ARL-MR-1054 ● JULY 2022

DEVCOM ARMY RESEARCH LABORATORY

Theory of Mind and Metareasoning for Artificial Intelligence: A Review

by Erin Zaroukian

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

ARL-MR-1054 ● JULY 2022

Theory of Mind and Metareasoning for Artificial Intelligence: A Review

Erin Zaroukian
DEVCOM Army Research Laboratory

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED (From - To)
July 2022	Memorandum Report	December 2021–June 2022

4. TITLE AND SUBTITLE
Theory of Mind and Metareasoning for Artificial Intelligence: A Review

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Erin Zaroukian

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
DEVCOM Army Research Laboratory
ATTN: FCDD-RLC-IT
Aberdeen Proving Ground, MD 21005

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-MR-1054

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release: distribution unlimited.

13. SUPPLEMENTARY NOTES
ORCID ID: Erin Zaroukian, 0000-0002-1381-085X

14. ABSTRACT
Theory of Mind and metareasoning present approaches to artificial intelligence that focus on unobserved mental states, and in doing so they hold promise for improving robustness and collaboration in multi-agent and human–agent systems. This report provides an overview of these approaches.

15. SUBJECT TERMS
multi-agent systems, Theory of Mind, metareasoning, artificial intelligence, decision making, Humans in Complex Systems

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Erin Zaroukian
Unclassified	Unclassified	Unclassified	UU	21	19b. TELEPHONE NUMBER (Include area code) (410) 278-3203

Contents

List of Figures

1. Introduction

Theory of Mind (ToM) and metareasoning, as discussed in the following, have become areas of interest in artificial intelligence (AI) and human–agent teaming. Both hold promise for developing more robust, more collaborative, and even more human-like systems by taking inspiration from humans: ToM, through focus on another's mental states (or their computer analogs), and metareasoning, as a form of reasoning over these mental states. The following two sections present an overview of ToM and metareasoning, including past and potential applications to multi-human/agent systems performing multi-domain operations in complex environments.

2. Theory of Mind

2.1 Theory of Mind within Humans

Theory of Mind (ToM) within psychology describes the human ability to represent and reason about the mental states of others (Premack and Woodruff 1978). A hallmark of this is the ability to recognize false beliefs in others, where a person uses ToM to recognize that the state of the world is inconsistent with another person's beliefs. The Sally–Anne test (Wimmer and Perner 1983) is a classic task to assess this ability to recognize false beliefs. In this task, represented in Fig. 1, a research participant watches a scene with Sally and Anne, where Sally places an item in one location and then leaves. While Sally is gone, Anne moves the item to a new location. When Sally returns, the participant is asked where she will look for the item. If the participant exercises ToM, they should recognize Sally's false belief that the item is still in its original location. Otherwise, they are likely to indicate that Sally will look in the item's actual moved location.

