors. However, some ambiguities can be subtle: compare *Nishi opened the book* with *Nishi opened the jar*.

Boleda (2020) argues that tensor-based compositional distributional models of meaning can reflect polysemy, and this assertion is borne out by the experimental results of Grefenstette and Sadrzadeh (2011), where a matrix-based method is compared with (among others) an additive model. A hand-crafted example of this is given in Grefenstette et al. (2010) where they show how to design a representation for *catch* so that the similarity of the phrases *catch ball* and *catch disease* is not

boosted by the word *catch*. Coecke and Lewis (2015) look at the pet fish problem and show how a representation for *pet* can be developed that sends *fish* to *goldfish* but leaves *dog* and *cat* mostly unchanged.

The compositional distributional model is focussed on corpus-based semantics and was not developed with any eye to neural plausibility or to learning language in a grounded fashion. We would like to be able to combine the greater flexibility of tensor-based compositional models with the neural plausibility of VSA-based architectures, and with the possibility of learning meanings that are grounded in an input outside of text. We provide a mapping from the compositional distributional model into a Smolensky-based architecture. We show how meanings of words can be learnt in a way that is grounded in inputs outside of text.

**Relations to Quantum Models of Concepts** The compositional distributional semantic model of meaning proposed in Coecke et al. (2010) has its roots in quantum theory. There is a wide range of interest in quantum modelling of concepts, examining how concepts behave in composition, and their application in artificial intelligence. Widdows et al. (2021) provide a thorough review of applications of quantum theory to artificial intelligence, part of which is to do with the representation of concepts or words by vectors, together with an analysis of the use of VSA-like architectures that we cover here. In the area of concept composition, it is proposed that concept composition can be well modelled as interference between two quantum states (Aerts 2009; Aerts et al. 2012). Aspects such as emergent meaning and vagueness are addressed in Bruza et al. (2012), Blutner et al. (2013), and reviews of the use of quantum theory in cognitive science are given in Pothos and Busemeyer (2013) and Lewis (2021). Further consideration of whether these phenomena can be well modelled within neural networks is an area of future work.

### 2.3 *Semantic Pointers for Concept Representation*

While deep neural architectures have had huge success in recent years, they are biologically implausible in the structure of the individual units, the overall architecture of networks, and in the learning algorithm implemented. There has therefore been interest in implementing more biologically plausible networks. One of these is Nengo (Eliasmith et al. 2012). Within this architecture, symbolic structure is implemented using the Semantic Pointer Architecture (SPA) (Eliasmith et al. 2012; Blouw et al. 2016). Semantic pointers can be thought of as vectors that are instantiated by patterns of neurons firing in a spiking neural network. Semantic pointers can be bound together using circular convolution and unbound using circular correlation (Plate 1994). A Smolensky-style form of concept composition can readily be implemented – with the caveat that these representations and binding and unbinding operations are noisy.

In what follows, we propose a way to view tensor-based compositional distributional semantics within a Smolensky-style framework and thereby propose a way

for tensor-based compositional distributional semantics to be implemented within a spiking neural network architecture.

### 3 Compositional Distributional Semantics in the Nengo Framework

The Nengo framework uses the Semantic Pointer Architecture to represent concepts. A usual proposal for the representation of concepts in this kind of framework is to consider features of a concept as roles, the value of those features as fillers, and form the representation of a concept as a sum of role–filler bindings. However, this then leads to the question of how function words like adjectives and verbs should be represented, and how they might be combined with noun concepts.

#### 3.1 A First Proposal

Recall that a key aspect of vector symbolic architectures is the existence of a binding operator and an unbinding operator. In Smolensky’s ICS, these are respectively tensor product and inner product, and in Plate (1994) these are circular convolution and circular correlation.

In compositional distributional semantics, we also make use of the inner product (or more generally tensor contraction) as a composition operator, and function words such as verbs and adjectives are tensors or matrices, i.e., weighted sums of tensor products of basis vectors. **We therefore have an immediate way of mapping to the semantic pointer architecture needed for implementation in Nengo**, by viewing inner product as an unbinding operator and tensor product as a binding operator.

In a little more detail: given a matrix **red** and a vector **car**, we form the composition **red car** via matrix multiplication. Writing this out explicitly, if we have a noun space  $N$  with basis  $\{\mathbf{e}_i\}_i$ , then

$$\mathbf{car} = \sum_i c_i \mathbf{e}_i, \quad \mathbf{red} = \sum_i r_{ij} \mathbf{e}_i \otimes \mathbf{e}_j \quad (15)$$

$$\mathbf{red car} = \sum_{ij} r_{ij} \mathbf{e}_i \langle \mathbf{e}_j, \mathbf{car} \rangle = \sum_{ijk} r_{ij} c_k \mathbf{e}_i \langle \mathbf{e}_j, \mathbf{e}_k \rangle = \sum_{ij} r_{ij} c_j \mathbf{e}_i \quad (16)$$

i.e., we can think of this operation as unbinding the vector **car** from the adjective **red**.

Now, to move to the semantic pointer setting, we map tensor product to circular convolution and inner product to circular correlation. We view each basis vector as a semantic pointer and encode a noun as a weighted sum of semantic pointers and an adjective as a weighted sum of convolved pairs of semantic pointers:

$$\mathbf{car} = \sum_i c_i \mathbf{p}_i, \quad \mathbf{red} = \sum_i r_{ij} \mathbf{p}_i \otimes \mathbf{p}_j \quad (17)$$

$$\mathbf{red\ car} = \sum_{ij} r_{ij} \mathbf{p}_i (\mathbf{p}_j \otimes \mathbf{car}) = \sum_{ijk} r_{ij} c_k \mathbf{p}_i (\mathbf{p}_j \otimes \mathbf{p}_k) = \sum_{ij} r_{ij} c_j \mathbf{p}_i + \text{noise} \quad (18)$$

The last step in the above relies on the semantic pointers  $\mathbf{p}_i$  being approximately orthogonal, which they are by design.

A toy implementation of this is available at <https://github.com/marthafinderslewis/nengo-disco>. We use the ‘pet fish’ problem as an example. In the pet fish problem, we want the adjective ‘pet’ to modify animals in certain ways. A ‘pet fish’ should modify ‘fish’ to make it similar to a goldfish; however, the representation of ‘cat’ and ‘dog’ should stay pretty similar: cats and dogs are already pretty archetypal pets. To implement a model of the ‘pet fish’ problem, we take inspiration from Coecke and Lewis (2015). We choose some features to describe our animals: *cared-for*, *vicious*, *fluffy*, *scaly*, *lives-house*, *lives-sea*, *lives-zoo*. These are rendered as semantic pointers, and we use the following notation: *cared-for*:  $\mathbf{c}$ , *vicious*:  $\mathbf{v}$ , *fluffy*:  $\mathbf{f}$ , *scaly*:  $\mathbf{s}$ , *lives-house*:  $\mathbf{h}$ , *lives-sea*:  $\mathbf{e}$ , *lives-zoo*:  $\mathbf{z}$ . Each animal is rendered as a weighted sum of semantic pointers with weights as in Table 1. We interpret these weights as the importance of each feature to the noun. Note that vectors are normalised.

We then design an adjective as the following sum of convolved semantic pointers:

$$\mathbf{pet} = \mathbf{c} \otimes (\mathbf{c} + \mathbf{v} + \mathbf{f} + \mathbf{s} + \mathbf{e} + \mathbf{h} + \mathbf{z}) + \mathbf{v} \otimes \mathbf{v} + \mathbf{f} \otimes \mathbf{f} + \mathbf{s} \otimes \mathbf{s} + \mathbf{h} \otimes (\mathbf{h} + \mathbf{e} + \mathbf{z}) \quad (19)$$

which in matrix format looks as in Table 2. We can interpret these weights as follows. The first row of the matrix is essentially saying that no matter what the features of the animal, after application of the *pet* adjective, the animal should be cared for. The next three rows are just identity. The rows corresponding to *lives-sea* and *lives-zoo* are zero: after application of the *pet* adjective, the pet animal should not have any weight on these features. Lastly, we see that the row corresponding to *lives-home* moves weight from other features to this feature.

**Table 1** Weights for semantic pointer representations of nouns

	Fish	Goldfish	Cat	Dog	Shark	Lion
$\mathbf{c}$	0.13	0.44	0.57	0.67	0.00	0.19
$\mathbf{v}$	0.51	0.00	0.13	0.37	0.57	0.62
$\mathbf{f}$	0.00	0.00	0.57	0.37	0.00	0.44
$\mathbf{s}$	0.63	0.62	0.00	0.00	0.57	0.00
$\mathbf{e}$	0.51	0.00	0.00	0.00	0.57	0.00
$\mathbf{h}$	0.19	0.62	0.57	0.52	0.00	0.00
$\mathbf{z}$	0.19	0.19	0.00	0.00	0.11	0.62

**Table 2** Weights for sum of convolved semantic pointers

	<b>c</b>	<b>v</b>	<b>f</b>	<b>s</b>	<b>e</b>	<b>h</b>	<b>z</b>
<b>c</b>	1	1	1	1	1	1	1
<b>v</b>	0	1	0	0	0	0	0
<b>f</b>	0	0	1	0	0	0	0
<b>s</b>	0	0	0	1	0	0	0
<b>e</b>	0	0	0	0	0	0	0
<b>h</b>	0	0	0	0	1	1	1
<b>z</b>	0	0	0	0	0	0	0

In Nengo, the nouns and adjective are implemented as weighted sums of semantic pointers or convolved semantic pointers, and the nouns and adjective are composed using the unbinding mechanism: the nouns are unbound from the adjective. Each adjective–noun combination is then queried against the nouns to retrieve the noun that is most similar. We wish that ‘pet fish’ is most similar to ‘goldfish’, ‘pet cat’ is most similar to ‘cat’, and so on. A video of the system can be seen at <https://github.com/marthaflinderslewis/nengo-disco>.

The above goes to show that compositional distributional semantics can be implemented within the semantic pointer architecture but does not give any indication about how features or weights could be learnt. In the following section, we give an alternative formulation that has the potential to provide a learning mechanism from labelled stimuli.

### 3.2 *Compositional Distributional Semantics as a Role–Filler Model of Meaning*

We now provide a slightly different perspective on compositional distributional semantics within a semantic pointer architecture. As we described in Sect. 2.2, Eq. (2), a semantic representation in ICS consists of a sum of role–filler pairs:

$$\mathbf{s} = \sum_i \mathbf{r}_i \otimes \mathbf{f}_i \quad (20)$$

In order to map this representation to the compositional vector representation, we consider the following. In the case of a noun, we say that the roles  $\mathbf{r}_i$  are a set of basis vectors spanning the noun space, and then the fillers are simply scalars attached to each role.

$$\mathbf{n} = \sum_i n_i \mathbf{r}_i \quad (21)$$

We view an adjective as a set of fillers bound to roles where the roles are possible nouns, and the fillers are vectors corresponding to the adjective–noun combination:

$$\mathbf{adj} = \sum_i \mathbf{a} \mathbf{n}_i \otimes \mathbf{n}_i \quad (22)$$

We view intransitive verbs as a set of fillers bound to roles where the roles are possible nouns and the fillers are the resulting sentences:

$$\mathbf{in-verb} = \sum_i \mathbf{n}_i \otimes \mathbf{sent}_i \quad (23)$$

Transitive verbs are a set of fillers bound to roles where the roles are pairs of possible nouns and the fillers are the resulting sentences:

$$\mathbf{tr-verb} = \sum_{ij} \mathbf{n}_i \otimes \mathbf{sent}_{ij} \otimes \mathbf{n}_j \quad (24)$$

The composition of an adjective and a noun is then found by unbinding the noun role from the adjective, and similarly for verb–noun composition. In the adjective–noun example, unbinding is just matrix–vector multiplication. For a very toy example, suppose we have some kind of vector representations of:  $\mathbf{red\ car} = \mathbf{rc}$ ,  $\mathbf{red\ apple} = \mathbf{ra}$ ,  $\mathbf{red\ wine} = \mathbf{rw}$ ,  $\mathbf{car} = \mathbf{c}$ ,  $\mathbf{apple} = \mathbf{a}$ ,  $\mathbf{wine} = \mathbf{w}$ .

Then,

$$\mathbf{red} = \mathbf{rc} \otimes \mathbf{c} + \mathbf{ra} \otimes \mathbf{a} + \mathbf{rw} \otimes \mathbf{w} \quad (25)$$

Computing  $\mathbf{red\ car}$  as  $\mathbf{red} \cdot \mathbf{c}$ , we obtain

$$\mathbf{red} \cdot \mathbf{c} = \mathbf{rc} \langle \mathbf{c}, \mathbf{c} \rangle + \mathbf{ra} \langle \mathbf{a}, \mathbf{c} \rangle + \mathbf{rw} \langle \mathbf{w}, \mathbf{c} \rangle \quad (26)$$

$$= \mathbf{rc} + \text{noise} \quad (27)$$

assuming that cars are not very similar to apples or wine.

Recall that in Eqs. (17) and (18) we argued that in order to implement a distributional semantic model within the spiking neural architecture, we could map from tensor product as binding operator and inner product as unbinding operator, to circular convolution as binding operator and circular correlation as unbinding operator. We use the same methodology to obtain a representation of nouns, adjectives, and verbs within the semantic pointer architecture. For example, in the semantic pointer architecture,

$$\mathbf{red} = \mathbf{rc} \circledast \mathbf{c} + \mathbf{ra} \circledast \mathbf{a} + \mathbf{rw} \circledast \mathbf{w} \quad (28)$$

and

$$\mathbf{red} \oslash \mathbf{c} = \mathbf{rc} \langle \mathbf{c} \oslash \mathbf{c} \rangle + \mathbf{ra} \langle \mathbf{a} \oslash \mathbf{c} \rangle + \mathbf{rw} \langle \mathbf{w} \oslash \mathbf{c} \rangle \quad (29)$$

$$= \mathbf{rc} + \text{more noise} \quad (30)$$

again, assuming that cars are not very similar to apples or wine.

This perspective on compositional distributional semantics as a role–filler binding architecture also gives us a potential way of learning representations for words.

### 3.3 Learning Strategy

We consider adjective–noun composition. Suppose we have a set of images labelled with adjective–noun combinations, and assume that within our semantic pointer architecture we have some kind of vision system that can produce semantic pointers for the images themselves.

**Supervised Learning Situation** We have a set of labelled inputs. Let us assume they are all of the form adj–noun. Suppose the system has a convolved semantic pointer representation of each adjective it has learnt so far, call them  $\{\mathbf{A}_i\}_i$ , and a semantic pointer representation of each noun it has learnt so far, call them  $\{\mathbf{n}_i\}_i$ . Suppose an input is labelled  $A_j n_k$  and the system has both these words in its vocabulary. We assume the system has a vector representation  $\mathbf{an}_i$  of each image.

