Artificial Intelligence & Theory of Mind

Technical Report · August 2016

DOI: 10.13140/RG.2.2.27105.71526

1 author:

Benjamin Erb
Ulm University
32 PUBLICATIONS **66** CITATIONS

SEE PROFILE

ulm university universität ulm

Artificial Intelligence & Theory of Mind

Benjamin Erb

Seminar **Cognition and Emotion**
Summer Term 2016

Dept. Applied Emotional and Motivational Psychology
Institute of Psychology and Education
Faculty of Engineering, Computer Science and Psychology
Ulm University

Artificial Intelligence & Theory of Mind

Author	Benjamin Erb
Course	Psychology (Bachelor)
Contact	benjamin.erb@uni-ulm.de

Lecture	Cognition and Emotion V
Term	Summer term 2016
Course ID	PSY71104.002
Department	Applied Emotional and Motivational Psychology
	Institute of Psychology and Education
	Faculty of Engineering, Computer Science and Psychology
	Ulm University

Version	September 26, 2016
Typesetting	LaTeX, `apa6` Package

Title Image

Untitled by unsplash, released under Public Domain (`CC0` license).

https://unsplash.com/photos/Maf7wdHCmvo/

Abstract

One of the aims of artificial intelligence is the reproduction of human cognition. With increasing complexity of artificial intelligence, advanced cognitive characteristics such as the theory of mind may become relevant in this scheme. Therefore, we briefly discuss the intersection of artificial intelligence and the theory of mind for human-machine interaction by exploring different perspectives on that matter.

Contents

Introduction

The broad-ranging and versatile cognitive abilities and skills of humankind have inspired humans to recreate intelligence in an artificial manner for a long time. This endeavor does not only require a very deep understanding of human cognition and the underlying processes thereof, but also conceptualized models that can be implemented and executed accordingly. Although the notion of artificially created intelligence dates back to ancient times (Russell, Norvig, Canny, Malik, & Edwards, 2003), the advent of computing machinery in the middle of the 20th century finally enabled humans to design, implement, and execute programs with predefined, rational logic. Since then, both expectations and possibilities have grown with technological advancements. At the same time, it has become apparent that the complexity of human cognition cannot be easily captured in its entirety in programming logic. While the computer has surpassed humans in mathematical calculations in a short time, other basal cognitive tasks such as object recognition still represent a major challenge for computers.

One of the cognitive capabilities that separates humankind from many other species is the so-called Theory of Mind (ToM), which captures an individual's ability to attribute mental states to oneself and other individuals (Premack & Woodruff, 1978). In an attempt to model human cognition, ToM capabilities become relevant for advanced artificial intelligences. Vice versa, the interaction of humans with intelligent, non-human agents affects human attribution of mental states to machines.

In the remainder of this paper, we briefly introduce the concepts of artificial intelligence, ToM, and the computational theory of mind. Next, we discuss the interplay of artificial intelligence and the ToM from different perspectives. Finally, we summarize our arguments in a multidisciplinary outlook.

Background

Artificial Intelligence

Artificial intelligence (AI) represents one of the oldest research fields of computer science and the idea of machine-based reasoning even predates actual physical comput-

ing machines (Russell et al., 2003). Based on the theoretical foundation of computation by Turing (1936), the idea of building an artificial brain inspired computer scientists early on. In a seminal event for the research field of AI, McCarthy, Minsky, Rochester, and Shannon (1955) established AI as the "science and engineering of making intelligent machines". A more recent categorization by Russell et al. (2003) suggests AI for systems that (a) *think rationally* (e.g., logical reasoning), (b) *act rationally* (e.g., rational agents), (c) *think like humans* (e.g., cognitive modeling), or (d) *act like humans* (e.g., computer vision and robotics).

Numerous researchers believed that the power of AI is only limited by (ever) increasing computing resources. However, AI research was quickly hit by a trough of disillusionment: While some tasks were easy to program, many general human skills were either far too complex to be implemented or nowhere near well enough understood to be modeled appropriately. Instead, researchers then focused on more specific tasks that could be aligned with prevalent concepts, methods, and technologies — including search, planning, optimization, and reasoning. This subdomain was later coined *weak AI*, as opposed to *strong AI* which represents a general AI capable of solving arbitrary, previously unknown problems based on generalized methods and knowledge.