If we assume that the adjective has the form  $\mathbf{adj} = \sum_i \mathbf{an}_i \otimes \mathbf{n}_i$ , then we get a simple update rule for the adjective, by mixing the current adjective with the convolution of  $\mathbf{an}$  and  $\mathbf{n}$ :

$$\mathbf{A}_j \mapsto (1 - h)\mathbf{A}_j + h\mathbf{an}_i \otimes \mathbf{n}_k \quad (31)$$

where  $h$  is some small value in  $[0, 1]$ . In order to update the noun, we propose unbinding the  $\mathbf{an}$  filler from  $\mathbf{A}$  to get the  $\mathbf{n}$  role, giving us

$$\mathbf{n}_k \mapsto (1 - h)\mathbf{n}_k + h(\mathbf{an}_i \oslash \mathbf{A}_j) \quad (32)$$

In the cases where the system does not yet have a representation of the adjective, we can initialise it as  $\mathbf{an} \otimes \mathbf{an}$ , and where there is no representation of the noun, it can be initialised as  $\mathbf{an}$ .

We can make a similar proposal for intransitive verbs. We assume that the verb has the form  $\mathbf{verb} = \sum_i \mathbf{n}_i \otimes \mathbf{nv}_i$ , where  $\mathbf{nv}$  is a vector representing an intransitive sentence like ‘Junpa walks’. Given an input labelled  $n_j V_k$  corresponding to a vector representation  $\mathbf{nv}_i$ , the verb  $V_k = \sum_i \mathbf{n}_i \otimes \mathbf{nv}_i$  is then updated by

$$\mathbf{V}_k \mapsto (1 - h)\mathbf{V}_k + h\mathbf{n}_j \otimes \mathbf{nv}_i \quad (33)$$

and the noun is updated by

$$\mathbf{n}_j \mapsto (1 - h)\mathbf{n}_j + h(\mathbf{V}_k \oslash \mathbf{nv}_i) \quad (34)$$

We have started to investigate the approach outlined above within a standard neural network model to learn compositional word representations from labelled images (Lewis et al. 2023), and future work will go on to extend this to the Nengo architecture.

## 4 Conclusions and Future Work

Compositional distributional semantics has a set of powerful machinery that can be used for composition. However, it does not have any particular cognitive grounding. In this chapter, we have given a proposal for implementation of compositional distributional semantics within the cognitive architecture Nengo. This architecture uses a biologically (more) realistic substrate to represent concepts as semantic pointers and is integrated with decision-making, vision, and other modules. This therefore has the potential to provide compositional distributional semantics with an environment in which meanings can be grounded. We have given a toy implementation to show that our proposal is possible and presented an alternative formulation of the approach together with a strategy for learning word representations.

Future work in this area is of course to take forward possible implementations of these ideas. We already have a strategy to begin implementation within the Nengo framework. Once implementation within this kind of architecture has been carried out, there is potential to examine what kind of representation best model human behavior – whether compositional distributional semantics is a useful representation. We gave arguments in Sect. 2.2 to argue that there is a need for more flexible composition than additive binding, so there is potential here.

We would also like to implement these ideas within a dialogue setting. While a supervised learning setting has been described, the kinds of representation proposed are very amenable to being learnt in a self-supervised fashion between two or more agents.

## References

- Diederik Aerts. Quantum structure in cognition. *Journal of Mathematical Psychology*, 53(5): 314–348, October 2009. ISSN 00222496. <https://doi.org/10.1016/j.jmp.2009.04.005>. <https://linkinghub.elsevier.com/retrieve/pii/S0022249609000558>.
- Diederik Aerts, Jan Broekaert, Liane Gabora, and Tomas Veloz. The Guppy Effect as Interference. In Jerome R. Busemeyer, François Dubois, Ariane Lambert-Mogiliansky, and Massimo Melucci, editors, *Quantum Interaction*, Lecture Notes in Computer Science, pages 36–47, Berlin, Heidelberg, 2012. Springer. ISBN 978-3-642-35659-9. [https://doi.org/10.1007/978-3-642-35659-9\\_4](https://doi.org/10.1007/978-3-642-35659-9_4).
- Marco Baroni and Roberto Zamparelli. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1183–1193, USA, October 2010. Association for Computational Linguistics.



- Jean-Philippe Bernardy and Stergios Chatzikyriakidis. What Kind of Natural Language Inference are NLP Systems Learning: Is this Enough? In *Special Session on Natural Language Processing in Artificial Intelligence*, volume 2, pages 919–931. SCITEPRESS, February 2019. ISBN 978-989-758-350-6. <https://doi.org/10.5220/0007683509190931>. <http://www.scitepress.org/Papers/2019/76835>.
- Peter Blouw, Eugene Solodkin, Paul Thagard, and Chris Eliasmith. Concepts as Semantic Pointers: A Framework and Computational Model. *Cognitive Science*, page 35, 2016.
- Reinhard Blutner, Emmanuel M. Pothos, and Peter Bruza. A Quantum Probability Perspective on Borderline Vagueness. *Topics in Cognitive Science*, pages n/a–n/a, September 2013. ISSN 17568757. <https://doi.org/10.1111/tops.12041>. <http://doi.wiley.com/10.1111/tops.12041>.
- Gemma Boleda. Distributional Semantics and Linguistic Theory. *arXiv:1905.01896 [cs]*, March 2020. <https://doi.org/10.1146/annurev-linguistics-011619-030303>. <http://arxiv.org/abs/1905.01896>.
- Samuel R. Bowman, Christopher Potts, and Christopher D. Manning. Recursive Neural Networks Can Learn Logical Semantics. In *Proceedings of the 3rd Workshop on Continuous Vector Space Models and Their Compositionality*, pages 12–21, Beijing, China, July 2015. Association for Computational Linguistics. <https://doi.org/10.18653/v1/W15-4002>. <https://www.aclweb.org/anthology/W15-4002>.
- P. D. Bruza, K. Kitto, B. Ramm, L. Sitbon, D. Song, and S. Blomberg. Quantum-like non-separability of concept combinations, emergent associates and abduction. *Logic Journal of the IGPL*, 20(2):445–457, April 2012. ISSN 1367-0751. <https://doi.org/10.1093/jigpal/jzq049>. <https://doi.org/10.1093/jigpal/jzq049>.
- Ryan Calmus, Benjamin Wilson, Yukiko Kikuchi, and Christopher I. Petkov. Structured sequence processing and combinatorial binding: Neurobiologically and computationally informed hypotheses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 375(1791):20190304, February 2020. <https://doi.org/10.1098/rstb.2019.0304>. <https://royalsocietypublishing.org/doi/full/10.1098/rstb.2019.0304>.
- Bob Coecke and Martha Lewis. A Compositional Explanation of the Pet Fish Phenomenon. *arXiv:1509.06594 [cs, math]*, September 2015. <http://arxiv.org/abs/1509.06594>.
- Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. Mathematical Foundations for a Compositional Distributional Model of Meaning. *arXiv:1003.4394 [cs, math]*, March 2010. <http://arxiv.org/abs/1003.4394>.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1423>. <https://aclanthology.org/N19-1423>.
- Leonidas A. A. Dumas and John E. Hummel. Computational models of higher cognition. In *The Oxford Handbook of Thinking and Reasoning*, Oxford Library of Psychology, pages 52–66. Oxford University Press, New York, NY, US, 2012. ISBN 978-0-19-973468-9.
- C. Eliasmith, T. C. Stewart, X. Choo, T. Bekolay, T. DeWolf, Y. Tang, and D. Rasmussen. A Large-Scale Model of the Functioning Brain. *Science*, 338(6111):1202–1205, November 2012. ISSN 0036-8075, 1095-9203. <https://doi.org/10.1126/science.1225266>. <https://www.sciencemag.org/lookup/doi/10.1126/science.1225266>.
- Chris Eliasmith. *How to Build a Brain: A Neural Architecture for Biological Cognition*. Oxford University Press, 2013. ISBN 978-0-19-934523-6. <https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780199794546.001.0001/acprof-9780199794546>.
- Ross W Gayler. Vector Symbolic Architectures Answer Jackendoff’s Challenges for Cognitive Neuroscience. In *Joint International Conference on Cognitive Science*, page 6, July 2003. <https://arxiv.org/abs/cs/0412059>.
- Edward Grefenstette and Mehrnoosh Sadrzadeh. Experimental Support for a Categorical Compositional Distributional Model of Meaning. In *Proceedings of the 2011 Conference on Empirical*

- Methods in Natural Language Processing*, pages 1394–1404, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics. <https://aclanthology.org/D11-1129>.
- Edward Grefenstette, Mehrnoosh Sadrzadeh, Stephen Clark, Bob Coecke, and Stephen Pulman. Concrete Sentence Spaces for Compositional Distributional Models of Meaning. *arXiv:1101.0309 [cs]*, December 2010. <http://arxiv.org/abs/1101.0309>.
- John E. Hummel. Getting symbols out of a neural architecture. *Connection Science*, 23(2):109–118, June 2011. ISSN 0954-0091. <https://doi.org/10.1080/09540091.2011.569880>. <https://doi.org/10.1080/09540091.2011.569880>.
- Martha Lewis. Quantum Computing and Cognitive Simulation, March 2021. <https://psyarxiv.com/hvbgf/>.
- Martha Lewis, Nihal V. Nayak, Peilin Yu, Qinan Yu, Jack Merullo, Stephen H. Bach, and Ellie Pavlick. Does CLIP Bind Concepts? Probing Compositionality in Large Image Models, March 2023. <http://arxiv.org/abs/2212.10537>.
- Andrea E. Martin and Leonidas A. A. Doumas. Tensors and compositionality in neural systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 375(1791):20190306, February 2020. <https://doi.org/10.1098/rstb.2019.0306>. <https://royalsocietypublishing.org/doi/10.1098/rstb.2019.0306>.
- Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics. <https://doi.org/10.18653/v1/P19-1334>. <https://www.aclweb.org/anthology/P19-1334>.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*, September 2013. <http://arxiv.org/abs/1301.3781>.
- Jeff Mitchell and Mirella Lapata. Composition in Distributional Models of Semantics. *Cognitive Science*, 34(8):1388–1429, November 2010. ISSN 03640213. <https://doi.org/10.1111/j.1551-6709.2010.01106.x>. <http://doi.wiley.com/10.1111/j.1551-6709.2010.01106.x>.
- Denis Paperno, Nghia The Pham, and Marco Baroni. A practical and linguistically-motivated approach to compositional distributional semantics. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 90–99, Baltimore, Maryland, June 2014. Association for Computational Linguistics. <https://doi.org/10.3115/v1/P14-1009>. <https://www.aclweb.org/anthology/P14-1009>.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, 2014. Association for Computational Linguistics. <https://doi.org/10.3115/v1/D14-1162>. <http://aclweb.org/anthology/D14-1162>.
- Tony A. Plate. *Distributed Representations and Nested Compositional Structure*. PhD thesis, 1994.
- Emmanuel M. Pothos and Jerome R. Busemeyer. Can quantum probability provide a new direction for cognitive modeling? *Behavioral and Brain Sciences*, 36(3):255–274, June 2013. ISSN 0140-525X, 1469-1825. <https://doi.org/10.1017/S0140525X12001525>. [https://www.cambridge.org/core/product/identifier/S0140525X12001525/type/journal\\_article](https://www.cambridge.org/core/product/identifier/S0140525X12001525/type/journal_article).
- Paul Smolensky. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, 46(1-2):159–216, November 1990. ISSN 00043702. [https://doi.org/10.1016/0004-3702\(90\)90007-M](https://doi.org/10.1016/0004-3702(90)90007-M). <https://linkinghub.elsevier.com/retrieve/pii/000437029090007M>.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive Deep Models for Semantic Compositionality over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. <https://www.aclweb.org/anthology/D13-1170>.
- Aarne Talman and Stergios Chatzikyriakidis. Testing the Generalization Power of Neural Network Models across NLI Benchmarks. In *Proceedings of the 2019 ACL Workshop*

*BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 85–94, Florence, Italy, August 2019. Association for Computational Linguistics. <https://doi.org/10.18653/v1/W19-4810>. <https://aclanthology.org/W19-4810>.

Dominic Widdows, Kirsty Kitto, and Trevor Cohen. Quantum Mathematics in Artificial Intelligence. *Journal of Artificial Intelligence Research*, 72:1307–1341, December 2021. ISSN 1076-9757. <https://doi.org/10.1613/jair.1.12702>. <https://www.jair.org/index.php/jair/article/view/12702>.

# Optimality, Prototypes, and Bilingualism



Igor Douven and Galina V. Paramei

**Abstract** There is an old debate about the status of basic color categories (such as blue, green, and red), the question being whether that status derives from the way the world is or whether it is culture-bound. In more scientific terminology, this amounts to the question of whether the categorical structure of color space (how it is carved up and where the color prototypes or foci are located) is fixed by the world or a matter of cultural conventions. Some recent work suggests that the categorical structure of color space is not a conventional matter without, however, being completely determined by the world; it is also subject to constraints deriving from the various ways in which our perceptual and cognitive capacities are limited. While there is recent evidence for this newer position, we report a study comparing Italian monolingual, English monolingual, and Italian–English bilingual speakers with regard to focal color choices in the BLUE region of color space suggesting that cultural and linguistic factors play a role in the categorical structuring of color space.

**Keywords** Color space · Cultural dependence · Color perception · Bilingual · Prototype

## 1 Introduction

Philosophers, psychologists, and linguists have long been debating the nature of natural kind terms. According to some theorists (called “realists”), these terms capture, or at least are meant to capture, the in-built structure of reality, while others (“nominalists”) maintain that which terms we regard as picking out natural

---

I. Douven (✉)  
IHPST/CNRS/Panthéon–Sorbonne University, Paris, France

G. V. Paramei  
Department of Psychology, Liverpool Hope University, Liverpool, UK

properties or concepts are essentially conventional, or at most motivated by pragmatic concerns. A less well-known and in a way middle position—sometimes called “conceptualism”—conceives of the world as imposing constraints on the terms we use to talk and think but maintains that, at the same time, there are constraints deriving from our particular cognitive makeup.

Recently, evidence for conceptualism has emerged, at least insofar as the position applies to the color domain, which has long been regarded as providing us with prime examples of natural kind terms (“blue,” “green,” “red,” etc.). Most notably, Regier et al. (2007) offer evidence for Jameson and D’Andrade’s (1997; also Jameson 2005) so-called Interpoint-Distance Model, according to which color concepts (i.e., denotative meanings of color terms) stem from a combination of the irregularities to be found in perceptual color space (i.e., color similarity space, which is often taken to be CIELAB space; see Fairchild 2013, and below) and a preference for informative naming systems, where this preference is perceptually and cognitively motivated, specifically, in terms of how our limited perceptual and cognitive capacities favor adding color categories, so that color differences between adjacent categories get maximized, while color differences within the new contiguous categories are minimized.