In the early 21st century, prevailing AI research was complemented with concepts for processing large-scale data sets with statistical methods, i.e., data mining and machine learning (Han, Pei, & Kamber, 2011). Especially deep learning, a machine learning approach that relies on artificial neural networks, recently gained much attention for its unprecedented success in applications such as object recognition, computer vision, and other highly complex tasks (e.g., Mnih et al., 2015).

Theory of Mind

The Theory of Mind pools a number of different attributions of mental states of an individual as part of its self (Ferstl, 2012). Starting with the research of Premack and Woodruff (1978) on primates, the ToM has been established as a cross-cutting issue in many other research fields of psychology, including cognitive, developmental, and social psychology. Although several paradigms have been suggested for testing

an individual's ToM, and deficiencies are already associated with certain disorders, the ToM still remains a theory and an active field of research. For instance, it is still inconclusive whether there are dedicated neural correlates for the ToM, or if the ToM rather emerges jointly from other, subordinate correlates such as consciousness, language processing, or prediction (Spreng, Mar, & Kim, 2009).

Computational Theory of Mind

The Computational Theory of Mind (CTM) is a philosophical movement which suggests the idea that the human mind represents a computational entity for information processing. In spite of the name, the CTM has no direct relationship to the original ToM, but addresses a general perspective on the mind instead. If at all, the ToM traits could be captured in the CTM as a property emerging from lower level processing functions of the mind. The CTM was introduced by Putnam (1976) and later enhanced by his scholar Fodor (1975). Computationalists consider the human brain to be a computing machinery consisting of hardware for computational processing and software that uses representational entities (Thagard, 2007). In the CTM, human experience and thought is solely based on a set of inputs (e.g., stimuli), a number of computing processes and internal states, and corresponding outputs (e.g., behavior).

The use of computationally-inspired models as proposed by the CTM is commonly applied in cognitive psychology, computational neuroscience, and related research areas. However, the philosophical essence of the CTM has been heavily criticized (e.g., Horst, 1999) — and in some instances even by Putnam (1991) and Fodor (2000) themselves in their later research.

The Interplay of Artificial Intelligence and the Theory of Mind

Technological paradigm shifts such as ubiquitous computing (Weiser, 1991) have brought computers into everyday lives. In fact, these computers increasingly disappear as arbitrary entities and become part of the environment, i.e., as part of pervasive and ambient intelligent environments.

Self-driving cars or humanoid assistance robots differ from previous computer applications once they provoke a stronger notion of an intelligent agent. A service robot can take advantage of affective computing and adapt its communication and behavior based on the detected emotional state of its user. A self-driving car adheres to a number of rules and constraints while following different layers of strategies in order to steer the vehicle. Such observations might yield a human feeling that these machines not only possess knowledge and situational awareness, but also follow certain intents (direct or indirect task goals) and own beliefs (probabilistic or uncertain knowledge). Thereby, humans eventually attribute mental states to machines — because of the complex, "intelligent", and human-like behavior of these machines. This raises the open research question whether neural correlates for the ToM can also be triggered by sufficiently advanced machines, and how this notion compares to the traditional ToM experienced in interpersonal interaction.

On the other hand, perceiving an intelligent agent might not necessarily require a machine with ToM traits. When an AI opponent challenges a human player in a game, the human might attribute certain strategies and tactics to the AI. In fact, the AI might only use an algorithm for exploring the mathematical decision space of all potential moves of the entire game and choosing optimal paths on each turn. While following a decision tree obviously does not represent a ToM trait, the human perception thereof might yield different impressions. Similar arguments apply for chat bots such as Eliza (Weizenbaum, 1966), which keep running a conversation by constantly answering a question with a question, or the famous Turing test (Turing, 1950), used to tell apart humans and machines pretending to be humans. Furthermore, most ToM paradigms can be solved by computers, either using rules, facts, and reasoning (e.g., *False Belief Task* using Prolog), or by machine learning (e.g., *Reading the Mind in the Eyes* test using trained deep neural networks). Eventually, these issues can be reduced to a philosophical argument, as suggested by Searle (1980) and his Chinese room thought experiment: A sufficiently complex program *symbolically* executing code that simulates human mind capabilities does not make this program

necessarily possess an actual human-like mind. Consequently, Searle opposed both CTM and the existence of a strong AI.

Regarding the potential emergence of ToM-like behaviors in future machines, singularitarians (e.g., Kurzweil, 2006; Vinge, 1993) — followers of a current of transhumanism — provide a different perspective. Based on the exponential growth according to Moore's Law (i.e., the doubling of performance every 18 months; Moore, 1965), singularitarians argue that computational capacity will reach and surpass the human brain in a few decades. Projecting current growth rates, it will be possible to build artificial neural networks in the scale of an actual human brain around the year 2045 (Kurzweil, 2006). The resulting artificial intelligence explosion could — directly or indirectly — introduce ToM as part of a general artificial intelligence.

Conclusion

In the pursuit of ever-increasing intelligence of machines, the ToM might become a relevant part in human-computer interactions. In intelligent, technical environments, humans may intrinsically apply ToM traits to their non-human interaction partners. When designing humanoid agents, self-driving cars, or ambient technologies, it becomes important to factor in the human ToM. Once humans assign intentions, motives, or beliefs to their non-human counterparts, these thoughts inherently become part of their interaction space and must be considered carefully.

Vice versa, the ToM itself might also become an essential feature of a strong AI, or at least a weak AI sufficiently advanced for interaction. Only the future can tell whether ToM-like traits could be implemented using a well-defined theoretical model thereof, which still has to be developed by then. Alternatively, ToM-based characteristics may also emerge as part of a sufficiently large artificial neural network once its complexity approaches the scale of the human brain. If possible at all, this development will require vast advances in electrical engineering, artificial intelligence, computational neuroscience, and cognitive psychology. Even then, it remains a philosophical question whether a machine emulating a self which includes a concept of the ToM could be regarded as a true individual with actual ToM capabilities, comparable to humans.

References

Ferstl, E. C. (2012). *Theory of Mind. Neurobiologie und Psychologie sozialen Verhaltens*. Springer.

Fodor, J. A. (1975). *The language of thought*. Harvard University Press.

Fodor, J. A. (2000). *The mind doesn't work that way: the scope and limits of computational psychology*. The MIT Press.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Horst, S. (1999). Symbols and computation a critique of the computational theory of mind. *Minds and Machines*, *9*(3), 347–381.

Kurzweil, R. (2006). *Singularity is near*. Penguin Group.

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. (1955). A proposal for the dartmouth summer research project on artificial intelligence.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533.

Moore, G. E. (1965). Cramming more components onto integrated circuits. *Electronics*, *38*(8).

Premack, D. & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*, 515–526.

Putnam, H. (1976). Psychological predicates. *Art, Mind, and Religion*, 429–440.

Putnam, H. (1991). *Representation and reality (representation and mind)*. A Bradford Book.

Russell, S. J., Norvig, P., Canny, J. F., Malik, J. M., & Edwards, D. D. (2003). *Artificial intelligence: a modern approach*. Prentice Hall Upper Saddle River.

Searle, J. R. (1980). Minds, brains, and programs. *The Behaviorial and Brain Sciences*, *3*, 417–424.

Spreng, R. N., Mar, R. A., & Kim, A. S. (2009). The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the

default mode: a quantitative meta-analysis. *Journal of cognitive neuroscience*, *21*(3), 489–510.

Thagard, P. (2007). *Philosophy of psychology and cognitive science (handbook of the philosophy of science)*. North Holland.

Turing, A. M. (1936). On computable numbers, with an application to the entscheidungsproblem. *J. of Math*, *58*(345-363), 5.

Turing, A. M. (1950). I.—computing machinery and intelligence. *Mind*, *59*(236), 433–460.

Vinge, V. (1993). The coming technological singularity: how to survive in the post-human era. *Interdisciplinary Science and Engineering in the Era of Cyberspace*.

Weiser, M. (1991). The computer for the 21st century. *Scientific american*, *265*(3), 94–104.

Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45.