Regier and colleagues’ (2007) work inspired Douven and Gärdenfors (2020) to propose that natural concepts (i.e., the meanings of natural kind terms) are given by the cells of optimally partitioned similarity spaces (such as color space when the kind terms concern color categories), where the notion of an optimal partition is defined by reference to principles of good engineering, that is, principles a good engineer would want to respect in designing a system of concepts for creatures with our perceptual and cognitive makeup. For instance, according to the principle Douven and Gärdenfors call “Representation,” an optimal partition will allow prototypes (Rosch 1973)—in the case of basic color categories referred to as “foci” or “focal colors”—to be placed such that each prototype is a good representative of all items falling under the concept. And according to the principle these authors call “Contrast,” an optimal partition will allow prototypes of different concepts to be so chosen that they are easily distinguishable from each other. These and the other principles Douven and Gärdenfors propose are jointly meant to lead to conceptual systems that facilitate learning and memorization and help avoid classification errors. As Douven and Gärdenfors (2020) show, evidence that at least some of these principles have indeed been at work in shaping our concepts, in particular our color concepts, is already to be found in the literature (e.g., Jameson 2005; Kemp and Regier 2012; Xu and Regier 2014; Xu et al. 2016; see Douven 2019, specifically for evidence for Representation and Contrast in the context of color categorization).

But while there is evidence for a version of conceptualism, some of this same evidence also suggests that conceptualism may not be the whole story. For instance, Regier and co-authors (2007) show that their computational model of Jameson and D’Andrade’s hypothesis does remarkably well in predicting color concepts in the languages with up to and including six basic color terms (BCTs; see below). As shown in Jraissati and Douven (2017), however, the same model does worse for

languages with more than six BCTs, even much worse for languages with 11 BCTs, such as English (see also Douven 2017). Something similar holds for evidence documented in Douven (2019). Douven looked at constellations of 11 possible color prototypes that do best, on balance, at satisfying Representation and Contrast. Since there is no unique best trade-off between these principles, the said constellations form what is known as a “Pareto front.”<sup>1</sup> And Douven (2019) found that, while the actual constellation of prototypes lies close to that front, it does not quite lie *on* the front.

To be sure, these discrepancies could just be a matter of not having discovered all principles of good engineering that define optimality in Douven and Gärdenfors’ (2020) proposal. For instance, the Pareto front in Douven (2019) corresponds to the best trade-off of two desiderata for optimality. And while the actual constellation of prototypes was not represented as a point *on that* front, it might well be on a front that results from taking into consideration further desiderata, perhaps ones yet to be discovered. Similarly, an extension of Regier and colleagues’ (2007) computational model that also incorporates criteria beyond informativeness (where we imagine these further criteria to be equally motivated by reference to our perceptual and cognitive capacities) might be able to achieve the same good fit for languages with more than six BCTs that the extant model achieves for languages with up to and including six BCTs.

Another possibility, which is not necessarily inconsistent with the previous one, is the presence of more local or contingent effects on how we use color terms to carve up and furnish (by placing prototypes) color space, and more generally on how we use natural kind terms to carve up and furnish the relevant similarity spaces.<sup>2</sup> Among such effects could be primacy effects (items encountered early on in the process of concept acquisition shape our concepts more than ones encountered later), recency effects (items encountered most recently have a relatively larger impact on the shape of our concepts), a combination of the two, and, crucially, cultural and linguistic effects, specifically, exposure to rival or partly rival conceptual systems as associated with other cultures or languages we happen to be acquainted with (see Ervin 1961; Caskey-Sirmons and Hickerson 1977).<sup>3</sup>

---

<sup>1</sup> Formally, the Pareto front is the curve or surface (or hyper-surface) in the space of possible solutions to a multi-objective optimization problem such that, for any solution represented on the curve/surface, one can only improve with regard to one of the given objectives by doing worse with regard to one or more of the others.

<sup>2</sup> It is an open question whether all concepts can be represented in similarity spaces. We will leave this broader question aside here, our main interest being in the categorical structure of color space. For more on the broader program of using similarity spaces to represent concepts, see Gärdenfors (2000).

<sup>3</sup> An anonymous referee made the important observation that yet another possibility (compatible with the ones mentioned) is that, because color space will have evolved over time, and is thus not to be conceived as the product of an optimization process from scratch, there have been anchoring effects of older, less fine-grained partitions of the space, which have served as a kind of starting point for the optimization process that eventually led to the space as we know it, partitioned into eleven basic color concepts.

In this chapter, we look at the second possibility, in particular, at the possible effects of exposure to partly different conceptualizations of the color domain in bilingual Italian–English speakers. Specifically, we analyze a data set from a study with such speakers, as well as with monolingual English and monolingual Italian speakers to investigate the influence of immersion into a non-native language and cultural factors on the placement of prototypes for color terms as well as on concept extensions in color space. Still more specifically, our study focused on the BLUE region of color space, given that Italian speakers are known to require more than one BCT to name the blue colors. Thus, if cultural factors impact conceptualization, one might expect to find in Italian–English bilinguals some interaction between the different conceptualizations of the BLUE region associated with English and Italian, where this interaction might be revealed by comparing where in the BLUE region the bilingual speakers locate focal colors with where monolingual Italian and monolingual English speakers do. Before we present the study, we provide some background on the “Italian blues” in relation to the dominant view of BCTs.

## 2 Italian Blues

Berlin and Kay (1969/1991, p. 6) define a color term as basic if it is monolexicemic, not included in any other BCTs, applied not only to a limited class of objects, and is psychologically salient for all informants. According to the Berlin and Kay model, languages with a developed color lexicon have maximally 11 BCTs. But the BCT upper limit tenet has been questioned in recent years (see Paramei and Bimler 2021). In particular, the BLUE region of color space was demonstrated to require two BCTs in a number of languages, with the two terms differentiating light and dark(er) shades of blue. The classic examples are Russian *sinij* “dark blue” and *goluboj* “light blue,” with the latter named by Berlin and Kay (1969/1991) as a potential 12th BCT. Recent reviews of linguistic and psycholinguistic studies of the two “Russian blues” underscored Berlin and Kay’s conjecture Paramei (2005, 2007).

The “BLUE challenge” encompasses also Italian. Earlier linguistic studies provided evidence of *azzurro*, *blu*, and *celeste* as salient terms in both spoken language (Giacalone Ramat 1978) and written language (Grossmann 1988) across various Italian dialects and, thus, as BCT candidates.

*Azzurro* is deeply entrenched in Italian and has been attested already in the ninth century, originally having denoted lapis lazuli (Frison and Brun 2016). According to De Mauro (1983), initially *azzurro* belonged solely to the written language and was absent in Italian dialects; it entered the spoken language after the political unification of Italy during the *Risorgimento* (1815–1871), with school education having become affordable to the general population. *Celeste* originates from Latin *caeruleus*, derived from *caelum* “sky,” and at the time did not have a color meaning. With a symbolic religious meaning, but also with a color sense, it is attested in the thirteenth century, denoting light shades of blue and being common in Italian dialects (Grossmann 1988). *Blu* was the last of the “blues” to enter Italian, at the

end of the seventeenth century, and was deployed to lexicalize deep (dark/navy) blue, conceivably as a result of its use in the cloth trade (Pastoureau 2001, p. 127). Nowadays, it is considered to be the most widespread “blue” term in Italian (Sandford 2015).

In addition to the linguistic studies addressed above, recent psycholinguistic studies carried out in different regions of Italy provide converging evidence that in Italian at least two BCTs are required for naming the BLUE region (e.g., Bimler and Uusküla 2014, 2018; Paramei et al. 2014, 2018; Uusküla 2014). Whether two suffices or whether three are needed is still a matter of controversy.

Across all linguistic and psycholinguistic studies, the authors are unanimous that *blu* is the counterpart of English “dark/navy blue.” However, the second Italian BCT is argued to be either *azzurro* (Paggetti et al. 2016) or *celeste* (Paramei et al. 2018); in particular, in some speakers’ opinion, *azzurro* denotes a shade in between *celeste* “light-blue” and *blu* “dark blue” (Grossmann 1988), whereas for others its meaning is similar to *celeste* and both are in opposition to *blu* (Albertazzi and Da Pos 2017).

The degree of use of the three terms in the spoken language is known to be subject to diatopic, diastratic, and diaphasic variation (e.g., Paramei et al. 2018). In particular, for Verona speakers (Veneto region), *azzurro* is the second BCT denoting the BLUE region and corresponds to “light-and-medium blue” (Paramei et al. 2014, 2018). In comparison, for Alghero speakers of the Catalan–Algherese dialect (Sardinia), *celeste* appears to be the second BCT and denotes the light and medium shades of blue, while *azzurro*, with the meaning of “dark medium blue,” is apparently not a basic term there (Paramei et al. 2018). Particularly relevantly for our study (given the provenance of the monolingual Italian participants we used for this chapter; see below), for Florence speakers (Tuscany) the BLUE region is “clothed in triple blues”: *blu* denotes “dark/navy blue,” *azzurro* “medium blue,” and *celeste* is reserved for “light blue” (Bimler and Uusküla 2014; Del Viva et al. 2022).

### 3 Study

We were interested in the effect of bilingualism on focal color judgments in the BLUE region of color space as this might provide evidence for cultural and/or linguistic effects on the placement of prototypes in color space. Studies on the effect of bilingualism on color conceptualization are sparse, but the few that exist (Ervin 1961; Caskey-Sirmons and Hickerson 1977; Athanasopoulos 2009; Paramei et al. 2016) suggest that, for Italian–English bilinguals, focal color choices in relation to color terms in one language could exert an effect on focal color choices in relation to color terms in the other language. The goal of the present chapter was to look for such an effect in comparing the responses from English monolingual speakers, Italian monolingual speakers, and Italian–English bilingual speakers to questions about focal colors in the BLUE region of color space.



## 3.1 Method

### 3.1.1 Participants

There were 92 participants in total, constituting three groups: British English monolingual speakers (EN), Italian monolingual speakers (IT), and Italian–English bilingual speakers (BI):

- (i) *EN speakers* ( $N = 32$ ;  $F = 20$ ) were from West-North England, aged 30.1 ( $\pm 8.9$ ) years. All had a college education, some at a graduate level, or obtained a doctoral degree. Their self-reported English-language proficiency scores were well above 9 for all categories (reading, writing, speaking, listening), on a scale from 1 to 10, with 10 being the highest.
- (ii) *IT speakers* ( $N = 31$ ;  $F = 17$ ) were students or graduates of the University Florence, born in Tuscany, 22.9 ( $\pm 2.1$ ) years old (for detailed demographic characteristics, see Del Viva et al. 2022). Italian-language proficiency data were not available.
- (iii) *BI speakers* ( $N = 30$ ;  $F = 16$ ) were aged 35 ( $\pm 9.8$ ) years. All but one of them were born in Italy, and all but one resided in the United Kingdom at the time they participated in the study. Apart from one, all BIs had a university education, including 19 holding a doctoral degree. The majority of participants were originally from Central Italy, as well as from Lombardy, Tuscany, Puglia, and from Sardinia; one early bilingual was born in North-West England.

The mean age at which they had started to learn English was 9.8 ( $\pm 5.6$ ) years, with the majority being late bilinguals ( $N_{\text{BI}} = 28$ ), which are bilinguals who started learning English under the age of 6 years (Wattendorf and Festman 2008). The average age from which the BIs had lived in an English-speaking country was 23.5 ( $\pm 8.0$ ); the duration of their residence in an English-speaking country was on average 7.6 ( $\pm 8.8$ ) years. At the time they participated in the study, BIs used English on average 72.9% ( $\pm 23.4\%$ ) of the time. Their average self-reported English proficiency scores, on a scale from 1 to 10, were: 8.3 ( $\pm 1.0$ ) for reading, 7.7 ( $\pm 1.3$ ) for writing, 7.5 ( $\pm 1.2$ ) for speaking, and 7.8 ( $\pm 1.3$ ) for listening. Their average proficiency scores for Italian were, respectively, 9.7 ( $\pm 0.7$ ), 9.2 ( $\pm 0.9$ ), 9.5 ( $\pm 0.9$ ), and 9.8 ( $\pm 0.6$ ). All participants completed the Nation Vocabulary Test (Nation 1980), scoring an average of 80.2 ( $\pm 8.2$ ) points out of 90.

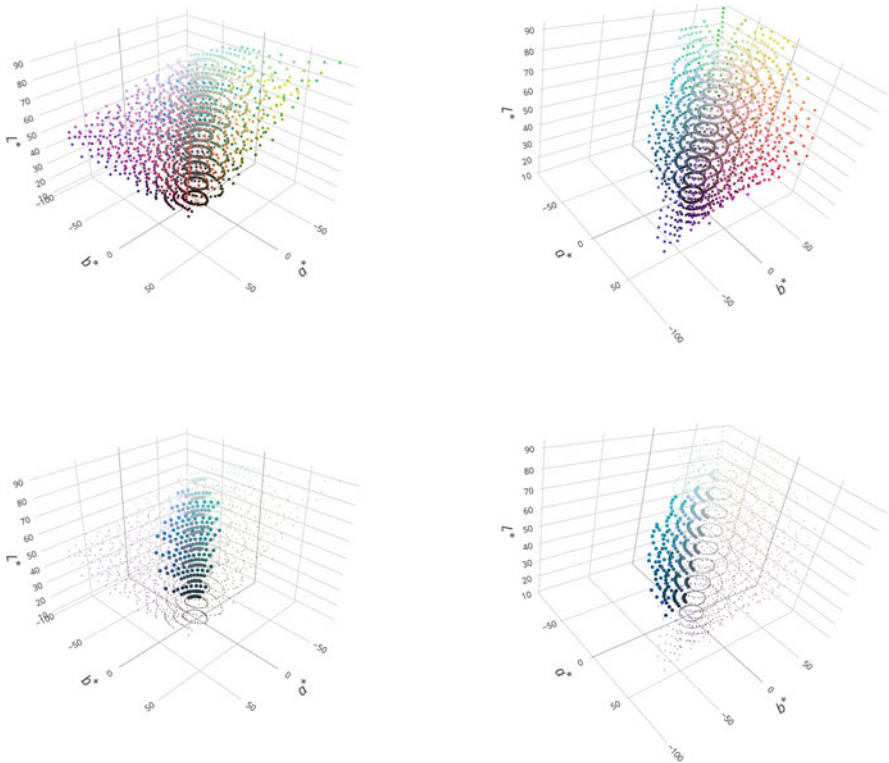
All participants had normal color vision tested with the Ishihara Pseudoisochromatic Plates (Ishihara 1973) and normal or corrected-to-normal visual acuity. None had reported any ocular disease, eye surgery, diabetes, or use of a medication that could have affected color vision.

Ethical approval was obtained from the Ethics Committees of the Departments of Psychology of Liverpool Hope University and the University of Florence. The study followed the ethical principles of the Declaration of Helsinki. All participants gave their written informed consent prior to participation in the study.

### 3.1.2 Materials

As stimuli served the 237 Munsell chips from eight glossy charts from *The Munsell Book of Color* (Munsell 1941), to wit, 7.5BG, 10BG, 2.5B, 5B, 7.5B, 10B, 2.5PB, and 5PB; together these charts encompass the BLUE region in Munsell color space. Value of the Munsell chips varied between 2 and 9, and Chroma varied (even number notation) between 2 and 10 or, for the charts 10B, 2.5PB, and 5PB, between 2 and 12.

For the visualizations and, also, for all of the statistics we carried out, we assumed the CIELAB coordinates of the Munsell chips, as online available at the website of the Munsell Color Science Laboratory from the Rochester Institute of Technology (RIT). While only a limited number of Munsell chips were used as stimuli (representing the BLUE region) in the study, Fig. 1 gives readers unfamiliar with CIELAB space a sense of its shape by depicting the CIELAB locations of all Munsell chips (top row). The bottom row of that figure highlights the stimuli that were used in the study.



**Fig. 1** Top row showing CIELAB space from different angles, with all Munsell chips from the RIT Munsell Color Science Laboratory placed into the space; bottom row highlighting the chips that were used as stimuli in the study

### 3.1.3 Procedure

At both testing locations, in Liverpool and Florence, we ensured identical procedure and illumination conditions. Participants were adapted to mesopic lighting in an otherwise dark room for at least 10 minutes, the temporal window ensuring dark adaptation of cones (Pirenne 1962). Following this, the charts were presented in a viewing booth under D65-metameric illumination (Just Normlicht Mini 5000; Fa. Color Confidence) suspended 40 cm above the chart and delivering a  $30 \times 25 \text{ cm}^2$  light area. At the chart surface, luminance was  $220 \text{ cd/m}^2$  (measured by the PR-650 SpectaScan Colorimeter; Photo Research, Inc.), corresponding to an illuminance of 1387 lux.

The response sheets consisted of a replica of the chip layout of the eight Munsell charts that constituted our materials. Participants were asked to write color names within white cells corresponding to the chart chips. The charts were presented one-by-one in a fixed order (as used in the description of the Materials). An unconstrained color-naming method was used for naming the Munsell chips, which, along with monolexic hue terms, also allowed to use compound, modified, or suffixed terms. Participants were further asked to indicate on the charts the most representative colors for *azzurro*, *blu*, and *celeste* in Italian or for *blue* and *light blue* in English.<sup>4</sup> Bilinguals were tested on separate days in a counter-balanced order of Italian and English sessions. The experimenter provided instructions and communicated with the participant in the language corresponding to the session.

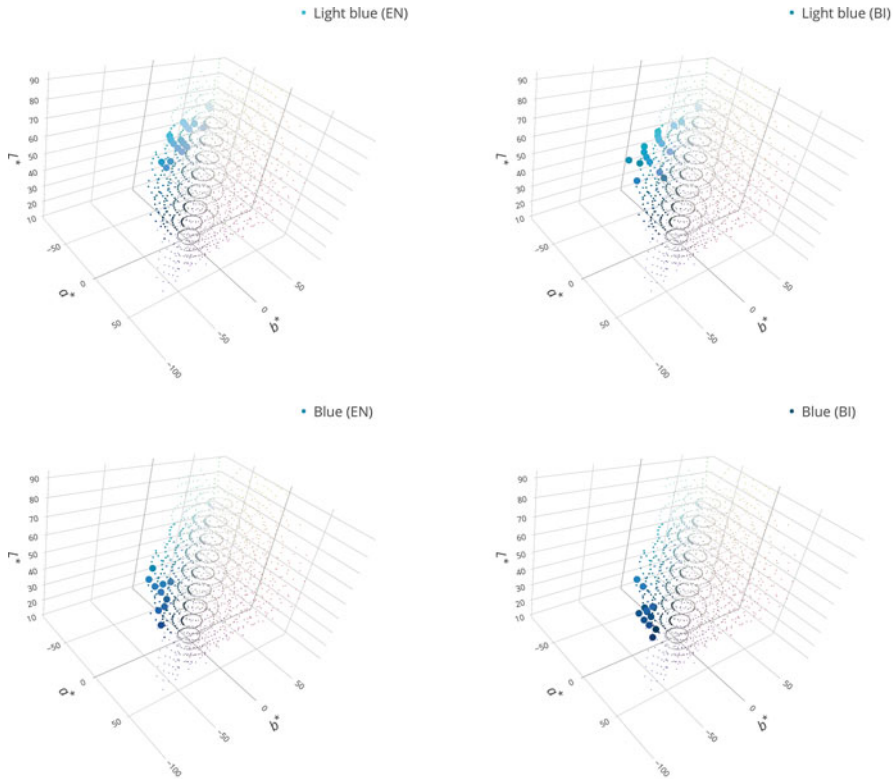
## 3.2 Results and Discussion

The data were gathered in the context of a broader comparison of color naming by bilingual Italian–English speakers. For our present purposes, only focal color responses matter; an analysis of the full data from the free-naming task is relegated to future work. Figures 2 and 3 give a visual summary of the focal color choices, the former exhibiting the *blue* and *light blue* foci, as designated by both the English monolinguals and the Italian–English bilinguals, and the latter doing the same for the Italian monolinguals’ and Italian–English bilinguals’ *azzurro*, *blu*, and *celeste* choices.

As a first step in our analysis, we used the `QHull.jl` library for the Julia language (Bezanson et al. 2017) to compute, per group of participants, convex hulls for their choices for each of the aforementioned color terms. Figure 4 shows the hulls for the bilinguals’ *blue* and *light blue* focal color choices. Then, using the same

---

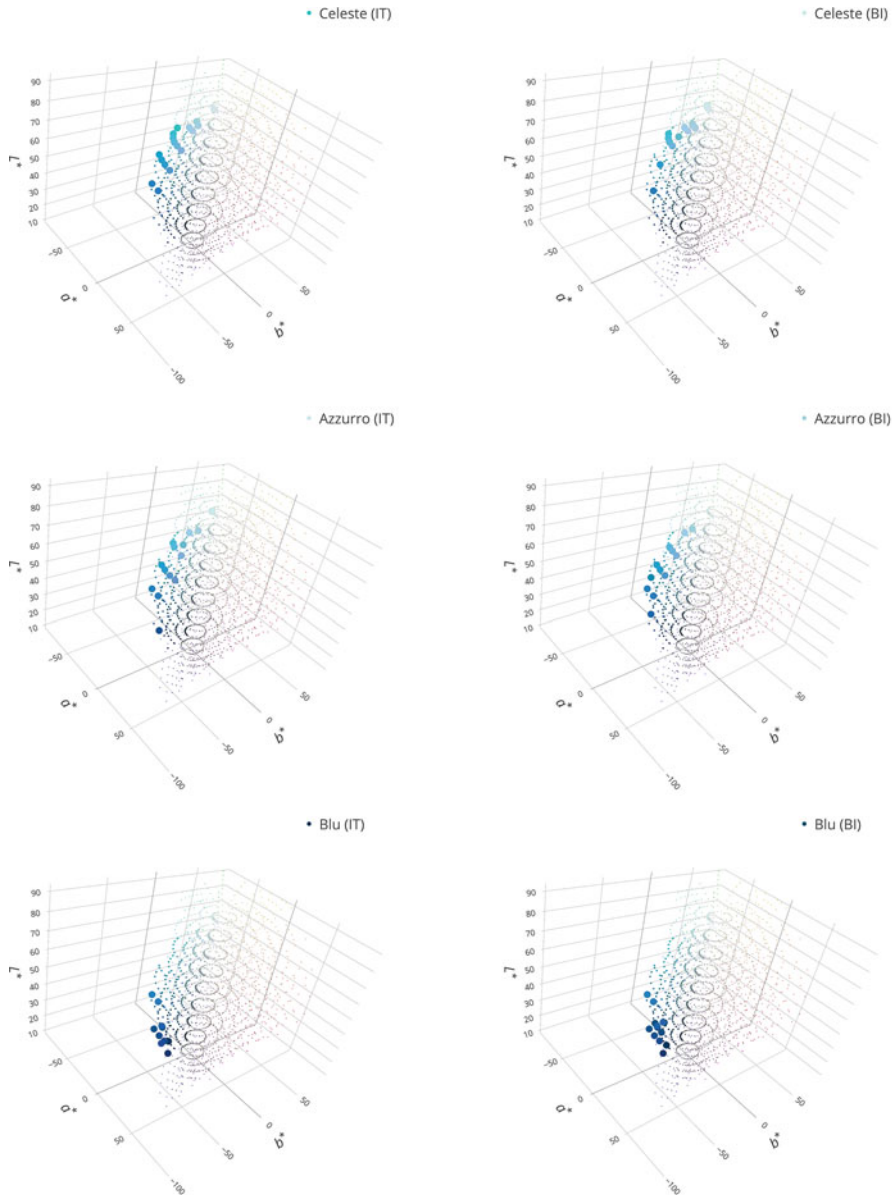
<sup>4</sup> We are aware that *light blue* is not a BCT in English; however, we decided to assess focal color of this non-basic category as a proxy to Italian *celeste*. Note also that *light blue* was found to be among the most frequently used English color terms in an online free-naming experiment (Jraissati and Douven 2018).



**Fig. 2** Focal *light blue* (top row) and *blue* (bottom row) choices of EN monolinguals (left column) and bilinguals (right column)

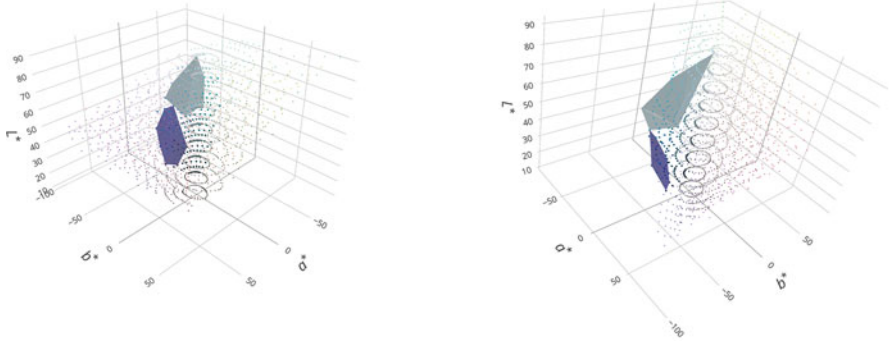
library, we computed the volumes of the various hulls, which are given in Table 1. Here, one notices some quite remarkable differences already. In English, the convex hull of the bilinguals’ choices of focal *light blue* covers more than three times the space covered by the convex hull of the EN monolinguals’ choices for that focal color. In Italian, much the same is true for the convex hulls for the IT monolinguals’ and bilinguals’ choices of focal *blu*. And while for *blue* the convex hull of the bilinguals’ choices is also larger than that of the EN monolinguals’ choices, for *azzurro* the opposite is the case: the hull of the IT monolinguals’ choices is about twice as large as that of the bilinguals’ choices.

We went on to calculate measures of central tendency for the various color-group combinations. Centroids (i.e., centers of mass) of the various convex hulls were calculated using the `QHull.jl` library once more. We also calculated centroids based on all focal color choices, both straight (i.e., unweighted)—by averaging the three coordinates ( $L^*$ ,  $a^*$ ,  $b^*$ ) of all chips that had been designated by at least one participant as being focal for the given color—and weighted, by also taking into account how many participants had designated a chip as being focal for the given



**Fig. 3** Focal *celeste* (top row), *azzurro* (middle row), and *blu* (bottom row) choices of IT monolinguals (left column) and bilinguals (right column)

color. In the same way, we calculated straight and weighted medoids. (Medoids stand to centroids as medians stand to means.) The different calculations did not yield dramatically different results. In Fig. 5, we are showing the weighted centroids

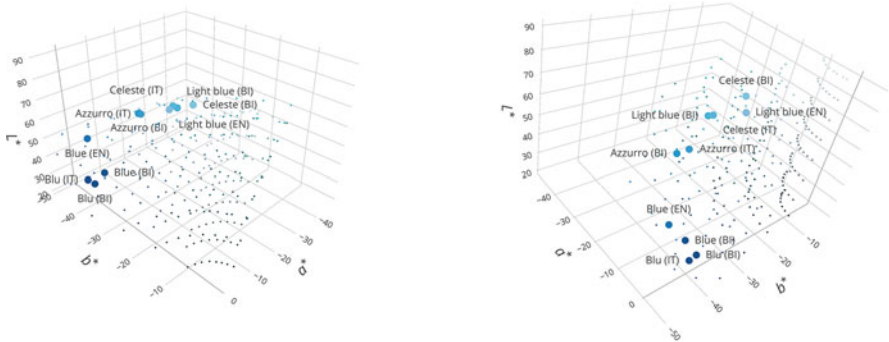


**Fig. 4** Convex hulls of the bilinguals’ focal *blue* and *light blue* choices, seen from different viewpoints

**Table 1** Volumes (in cubic units) of the convex hulls of focal color choices for the various color-group combinations

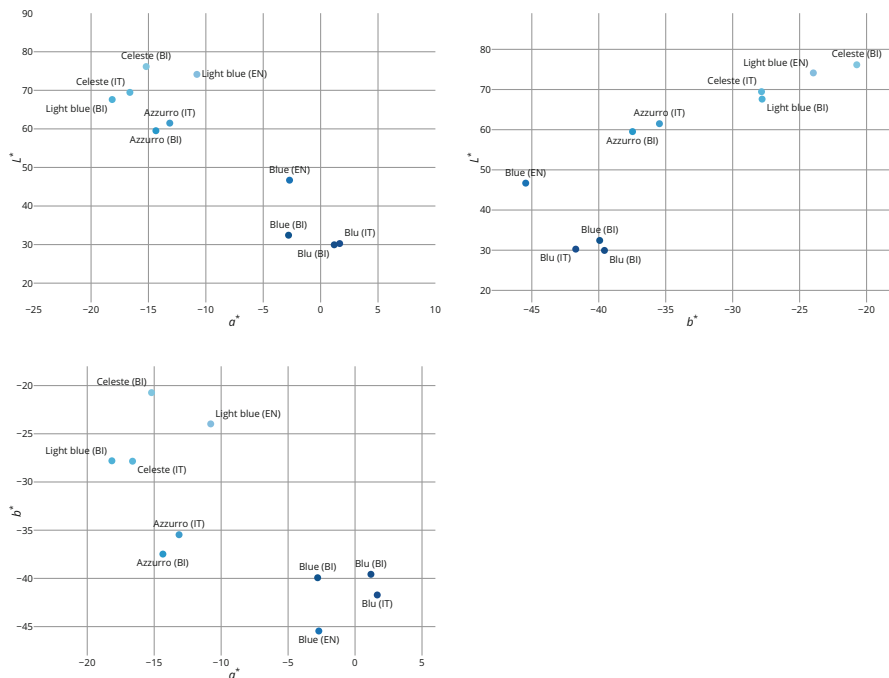
	EN	BI	IT
Blue	2877.9	4006.2	
Light blue	3097.0	11570.8	
Azzurro		4089.6	8100.0
Blu		6765.3	2270.7
Celeste		3326.4	4275.1

*Note:* *EN* English monolinguals, *BI* bilinguals, *IT* Italian monolinguals. For reference, the volume of the convex hull encompassing all Munsell chips is 887469.5



**Fig. 5** Weighted centroids for all color-group combinations. (To facilitate orientation, we also show the locations of the BLUE region stimuli from the study, but not of all Munsell chips, as that would clutter the graphs too much in this case)

in the full, three-dimensional CIELAB space. Because three-dimensional graphics can be hard to interpret, Fig. 6 also gives the three possible two-dimensional views of the space with centroids placed in it.



**Fig. 6** Weighted centroids for all color-group combinations. (See the text for explanation)

Just eyeballing the results, we see some clear patterns. The bilinguals' *blue* centroid appears closer to both the IT monolinguals' and the bilinguals' *blu* centroid than to the EN monolinguals' *blue* centroid. Similarly, the bilinguals' *light blue* centroid appears much closer to the IT monolinguals' *celeste* centroid than to the EN monolinguals' *light blue* centroid. These impressions are confirmed by looking at the distance matrix, given in Table 2. It is seen, for instance, that the bilinguals' focal *blue* centroid is about three times farther removed from the EN monolinguals' focal *blue* centroid than from the IT monolinguals' focal *blu* centroid. And the bilinguals' focal *light blue* centroid is more than four times closer to the IT monolinguals' focal *celeste* centroid than to the EN monolinguals' focal *light blue* centroid.

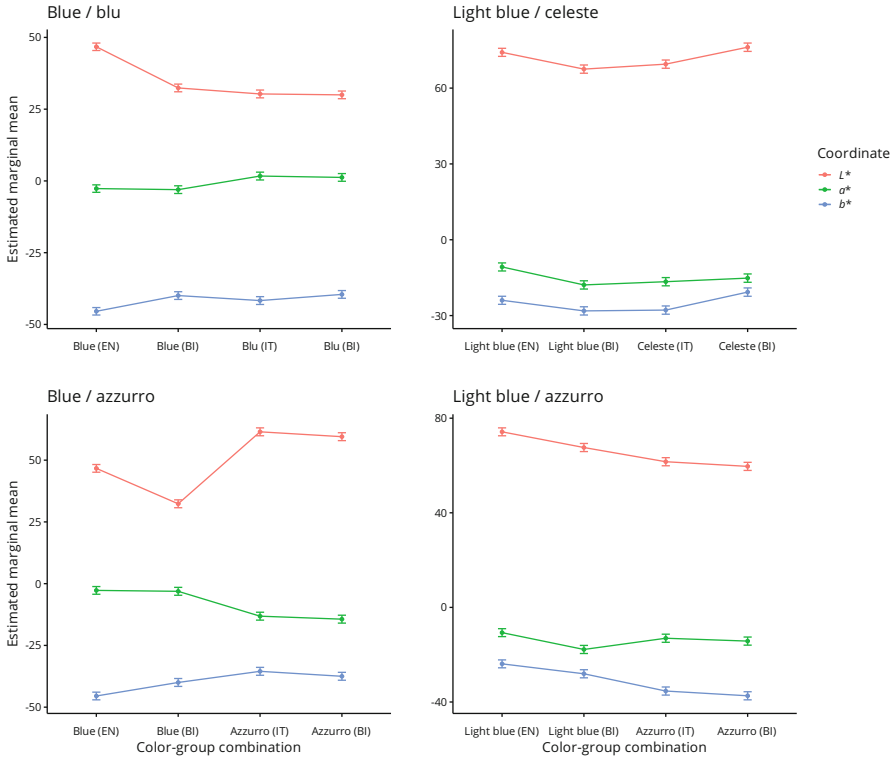
To find out whether the observable differences in the choices of the foci from the three groups have any statistical significance, we conducted four repeated-measures ANOVAs with color-group combination as a between-participants variable and the coordinate ( $L^*$ ,  $a^*$ ,  $b^*$ ) as a within-participants variable. One ANOVA focused on the *blue/blu* comparison, comparing the combinations *blue* (EN), *blue* (BI), *blu* (IT), and *blu* (BI); the second ANOVA focused on a *light blue/celeste* comparison, comparing *light blue* (EN), *light blue* (BI), *celeste* (IT), and *celeste* (BI); the third ANOVA did the same for *blue* and *azzurro*; and the fourth did the same for *light blue* and *azzurro*.

**Table 2** Distance matrix of weighted centroids of focal colors for the various color-group combinations

	EN			IT			BI				
	Light blue	Blue	Blu	Light blue	Blue	Blu	Light blue	Blue	Blu		
EN	Light blue	0.0	35.8	8.4	17.2	48.9	10.6	45.4	48.4	20.2	5.8
	Blue	35.8	0.0	32.0	20.7	17.4	31.0	15.4	18.2	19.1	40.4
IT	Celeste	8.4	32.0	0.0	11.6	45.4	2.4	41.3	44.9	14.0	9.9
	Azzurro	17.2	20.7	11.6	0.0	35.1	10.6	31.2	34.9	3.1	20.9
	Blu	48.9	17.4	45.4	35.1	0.0	44.2	5.5	2.2	33.6	53.2
BI	Light blue	10.6	31.0	2.4	10.6	44.2	0.0	39.9	43.6	12.8	11.7
	Blue	45.4	15.4	41.3	31.2	5.5	39.9	0.0	4.9	29.5	49.4
	Celeste	48.4	18.2	44.9	34.9	2.2	43.6	4.9	0.0	33.5	52.5
	Azzurro	20.2	19.1	14.0	3.1	33.6	12.8	29.5	33.5	0.0	23.6
	Blu	5.8	40.4	9.9	20.9	53.2	11.7	49.4	52.5	23.6	0.0

*Note:* The numbers represent Euclidean distances ( $\Delta E$ ) in CIELAB space





**Fig. 7** Marginal mean coordinates for the various color-group combinations as estimated in the repeated-measures ANOVAs described in the text

The first ANOVA revealed a main effect of the color-group combination,  $F(3, 117) = 15.61$ ,  $p < 0.0001$ , a main effect of the coordinate,  $F(2, 234) = 2615.84$ ,  $p < 0.0001$ , as well as an interaction between the coordinate and the color-group combination,  $F(6, 234) = 14.88$ ,  $p < 0.0001$ . As for the main effect of the color-group combination (the variable most directly of interest to our research), pairwise comparisons showed that the *blue* (EN) foci differed significantly from all other combinations, all  $ps < 0.0001$ , but that the other combinations did not differ significantly among each other, all  $ps > 0.46$ . See the left panel in the top row of Fig. 7 for the estimated marginal mean coordinates for the various color-group combinations.

The second ANOVA also revealed a main effect of the color-group combination,  $F(3, 117) = 6.57$ ,  $p < 0.0005$ , and a main effect of the coordinate,  $F(2, 234) = 7132.77$ ,  $p < 0.0001$ , but the interaction between the coordinate and the color-group combination was only borderline significant,  $F(6, 234) = 2.12$ ,  $p = 0.05$ . Pairwise comparisons showed that the *light blue* (EN) foci differed significantly from the *light blue* (BI) foci,  $p = 0.005$ , and from the *celeste* (IT) foci,  $p = 0.037$ ; furthermore, the *light blue* (BI) foci differed significantly from the *celeste* (IT) foci,

$p = 0.004$ ; finally, the *celeste* (IT) foci differed significantly from the *celeste* (BI) foci,  $p = 0.03$ . The right panel in the top row of Fig. 7 shows the marginal mean coordinates for the various color-group combinations as estimated in this ANOVA.

The third ANOVA found, again, a main effect of both the color-group combination,  $F(3, 118) = 12.86$ ,  $p < 0.0001$ , and the coordinate,  $F(2, 236) = 3443.40$ ,  $p < 0.0001$ , as well as a significant interaction between these variables,  $F(6, 236) = 42.08$ ,  $p < 0.0001$ . Pairwise comparisons revealed significant differences between *blue* (EN) and *azzurro* (IT),  $p = 0.003$ , between *blue* (BI) and *azzurro* (IT),  $p < 0.0001$ , and between *blue* (BI) and *azzurro* (BI),  $p = 0.0001$ . See the left bottom panel in Fig. 7 for the corresponding estimated marginal mean coordinates.

The fourth ANOVA, finally, showed a very similar pattern, with a main effect of the color-group combination,  $F(3, 118) = 13.78$ ,  $p < 0.0001$ , as well as of the coordinate,  $F(2, 236) = 5448.99$ ,  $p < 0.0001$ ; there was also a significant interaction between these variables,  $F(6, 236) = 6.97$ ,  $p < 0.0001$ . Pairwise comparisons showed there to be significant differences between focal colors for *light blue* (EN) and each of the other color-group combinations, all  $ps < 0.005$ . The right bottom panel in Fig. 7 plots the corresponding estimated marginal mean coordinates.

### 3.2.1 Discussion

We saw that bilinguals' choices of focal colors in English were more diffuse (i.e., spread out in color space) than the EN monolinguals' choices, in accord with instability and shifts of BCT prototypes in bilinguals reported previously by Ervin (1961), Caskey-Sirmons and Hickerson (1977), and Athanasopoulos (2009). By contrast, bilinguals' focal color choices in Italian were more concentrated than the IT monolinguals' choices for two out of the three Italian "blue" terms. That for *blu* the bilinguals' choices were actually much more diffuse than those of the IT monolinguals is plausibly due to the phonological and orthographic similarity between the English word *blue* and the Italian word *blu* (see Kroll et al. 2010). This may have led some bilinguals to identify colors as being typically *blu* because, speaking English, they would identify them as typically blue, and thereby to expand the convex hull for focal *blu*, which also encompassed colors more likely to be identified as being typically *blu* by IT monolingual speakers (see also Paramei et al. 2016). The aforementioned similarity between homophone *blue* and *blu* may also explain the clear shift of the *blue* prototype as identified for the bilingual speakers, which was found to be significantly darker and closer to their and IT monolinguals' prototype for *blu* than to the *blue* prototype as determined for the EN monolingual speakers. We also saw evidence for a shift in the bilinguals' *celeste* prototype toward their *light blue* prototype and away from the IT monolingual speakers' *celeste* prototype. All in all, these findings point to cultural and linguistic effects in the mental representation of cognates and semantic equivalents of "blue" in Italian-English bilingual speakers.

## 4 General Discussion

As mentioned in the Introduction, there is clear evidence in favor of the view that principles of rational design (e.g., principles having to do with the informativeness of naming systems) underlie the conceptual structure of color space, more specifically, how we carve up that space categorically and linguistically, and where, in it, we place the color category foci. But the same evidence also shows that the design principles that have been put forward as such in the literature may not tell the whole story about how perceptual color space gets its linguistically defined structure. The whole story may, of course, encompass hitherto unidentified design principles. But the study presented in the foregoing gives reason to believe that, whatever the true collection of design principles may be, cultural and linguistic factors also play some part in the categorical structuring of color space. Some may want to see these further factors as noise, detracting from an ideal structure fully fixed by principles of rational design. Even then, it is important to be at least aware of them.

At a more methodological level, our study showed how work with bilingual speakers can help to identify cultural and linguistic influences on color conceptualization. In many ways, the present study just scratched the surface. One important limitation of the study was that our bilingual participants were all native speakers of Italian but not of English, even if they were highly to very highly proficient in English, and at the time this study was conducted tended to speak English most of the time. Ideally, the study would be complemented by a “symmetric” group of native English speakers highly proficient in Italian and residing in Italy for a number of years. Our results indicate that bilinguals’ focal colors for English *blue* and *light blue* were closer to the Italian monolinguals’ focal colors for *blu* and *celeste*, respectively, than to the English monolinguals’ focal colors for *blue* and *light blue*. It appears reasonable to predict that the opposite would be the case for bilinguals of the kind just described, that is, native English speakers sufficiently immersed in Italian. We mention this here as an avenue for future research.

**Acknowledgments** The study was conducted in the framework of Erasmus Exchange Agreements between Liverpool Hope University with University of Sassari (2009–2011) and University of Florence (2019–2021). We thank Carmen De Caro, Maria Michela Del Viva, and Ilaria Mariani for providing data for Italian monolinguals (Florence), as well as students for assisting in data collection: Nadia Al-Mahrouky and Ricky Morton (English monolingual speakers); Cristina Stara and Ricky Morton (Italian–English bilinguals), both groups in North-West England (Liverpool and Manchester). We are greatly indebted to Panos Athanasopoulos for sharing the Nation Vocabulary Test; John Mollon for the advice on the standardized illumination; Deborah Roberson for guidance on the unconstrained color-naming method; Guido Frison for consultation on early use of Italian terms for “blue”; and Manila Soffici for consultation on the Florentine dialect and Standard Italian. We thank Robert Hewertson for technical assistance, Kaida Xiao for illuminance measurement, and all participants for their time and good will. We are also greatly indebted to two anonymous referees for valuable comments on a previous version of this chapter.

## References

- Albertazzi, L. & Da Pos, O. (2017). Color names, stimulus color, and their subjective links. *Color Research and Application*, 42, 89–101.
- Athanasopoulos, P. (2009). Cognitive representation of colour in bilinguals: The case of the Greek blues. *Bilingualism: Language and Cognition*, 12, 83–95.
- Berlin, B. & Kay, P. (1969/1991). *Basic color terms*. Stanford CA: CSLI Publications.
- Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Review*, 59, 65–98.
- Bimler, D. & Uusküla, M. (2014). “Clothed in triple blues”: Sorting out the Italian blues. *Journal of the Optical Society of America A*, 31, A332–A340.
- Bimler, D. & Uusküla, M. (2018). Individual variations in color-concept space replicate across languages. *Journal of the Optical Society of America A*, 35, B184–B191.
- Caskey-Sirmons, L. A. & Hickerson, N. P. (1977). Semantic shift and bilingualism: Variation in the color terms of five languages. *Anthropological Linguistics*, 19, 358–367.
- Del Viva, M. M., Mariani, I., De Caro, C., & Paramei, G. V. (2022). Florence “blues” are clothed in triple basic terms. *i-Perception*, 13, 20416695221124964. <https://doi.org/10.1177/20416695221124964>
- De Mauro, T. (1983). *Storia linguistica dell'Italia unita* [Linguistic history of the unified Italy]. Rome/Bari: Laterza & Figli.
- Douven, I. (2017). Clustering colors. *Cognitive Systems Research*, 45, 70–81.
- Douven, I. (2019). Putting prototypes in place. *Cognition*, 193, 104007. <https://doi.org/10.1016/j.cognition.2019.104007>
- Douven, I. & Gärdenfors, P. (2020). What are natural concepts? A design perspective. *Mind & Language*, 35, 313–334.
- Ervin, S. M. (1961). Semantic shift in bilingualism. *American Journal of Psychology*, 74, 233–241.
- Fairchild, M. D. (2013). *Color appearance models*. Hoboken NJ: Wiley.
- Frison, G. & Brun, G. (2016). Lapis lazuli, lazurite, ultramarine “blue,” and the colour term “azure” up to the 13th century. *Journal of the International Colour Association*, 16, 41–55.
- Gärdenfors, P. (2000). *Conceptual spaces*. Cambridge MA: MIT Press.
- Giacalone Ramat, A. (1978). Strutturazione della terminologia dei colori nei dialetti sardi. Italia linguistica nuova ed antica [A structure of color terminology in Sardinian dialects. Modern and antique linguistic Italy]. In V. Pisani & C. Santoro (Eds.), *Studi linguistici in memoria di Oronzo Parlangeli II* (pp. 163–181). Lecce/Milan: Congedo editore.
- Grossmann, M. (1988). *Colori e lessico: Studi sulla struttura semantica degli aggettivi di colore in catalano, castigliano, italiano, romeno, latino ed ungherese* [Colors and lexicon: Studies on semantic structure of color adjectives in Catalan, Castilian, Italian, Romanian, Latin and Hungarian]. *Tübinger Beiträge zur Linguistik* (Vol. 310). Tübingen: Gunter Narr Verlag.
- Ishihara, S. (1973). *Test for colour-blindness* (24 plates edition). Tokyo: Kanehara Shuppan.
- Jameson, K. A. (2005). Why GRUE? An Interpoint-Distance Model analysis of composite color categories. *Cross-Cultural Research*, 39, 159–204.
- Jameson, K. A. & D’Andrade, R. (1997). It’s not really red, green, yellow, blue: An inquiry into perceptual color space. In C. L. Hardin & L. Maffi (Eds.), *Color categories in thought and language* (pp. 295–319). Cambridge UK: Cambridge University Press.
- Jraissati, Y. & Douven, I. (2017). Does optimal partitioning of color space account for universal color categorization? *PLoS ONE*, 12, e0178083. <https://doi.org/10.1371/journal.pone.0178083>
- Jraissati, Y. & Douven, I. (2018). Delving deeper into color space. *i-Perception*, 9, 204166951879206. <https://doi.org/10.1177/2041669518792062>
- Kemp, C. & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, 336, 1049–1054.
- Kroll, J. F., van Hell, J. G., Tokowicz, N., & Green, D. W. (2010). The Revised Hierarchical Model: A critical review and assessment. *Bilingualism: Language and Cognition*, 13, 373–381.

- Munsell, A. H. (1941). *A color notation: An illustrated system defining all colors and their relations*. Boston: The Hoffman Brothers Co.
- Nation, I. S. P. (1980). *Teaching and learning vocabulary*. New York: Newbury House.
- Paggetti, G., Menegaz, G., & Paramei, G. V. (2016). Color naming in Italian language. *Color Research and Application*, 41, 402–415.
- Paramei, G. V. (2005). Singing the Russian blues: An argument for culturally basic color terms. *Cross-Cultural Research*, 39, 10–38.
- Paramei, G. V. (2007). Russian “blues”: Controversies of basicness. In R. E. MacLaury, G. V. Paramei, & D. Dedrick (Eds.), *Anthropology of color: Interdisciplinary multilevel modeling* (pp. 75–106). Amsterdam/Philadelphia: John Benjamins.
- Paramei, G. V. & Bimler D. L. (2021). Language and psychology. In A. Steinvall & S. Street (Eds.), *A cultural history of color; vol. 6, The Modern Age: From 1920 to present* (Ch. 6, pp. 117–134). London: Bloomsbury.
- Paramei, G. V., D’Orsi, M., & Menegaz, G. (2014). “Italian blues”: A challenge to the universal inventory of basic colour terms. *Journal of the International Colour Association*, 13, 27–35.
- Paramei, G. V., D’Orsi, M., & Menegaz, G. (2016). Cross-linguistic similarity affects L2 cognate representation: *blu* vs. *blue* in Italian–English bilinguals. *Journal of the International Colour Association*, 16, 69–81.
- Paramei, G. V., D’Orsi, M., & Menegaz, G. (2018). Diatopic variation in referential meaning of “Italian blues.” In L. W. MacDonald, C. P. Biggam, & G. V. Paramei (Eds.), *Progress in colour studies: Cognition, language and beyond* (pp. 83–105). Amsterdam/Philadelphia: John Benjamins.
- Pastoureau, M. (2001). *Blue: The history of a color*. Princeton NJ: Princeton University Press.
- Pirenne, M. H. (1962). Dark-adaptation and night vision. In H. Davson (Ed.), *The visual process* (pp. 93–122). Cambridge MA: Academic Press.
- Regier, T., Kay, P., & Khetarpal, N. (2007.) Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences of the U.S.A.*, 104, 1436–1441.
- Rosch, E. (1973). Natural categories. *Cognitive Psychology*, 4, 328–350.
- Sandford, J. L. (2015). A cognitive linguistic usage perspective: What is Italian *blu*, *azzurro*, *celeste*? Do English speakers agree on BLUE semantics? *Cultura e Scienza del Colore—Color, Culture and Science*, 4, 22–30.
- Uusküla, M. (2014). Linguistic categorization of blue in Standard Italian. In C. J. Kay, C. A. Hough, & C. P. Biggam (Eds.), *Colour studies: A broad spectrum* (pp. 67–78). Amsterdam/Philadelphia: John Benjamins.
- Wattendorf, E. & Festman, J. (2008). Images of the multilingual brain: The effect of age of second language acquisition. *Annual Review of Applied Linguistics*, 28, 3–24.
- Xu, Y. & Regier, T. (2014). Numeral systems across languages support efficient communication: From approximate numerosity to recursion. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Meeting of the Cognitive Science Society* (pp. 1802–1807). Austin TX: Cognitive Science Society.
- Xu, Y., Regier, T., & Malt, B. C. (2016). Historical semantic chaining and efficient communication: The case of container names. *Cognitive Science*, 40, 2081–2094.

# The Dimensionality of Color Perception



Javier Fdez, Oneris Rico, and Olaf Witkowski

**Abstract** Chromatics, or the science of color, not only studies the description of colors in terms of the physics of electromagnetic radiations, but also their perception through the human eye and cognitive apparatus. Although in purely physical terms colors may be described by as few as three dimensions – such as hue, saturation, and brightness – an open debate remains about how our cognition maps colors and in how many dimensions they encode the distinction between colors according to our perspective. In this chapter, we study the trade-off between finding an embedding for color perception with the minimal number of dimensions, while maximizing the discrimination between colors. To do so, we designed an experiment where thirteen subjects reported the similarity between twenty colors randomly generated using the Munsell color system. For each subject, we mapped perceived colors in an  $n$ -dimensional space, where distances between two colors reflect how different they are according to the subject. We used a least squares optimization to minimize the difference between subject-reported and mapped distances between colors with that dimensionality. We then repeated the process for values from one to nine dimensions. Our results showed an optimal number of dimensions of three when using a cosine similarity measure, which indicates a resemblance to the way the perception of colors is cognitively encoded from mere physical properties of color maps. We discuss the implications and limitations of these results in the light of color theory, and their relevance in both our understanding of the topology of mental concepts and major applications in fields where color theory is important, including composing color scales for designer tools, color psychology in marketing, color matching in interior architecture, and chromatic treatments in post-production of film-making.

**Keywords** Color perception · Color theory · Dimensions of color · Visual perception · Human perception

---

J. Fdez (✉) · O. Rico · O. Witkowski  
Cross Labs, Cross Compass Ltd., Kyoto, Japan

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023  
T. Veloz et al. (eds.), *Trends and Challenges in Cognitive Modeling*, STEAM-H:  
Science, Technology, Engineering, Agriculture, Mathematics & Health,  
[https://doi.org/10.1007/978-3-031-41862-4\\_12](https://doi.org/10.1007/978-3-031-41862-4_12)

165

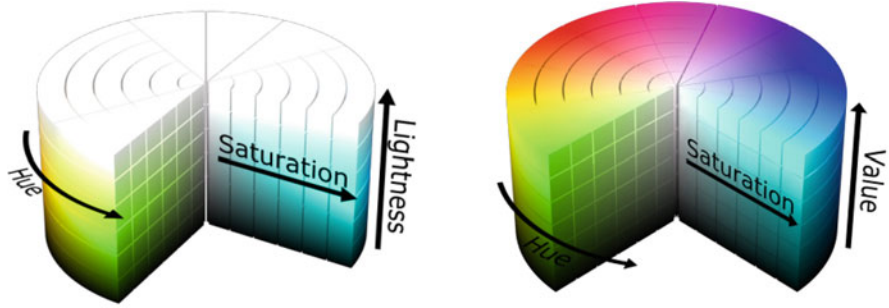
## 1 Introduction

The study of color is a fascinating field as it is connected with major applications such as composing color scales for designer tools (OU et al. 2008), color psychology in marketing (KHATTAK et al. 2018), color matching in interior architecture (Ulusoy et al. 2020), and chromatic treatments in post-production of film-making (Misek 2010). The science of color is called *chromatics*, and it not only studies how to describe colors based on the physics of electromagnetic radiation, but it also covers their perception by the human eye and cognitive apparatus.

One of the fundamental problems in cognitive science is how human cognition categorizes the visible color spectrum. Back in 1969, Berlin and Kay (1969) proposed that the basic color terms for each culture are predictable by the number of color terms the culture has. This work was later extended in the late 1970s (Cook et al. 2005), analyzing the color-naming data from speakers of 110 unwritten languages and assuring the existence of a partially fixed evolutionary progression. Moreover, the Sapir–Whorf hypothesis (Malkiel 1956), also known as linguistic relativity, suggests that the language we speak influences the way we think and perceive the world around us. In other words, the structure and vocabulary of a language shape the cognitive processes of its speakers, predicting that language-specific basic color categories are more likely to have a behavioral advantage at their additional cross-category boundary compared to other languages that do not differentiate categorically that area of color space (Bignozzi 2021; Martinovic et al. 2020). Also, Baronchelli et al. (2010) and Loreto et al. (2012) used multiagent simulations to categorize and name colors through a purely cultural negotiation in the form of language games, featuring an excellent quantitative agreement with the empirical observations of the World Color Survey (WCS).

One of the primary methods of classifying colors is based on their image processing applications (Ibraheem et al. 2012). One of the categories is user-oriented, which focuses on how the user perceives colors (Plataniotis and Venetsanopoulos 2000). Within the user-oriented category, there are two main color models: HSL (hue, saturation, lightness) (Joblove and Greenberg 1978) and HSV (hue, saturation, value, also known as HSB, where the *B* refers to brightness) (Smith 1978). While HSL representation models the way different paints are mixed to create color in the real world, the HSB representation models how colors appear under light. Both color representations are displayed in Fig. 1.

Both representations display hue and saturation as polar coordinates of a color circle (Hanbury 2008; Ibraheem et al. 2012). The former refers to the dominant wavelength of the spectrum, defining the classification of the color as red, green, blue, or an intermediate color. Instead, saturation relates to the chromatic intensity, and it is determined by how the light intensity is distributed across the spectrum of different wavelengths. The most saturated color is achieved with just one wavelength at a high intensity (e.g., laser light). Nevertheless, the saturation component varies for both models (Joblove and Greenberg 1978). For the HSL color space, saturation exists independently of lightness where both a very light color and a very



**Fig. 1** On the left, HSL (Hue, Saturation, Lightness) representation, from Wikimedia Commons (2010a). On the right, HSV (Hue, Saturation, Value) representation, from Wikimedia Commons (2010b)

dark color can be heavily saturated. Instead, all colors approaching white feature low saturation for HSV color space. Hence, saturation goes from fully saturated color to the equivalent gray for the HSL model, and from saturated color to white for the HSV model. Regarding the last dimension for both models, while the lightness dimension relates to the varying amounts of black or white paint in the mixture, the value is analogous to shining a white light on a colored object. The lightness in HSL always spans the entire range from black to white, while the value only goes half that way, from black to the chosen hue.

Research has shown how both HSV and HSL models do not effectively separate colors in the way human perception of color does. For example, MacLeod (2003) suggests that the standard three-dimensional conception of perceived color is inadequate. Specifically, the author indicates how the signal arriving to the retina is mainly determined, not by the local light stimulus, but also by its immediate surrounding. Likewise, color representations in two dimensions based on physics, such as CIE, have been proven inadequate in depicting colors as we experience them. A research conducted by Wright (1941) investigated the size of small color steps for different lines on the CIE color chart. The findings indicated significant variations in the perceived step size among observers situated in various regions of the CIE chart.

Moreover, there are visual effects that reflect that color perception changes under different conditions. One example that reflects how perception cannot only be described with these approaches is the light/dark adaptive state, also known as Purkinje shift, which is the tendency for the peak luminance sensitivity of the eye to shift toward the blue end of the color spectrum at low illumination levels as part of dark adaptation. A work that reflects this effect is the one by Anstis (2002), where they filled an iso-eccentric annulus with radial red/blue sectors and arranged that if the blue sectors looked darker (lighter) than the red sectors, the annulus would appear to rotate to the left (right).

Another example is the adaptation after-effects, where prolonged exposure to one object influences the perception of the next. Reindl et al. (2018) conducted



three experiments in which images of crabs and lobsters were presented in two versions: as complex, naturalistic images, or reduced to their simplified geometric shapes. They found out that the magnitude of adaptation after-effects depends on the complexity of the adaptor, but not on that of the test stimuli.

A third example is the simultaneous contrast effects, where simultaneous contrast affects our perception of color. Kaneko et al. (2017) show it by demonstrating that simultaneous contrast for brightness and color (chromatic saturation) were enhanced by flashing the stimulus very briefly (for 10 ms).

One last example is color constancy, which happens when the perceived color of objects does not vary much with changes in the illumination, despite these changes causing big shifts in the spectral light entering the eye. Xiao et al. (2017) studied conditions both where all cues were consistent with the simulated illuminant change (consistent-cue conditions) and where local contrast was silenced as a cue (reduced-cue conditions), suggesting a reliable interaction between test object type and cue condition.

In summary, even though there are several models in the literature that seem to be related to human perception of colors, both the adaptation after-effects and other studies found in the literature open a debate about how our cognition maps color and in how many dimensions they encode the distinction between colors according to our perspective. In this chapter, we investigate the dimensions of color perception by asking thirteen participants to label the similarities between colors to then study the trade-off between minimizing the number of dimensions for color perception while maximizing the discrimination between them.

To encourage further research on this topic, we have made the source code and the interface freely accessible to all.<sup>1</sup>

## 2 Materials and Methods

Our methodology unfolds as follows. First, we designed an experiment to capture the human perception of colors (c.f., Sect. 2.1). Second, we implemented the least squares optimization method to estimate the color embedding for various dimensions (c.f., Sect. 2.2). Lastly, we used the cosine similarity method to assess the optimal number of dimensions that describes the color perception (c.f., 2.3).

---

<sup>1</sup> <https://github.com/javiferfer/color-dimensionality>.

## **2.1 Experimental Design**

### **2.1.1 Materials and Setup**

The interface was developed using the Unity Real-Time Development Platform (Haas 2014) and was specifically designed for a MacBook Pro 13. All experiments were conducted in person using the same MacBook Pro 13 whose physical screen resolution is  $2560 \times 1600$  (square pixels) with a diagonal length of 13.3 inches. To ensure the accuracy of colors, all operating system color modifications were disabled, and the user interface was set to an 18% gray background.

The experiments were carried out in a controlled environment with only artificial room lighting. Specifically, the experiments were conducted in the evening without natural light, and under a neutral LED light with a color temperature of approximately 4000K and a CRI of 89. Participants completed the experiment on a computer screen positioned to minimize reflections from the light. The screen had a brightness rating of 1000 nits and was calibrated to produce a constant 457 LUX across its entire area. Participants sat on a wooden Scandinavian dining chair, and the laptop was placed on a table with a height of 70 cm. The distance between the participants and the computer screen was 50–60 cm, and the elbow angle when writing on the computer was around  $85^{\circ}$ – $90^{\circ}$ . The participants seated at an average distance of 55 cm from the screen, resulting in a test area within the parafoveal visual field of 7.3 degrees.

### **2.1.2 Participants**

Thirteen participants took part in the experiment, which is the required sample size to achieve a 95% confidence level that the measured value is within  $\pm 5.5\%$  of the true value with a 1% population proportion assumption and an unlimited population size.

All participants had normal vision based on self-reports. The test lasted an average of 22 minutes per participant, with 5 minutes and 10 seconds (ranging from 3 minutes and 50 seconds to 6 minutes and 10 seconds) allocated for instructions, and 17 minutes and 50 seconds (ranging from 15 minutes and 30 seconds to 21 minutes) for the experiment itself.

### **2.1.3 Experiment Protocol**

To assess the dimensionality of the colors, we implemented an experiment where participants had to manually label, using a slider, the similarity between the two colors displayed in the trial.

The experiment consisted on manually labeling, using a slider, the similarity between the two colors displayed in the trial. Placing the slider at one extreme indicated that the colors were dissimilar, meaning that they had nothing in common.

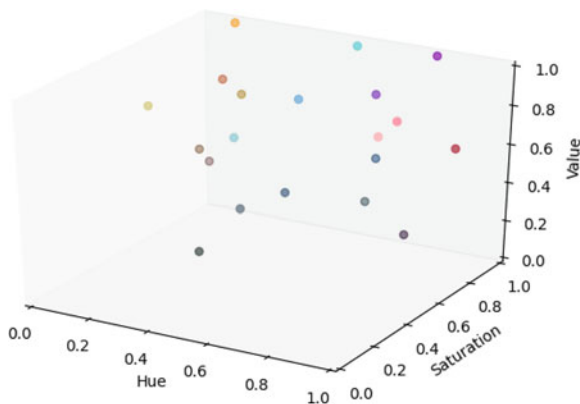
In contrast, by placing it in the opposite extreme meant that they were completely similar. What was fully similar and dissimilar was totally decided by the participants.

The experiment began with the instructions, which first instruct participants to sit comfortably in the chair and find an environment with no distractions. Then, the instructions describe the structure of the experiment, indicating that the experiments consist of four practice trials and 190 recorded trials. The instructions also mention that the trials are grouped in sets of four and that the next group of trials will not be displayed until you have labeled each of the trials. Thereafter, the instructions state the participants that their task is to indicate the similarity between the two colors displayed using a slider. The instructions also mention that moving the slider to the far left indicates that the colors are dissimilar and have no common features. Conversely, positioning it to the far right suggests that the colors are similar. A scale ranging from zero to ten is provided to rate the degree of color similarity. Lastly, the instructions mention that any question should be asked then or, otherwise, no questions could be asked afterward. No questions were asked by any participant.

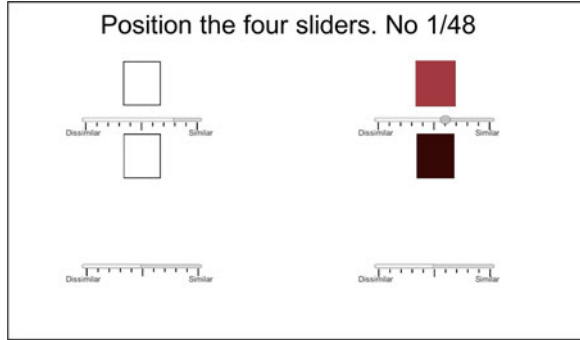
The second part of the experiment consists in a set of four randomly selected practice trials, followed by the 190 recorded trials. The trials were based on the combinations from the twenty colors randomly extracted using the Munsell color system. In colorimetry, the Munsell color system is a color space that specifies colors based on hue, value, and chroma. The randomly extracted colors are the ones displayed in Fig. 2. To minimize any order effects and balance out the impact of potential confounding factors that may affect participants differently across different conditions or treatments, the color combinations and positions were randomly assigned at the beginning of each experiment and for each participant.

Initially, the only objects shown to the participants were the labeling bars. To report a trial, the participant would position the mouse on top of the region of the trial, and only then the colors would appear. Once the trial was labeled and the mouse was placed out of the region of the trial, the colors would disappear to avoid any distraction when reporting other trials, only leaving a white square

**Fig. 2** The twenty randomly selected colors displayed into the HSV (hue, saturation, value) color system



**Fig. 3** Example of the interface where the user reports the similarity between colors. In this scenario, the participant has already labeled the top-left trial and is currently labeling the top-right trial. Hence, the white squares appear for the top-left trial, while the squared colors appear for the top-right trial



on the screen. The trial labeling was done using a scale from zero to ten, with a zero meaning they are completely dissimilar and a ten meaning they are completely similar. Furthermore, the labeling was always a whole number restricting the option of labeling the similarity with a decimal number. Lastly, the labeling bar was set by default to five (mid-value). The button to continue to the next set of trials appeared only when the four trials had been rated. Figure 3 shows a screenshot of the interface where the participant had already labeled the top-left trial and is labeling the top-right one. For this scenario, the participant still needs to answer the last two trials to unlock the button that allows the participant to continue to the next set of trials. The distance between the squares’ centers on the MacBook Pro 13 was 4.5 cm, and the vertical distance from the upper boundary of one square to the bottom boundary of the other square was 7 cm.

## 2.2 Optimization Method

Once the participants reported the similarity for the trials, the next step was to estimate the color embeddings that minimize the error between the computed distances and the reported ones. Since there were twenty colors in total, the maximum number of combinations without repetition was 190, which matches the number of combinations asked of the participants in the experiment.

The least squares algorithm was used as the optimization method. This finds the set of parameters that minimize the error function. The objective of the algorithm was to get an estimate of the color embeddings in an n-dimensional space that minimizes the mean squared error. This was iteratively computed over several dimensions, with constraint to nine dimensions, as this is the maximum number of dimensions before obtaining an underdetermined system of equations. An underdetermined system of equations occurs when the number of unknown variables is greater than the number of equations in the system. In other words, there are not enough constraints to determine a unique solution. This results in either no solution or an infinite number of solutions that can satisfy the given equations. In

our case, an underdetermined system of equations occurs with 10 dimensions, as there would be 200 variables (20 colors x 10 dimensions) and only 190 equations. These dimensions are not related to any specific color models but rather are a way to represent the colors in a higher dimensional space based on the reports of the participants.

Stating the error between the reported and estimated distances in an  $n$ -dimensional space is as follows:

$$\begin{aligned} d(i, j) &= \sqrt{(x_i - x_j)^2} \\ error(i, j) &= |d_{ij} - d(i, j)|^2 \end{aligned} \quad (1)$$

where  $i$  and  $j$  indicate the two colors,  $x$  the position for each color in the embedding space, and  $d_{ij}$  is the reported distance by the participant.

The objective of the algorithm is to calculate the positions of the colors that minimize the total error for each dimension, defined as:

$$\min \left( \sum_{i=1}^{20} \sum_{j=i+1}^{20} error(i, j) \right) \quad (2)$$

The reasons for choosing this algorithm are: (1) the algorithm is sensitive to the outliers, which is important as it gives more weight to the most sensitive inputs from the participants, (2) it is not necessary to set any hyperparameters, so the results are easily reproducible, and (3) the result only depends on the initial positions of the colors.

### 2.3 Dimensionality Selection

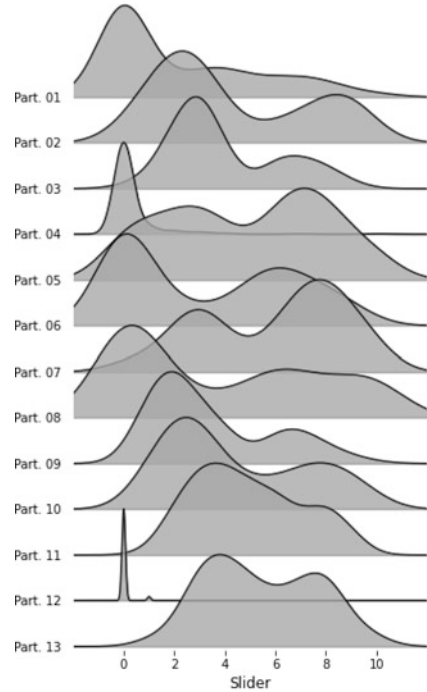
To analyze the optimal dimensional space that describes the color perception, we have used the cosine similarity method, which measures the similarity between two vectors of an inner product space. It is measured by the cosine of the angle between two vectors and determines whether two vectors are pointing in roughly the same direction.

Once the positions for each of the dimensions were optimized, the mean squared error for each dimension  $n$  was computed as

$$MSE_n = \frac{\sum_{i=1}^{20} \sum_{j=i+1}^{20} error(i, j)}{190} \quad (3)$$

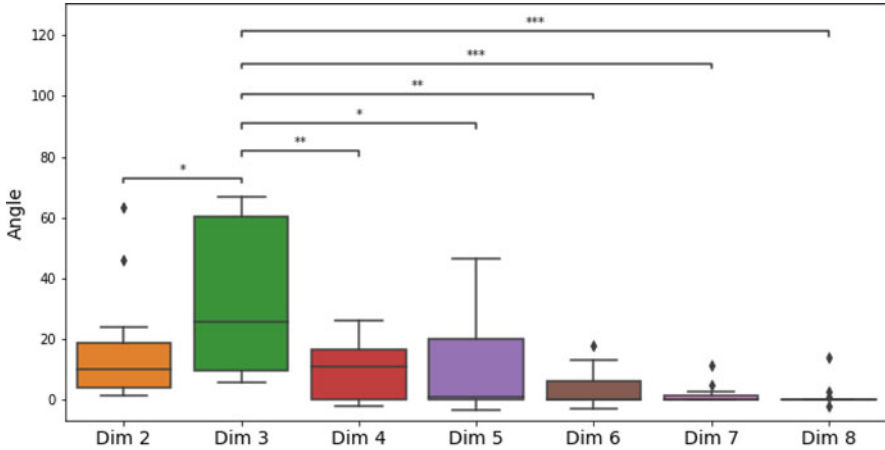


**Fig. 5** Distributions of the reported labels by each participant, indicated as *Part.* The y-axis is scaled for each distribution, instead of having a common scale for all



Nevertheless, two of the participants displayed a distinct pattern in their error line when compared to the other participants, shown in Fig. 5, which displays the distributions of the reported labels of the thirteen participants. Specifically, participants four and twelve, which are the ones with the different patterns, reported significantly different than the rest of the participants. Those two participants labeled most of their trials as completely dissimilar, with a ratio for the rating zero over all the trials of 85.79% and 95.79%, respectively. In other words, these participants indicated that most of the colors had no similarity with each other. As a reference, the next two highest ratios were 55.26% for participant one and 41.58% for participant six for that same label. In addition, we used the Hartigans' dip test (Hartigan and Hartigan 1985) to determine if the samples followed an uni-modal distribution. Results indicated that participants four and twelve were the only two participants with uni-modality in their labeling. Taking into account that the dip test's  $p$ -value increases when the distribution is deviant from a uni-modal distribution, the  $p$ -value for the former participant was of 0.034, while for the latter was 0.021.

From the obtained errors, we then ran the cosine similarity method to measure the angle between two slopes. This was possible for all the dimensions but the first and the last ones, which correspond to dimensions one and nine, as they do not have a prior or posterior slope to compare with. The distributions of the computed angles are shown in Fig. 6.



**Fig. 6** The boxplots indicate the distribution of the angle computed by running the cosine similarity method over the error for each participant. The value inside the boxplots is the median value of the distribution. The figure also reports the *p*-value of the paired *t*-test between the different dimensions for which we ran the least squares method

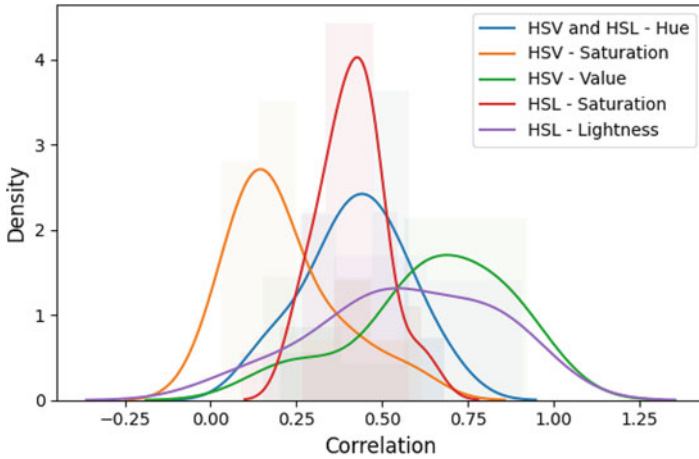
Besides, we ran a paired *t*-test to find an effect on the distribution of the angles for the different number of dimensions. Results have shown that the optimal number of dimensions that encodes optimally the human perception of colors while encoding the maximum information with the lowest complexity is three dimensions. Compared with its closed neighbors, two and four dimensions, the *p*-values were 0.032 for the former and 0.0077 for the latter. Also, participants with different labeling criteria influenced the angle for five dimensions, whose mean was higher than the four dimensions one. But, the distribution for this number of dimensions was also found statistically lower than the one for three dimensions with a *p*-value of 0.042.

These results resemble the way the perception of colors is cognitively encoded from mere physical properties of color maps, as both encode the perception in three dimensions. Then, we compared the relationship between the embedding positions estimated for each participant with the dimensions of the two most important color representations: HSV (hue, saturation, and value) and HSL (hue, saturation, and lightness).

For that comparison, we first computed the HSV and HSL for the twenty colors used in the experiment. Then, we calculated the Pearson correlation coefficient (Freedman et al. 2007) between each embedding dimension for each participant with each of the dimensions of the color models. The results are displayed in Fig. 7, where each of the distributions corresponds to the maximum correlation in absolute value for the thirteen participants.

The dimension with the highest mean was the value dimension of the HSV model ( $M = 0.64$ ,  $SD = 0.21$ ), followed by the lightness dimension of the HSL model ( $M = 0.57$ ,  $SD = 0.24$ ). The former displayed a statistically significant





**Fig. 7** Distribution plot for the correlations in absolute value for the dimensions of the HSV and HSL models. The distribution plot displays a histogram superimposed with a density curve, whose area under the curve must be one. Both models share the hue; hence, it is combined when visualizing it in the figure

higher correlation compared to the other dimensions ( $p$ -value  $< 0.05$ ). Similarly, the latter was found to be statistically significant higher than the remaining dimensions, except for the value and saturation of the HSL model, with a  $p$ -value of 0.11. Both outcomes were observed after conducting a paired  $t$ -test. Whereas the lightness dimension refers to the varying amounts of black or white paint in the mixture as it is related to paintings, the value, also named brightness, is analogous to shining a white light on a colored object. Therefore, both dimensions are closely linked with the effect of light on colors.

## 4 Discussion

Color perception is a fascinating and complex topic that has been studied extensively in various fields such as psychology, computer graphics, and printing. One of the key challenges in color representation is choosing an appropriate color model that accurately captures the characteristics of human perception. In this regard, a three-dimensional model has emerged as a popular choice due to its ability to represent color in terms of hue, saturation, and brightness or value. Interestingly, the optimal model was found to be the three-dimensional model. Furthermore, our correlation analysis between the obtained three-dimensional model and the HSL and HSV models has shown that the dimensions linked to the quantity of light in colors exhibit the strongest correlation. This means that there is a main part of the human

perception related to the light, more than, for example, the hue or saturation of colors.

The finding that the amount of light in colors has the strongest correlation with human perception is consistent with previous research in the field of color science. For example, a study by Fairchild (2013) found that lightness, which is related to the amount of light in a color, is one of the most important perceptual dimensions of color. Similarly, a study by Brainard and Wandell (2019) found that variations in lightness are more easily perceived by humans than variations in hue or saturation. The study by Valdez and Mehrabian (1994) found that the perceived emotional response to colors is influenced by their lightness, with lighter colors being associated with more positive emotions.

Taken together, these studies suggest that the amount of light in a color plays a crucial role in human perception and emotional response to color. This finding has important implications for various fields that rely on accurate representation of color, such as graphic design and printing. By using a color model that accurately captures the relationship between light and human perception, designers and printers can create more effective and visually appealing designs.

However, some factors are important to discuss here, as they may affect the results of future studies. These factors are discussed below.

First, the analysis was carried out using twenty randomly selected colors from the Munsell color system. This was implemented in this way to prevent introducing a bias in our experiment as we wanted to avoid choosing colors from an already predefined color model in the literature. Also, we selected “just” twenty colors to avoid having a tedious experiment, as with twenty colors there are already 190 trials to be asked of the participant. However, selecting the colors randomly has caused to have gaps in some ranges of the hue dimension. Specifically, we have found three main gaps: 0.15–0.4, 0.59–0.75, and 0.79–0.98. These ranges correspond to chartreuse green and green colors, blue colors, and magenta and rose colors, respectively. Subsequent research could expand the range of colors and investigate whether the correlation findings remain consistent with the present chapter.

Besides, the labeling method used for the experiment has been using a slider ranging from zero to ten to indicate the similarity between the two colors. This method has induced an inherent error within the experiment as the measures do not necessarily match when selecting the measure between two colors and comparing it with a third one. When measuring this inherent error in the most conservative way, the inherent error has been found to represent 22.32% of the error of the least squares method. This error was measured by: (1) selecting iteratively three colors, (2) obtaining the maximal and minimal distance allowed for the third combination using the first two combinations, and (3) computing the error when the color went out of that range. Hence, even though all the labeling methods present their own drawbacks, perhaps selecting another labeling approach would decrease these flaws, such as having a three-color labeling method where one of them is an anchor and the participant has to indicate which of the other two colors is more similar to the anchor color.

Lastly, this chapter has conducted multiple  $t$ -tests without adjusting for the possibility of false positive results (i.e., Type I errors). When conducting multiple statistical tests, the probability of finding significant results by chance alone increases, and so it is standard practice to adjust the statistical significance threshold to control the overall probability of Type I errors. Hence, for future studies, increasing the number of participants would certainly help to reduce the problem of multiple comparisons, as it can increase the statistical power of the study, as well as applying appropriate correction methods to control the Type I error rate.

## 5 Conclusion

The study in this chapter was motivated by color perception being usually modeled as a three-dimensional model, primarily based on physical considerations, while the way human cognition maps colors remains an open-ended question. Our goal was to shed some light on the topological constraints for color perception in the human mind, by studying the trade-off between maximizing the discrimination between colors and minimizing the number of dimensions of the embedding space for perceived colors.

First, we implemented an experiment where thirteen participants were asked to report the similarity in pair of two of twenty randomly generated colors. Then, from the results reported, we computed the color embeddings for dimensionalities from one to nine using a least squares optimization method. From there, we used cosine similarity to determine the change in slope for each dimension and participant. A paired  $t$ -test allows us to determine for which dimensionality the slope changes most significantly, indicating that the dimensions involved before the change manage to capture a lot of the topological constraints at play.

Results indicate that the optimal number of dimensions that can encode the human perception of colors while maintaining the highest amount of information with the lowest complexity is three dimensions, determined by measuring the change in the slope of the error. Interestingly, this number matches the literature on this topic (Joblove and Greenberg 1978; Smith 1978), albeit with a focus on human perception rather than a purely physics-based approach. Since these dimensions matched the ones described in the HSL and HSV models, we assessed the relationship between the embedded dimensions for the model estimated and the ones in those models, concluding that the most highly correlated dimensions were the ones linked with the light ones as they were the lightness of the HSL and the value for the HSV model. Also, we discussed the limitation of the color selection and the labeling method, proposing new approaches for future studies. This chapter indicates that a model describing a cognitive map for color perception fundamentally differs from previous models found in the literature.

In our future research, we intend to further explore the influence of color names, building upon previous studies conducted by Baronchelli et al. (2010) and Loreto et al. (2012). A thorough analysis will be conducted to evaluate similarities and

discrepancies across different languages and cultures, with particular attention paid to the ongoing debate surrounding the Sapir–Whorf hypothesis (Hoijer 1954; Malkiel 1956). This hypothesis proposes that the structure and vocabulary of a language can shape the cognitive processes of its speakers, influencing their perceptions, thoughts, and behaviors. We also intend to investigate further the physical meanings of each of the three dimensions discovered in this chapter, which would complete the picture of a physiological model of perception, with concrete consequences on their utilization for countless applications in media, marketing, or architecture, among others. This methodology could then be extended to cognitive representation of other conceptual spaces, opening up for modeling geometric shapes, faces, language, emotions, and more.

**Acknowledgments** We would like to express our sincere appreciation to all those who have contributed to the completion of this chapter.

First and foremost, we would like to acknowledge the contributions of the authors of this chapter. JF implemented the algorithm, analyzed the results, and wrote the manuscript. OR designed and ran the experiment. OW was the supervisor of the project and provided valuable revision suggestions. The collaboration of these three authors was instrumental in the success of this project.

We would also like to acknowledge the participants who generously gave their time and effort to take part in our experiment. Their contributions were critical to the success of this research.

Lastly, we would like to acknowledge the support and encouragement of our colleagues at Cross Labs, who provided valuable feedback and resources throughout this project.

## References

- Anstis, S. (2002), ‘The Purkinje rod-cone shift as a function of luminance and retinal eccentricity’, *Vision Research* **42**(22), 2485–2491.
- Baronchelli, A., Gong, T., Puglisi, A. and Loreto, V. (2010), ‘Modeling the emergence of universality in color naming patterns’, *Proceedings of the National Academy of Sciences* **107**(6), 2403–2407. <https://www.pnas.org/doi/abs/10.1073/pnas.0908533107>
- Berlin, B. and Kay, P. (1969), ‘Basic Color Terms: Their Universality and Evolution’, *Berkeley & Los Angeles: University of California Press*.
- Bignozzi, C. (2021), ‘The Sapir-Whorf’s Hypothesis: a comparison on colour definition and colour perception in English and Russian speakers.’.
- Brainard, D. H. and Wandell, B. A. (2019), *Color vision: From genes to perception*, Oxford University Press.
- Cook, R. S., Kay, P. and Regier, T. (2005), ‘Chapter 9 - The world color survey database’, pp. 223–241. <https://www.sciencedirect.com/science/article/pii/B9780080446127500640>
- Fairchild, M. D. (2013), *Color appearance models*, 3rd edn, Wiley.
- Freedman, D., Pisani, R. and Purves, R. (2007), ‘Statistics (international student edition)’, *Pisani, R. Purves, 4th edn. WW Norton & Company, New York*.
- Haas, J. K. (2014), ‘A history of the unity game engine’.
- Hanbury, A. (2008), ‘Constructing cylindrical coordinate colour spaces’, *Pattern Recognition Letters* **29**(4), 494–500. <https://www.sciencedirect.com/science/article/pii/S0167865507003601>
- Hartigan, J. A. and Hartigan, P. M. (1985), ‘The Dip Test of Unimodality’, *The Annals of Statistics* **13**(1), 70 – 84. <https://doi.org/10.1214/aos/1176346577>

- Hoijer, H. E. (1954), 'Language in culture; conference on the interrelations of language and other aspects of culture.'
- Ibraheem, N., Hasan, M., Khan, R. Z. and Mishra, P. (2012), 'Understanding Color Models: A Review', *ARPN Journal of Science and Technology* **2**.
- Joblove, G. H. and Greenberg, D. (1978), 'Color Spaces for Computer Graphics', *SIGGRAPH Comput. Graph.* **12**(3), 20–25. <https://doi.org/10.1145/965139.807362>
- Kaneko, S., Anstis, S. and Kuriki, I. (2017), 'Brief presentation enhances various simultaneous contrast effects', *Journal of Vision.* <https://jov.arvojournals.org/article.aspx?articleid=2621973>
- KHATTAK, D. S. R., Ali, H., Khan, Y. and Shah, M. (2018), 'Color psychology in marketing', *Journal of Business & Tourism* **4**(1), 183–190.
- Loreto, V., Mukherjee, A. and Tria, F. (2012), 'On the origin of the hierarchy of color names', *Proceedings of the National Academy of Sciences* **109**(18), 6819–6824. <https://www.pnas.org/doi/abs/10.1073/pnas.1113347109>
- MacLeod, D. I. A. (2003), 'New dimensions in color perception', *Trends in Cognitive Sciences* pp. 97–99. <https://pubmed.ncbi.nlm.nih.gov/21902878/>
- Malkiel, Y. (1956), 'Language in Culture. Conference on the Interrelations of Language and Other Aspects of Culture. Harry Hoijer', *International Journal of American Linguistics* **22**(1), 77–84. <https://doi.org/10.1086/464350>
- Martinovic, J., Paramei, G. V. and MacInnes, W. J. (2020), 'Russian blues reveal the limits of language influencing colour discrimination', *Cognition* **201**, 104281. <https://www.sciencedirect.com/science/article/pii/S0010027720301001>
- Misek, R. (2010), 'Chromatic Cinema: A History of Screen Color'.
- OU, L., Luo, M. and Cui, G. (2008), 'A Colour Design Tool Based on Empirical Studies', *Design Research Society International Conference*.
- Plataniotis, K. N. and Venetsanopoulos, A. N. (2000), 'Color Image Processing and Applications', *SpringerVerlag*.
- Reindl, A., Schubert, T., Strobach, T., Becker, C. and Scholtz, G. (2018), 'Adaptation Aftereffects in the Perception of Crabs and Lobsters as Examples of Complex Natural Objects', *Frontiers in Psychology* **9**. <https://www.frontiersin.org/article/10.3389/fpsyg.2018.01905>
- Smith, A. R. (1978), Color Gamut Transform Pairs, in 'Proceedings of the 5th Annual Conference on Computer Graphics and Interactive Techniques', SIGGRAPH '78, Association for Computing Machinery, New York, NY, USA, p. 12–19. <https://doi.org/10.1145/800248.807361>
- Ulusoy, B., Olguntürk, N. and Kocaoglu Aslanoğlu, R. (2020), 'Colour semantics in residential interior architecture on different interior types', *Color Research and Application* **45**.
- Valdez, P. and Mehrabian, A. (1994), 'Effects of color on emotions', *Journal of Experimental Psychology: General* **123**(4), 394–409.
- Wikimedia Commons (2010a), 'HSL color solid cylinder saturation gray'. [https://commons.wikimedia.org/wiki/File:HSL\\_color\\_solid\\_cylinder\\_saturation\\_gray.png](https://commons.wikimedia.org/wiki/File:HSL_color_solid_cylinder_saturation_gray.png)
- Wikimedia Commons (2010b), 'HSV color solid cylinder saturation gray'. [https://commons.wikimedia.org/wiki/File:HSV\\_color\\_solid\\_cylinder\\_saturation\\_gray.png](https://commons.wikimedia.org/wiki/File:HSV_color_solid_cylinder_saturation_gray.png)
- Wright, W. D. (1941), 'The sensitivity of the eye to small colour differences', *Proceedings of the Physical Society* **53**(2), 93.
- Xiao, B., Hurst, B., MacIntyre, L. and Brainard, D. H. (2017), 'The color constancy of three-dimensional objects', *Journal of Vision.*

# Index

## A

Agent-based modelling (ABM), 2, 7–19  
Analog memory, 5, 33–50

## B

Bell test, 5, 114, 115, 122, 123, 125–128  
Bilingual, 6, 150–152, 154–158, 161, 162

## C

Carnap, 5, 22, 24, 29  
Category learning, 166  
Causality, 113–128  
Cognition, 1–6, 54, 74, 85–96, 102, 109, 110, 117, 165, 166, 168, 178  
Cognitive dynamic, 94  
Cognitive models, 1–4, 78  
Cognitive science, 30, 54, 85, 94, 102, 131, 137, 166  
Color perception, 6, 165–179  
Color space, 6, 148–151, 153, 161, 162, 166, 167, 170  
Compositional semantics, 132  
Compositional vector semantics, 5, 131–143  
Computation, 85–96, 123, 125  
Conjunction fallacy, 5, 101–110  
Contextuality, 114, 124  
COVID-19 protests, 54, 66–68

## D

Data review, 106

Decision making, 1–3, 5, 9, 13, 74, 77, 79, 102, 104, 143  
Dimensions of color, 6, 165–179  
Dissipation, 5, 89–91, 96

## E

Emergence, 2, 5, 76, 85–96  
Entanglement, v, 114, 117, 121–123, 125, 126  
Evolution, 3, 4, 17, 89–92, 94, 96  
Experimental setting, 105, 124

## F

Feedback, 4, 5, 9, 15, 17, 63, 65, 74–79, 81, 90, 179  
Fine-tuning, 120, 121, 124, 127

## H

Human cognition, 1–6, 91, 95, 166, 178  
Human perception, 167, 168, 175–178

## I

Indistinguishability, 57, 58, 65, 70  
Inference, 5, 12, 21–23, 25, 28, 29, 74, 92

## N

Nyayasutra inference, 5  
*Nyāyasūtra*, 4, 5, 21–30

**O**

Optimality, 6, 147–162

**P**

Pheromone trail algorithm, 5, 33–50

Physicalistic perspective, 5, 85–96

Possible experiences, 105–107

Primordial chaos, 92

Probabilistic learning, 5, 73–81

Prototype, 6, 147–162

Pro-war and pro-peace beaming, 54, 67–68

Psychological phenomena, 13

**Q**

Qualitative modeling, 4

Quantum cognition, 117

Quantum computing, 36, 50

Quantum information theory, 122

Quantum modeling, 5, 113–128, 137

Quantum statistics, 57–58

**R**

Reasoning, v, 2, 5, 8, 14, 22, 26–27, 29, 30, 74, 79, 105, 118, 120, 123, 127

Reduction sentence, 5, 22, 27–30

**S**

Self-referential perception, 91

Semantic pointer architecture (SPA), 137, 138, 140–142

Social atom, 5, 54, 55, 60, 69

Social energy, 54–56, 67–68

Social laser, 5, 53–71

Spiking neural networks, 5, 131–143

**V**

Visual perception, 167

**W**

Word representation, 5, 131, 132, 142, 143