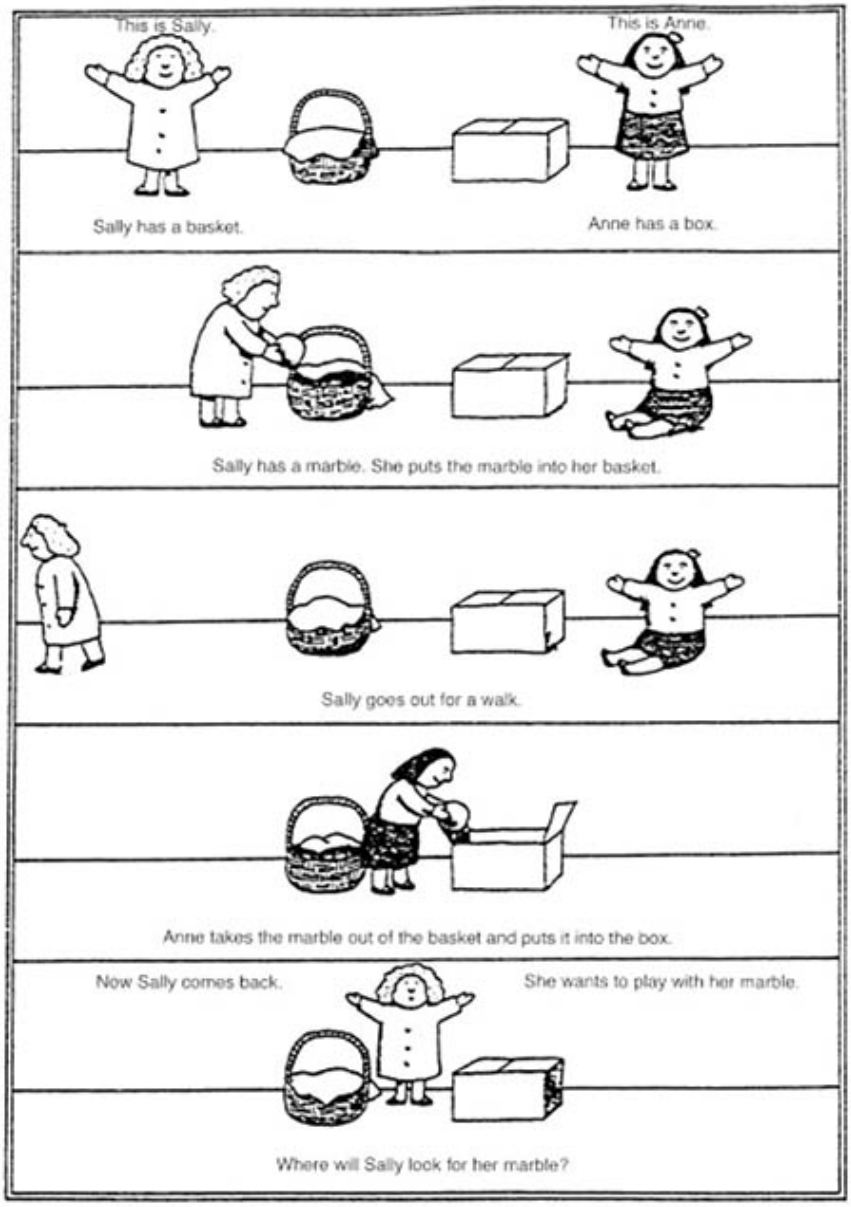

Fig. 1 A depiction of the Sally–Anne task for assessing ToM through a false belief task. (Baron-Cohen et al. 1985).

Tasks like the Sally–Anne task have been used to suggest that ToM is inaccurate or not available in very young children (Wellman et al. 2001), in people with autism (Baron-Cohen et al. 1985), and in nonhuman animals. However, nondeclarative tasks (ones that do not require an explicit answer as in the Sally–Anne task but that measure, e.g., looking time or first looks to locations in a scene) suggest that this type of reasoning is available at younger ages and even in nonhuman primates and corvids (Baillargeon et al. 2010; Horschler et al. 2020; Hampton 2021). Researchers have also emphasized the importance of uncertainty and interactivity

in tests, where ToM is more likely to be exercised in environments with greater uncertainty and asymmetries in knowledge between the research participant and the subject of inference as well as in tests with greater interaction between the participant and the subject of inference (Rusch et al. 2020). Conversely, tests of ToM that do not involve sufficient uncertainty or knowledge asymmetries may fail to find evidence of ToM in participants who do indeed exhibit it in other environments, presumably because they do not sufficiently prompt a participant to consider another's perspective.

Finally, as emphasized in Blaha et al. (2022), those capable of showing evidence of ToM in tests often do not do the same in real-world interactions. In a communication game, Keysar et al. (2003) found that neurotypical adult participants, when given directions from an instructor, would act as though the instructor had accurate knowledge that they (the instructor) were known to lack or even hold false beliefs about. Similarly, Bryant et al. (2013) sampled participants randomly throughout their day to assess how frequently and under what circumstance they considered the mental states of others, and they found that participants rarely thought about mental states, were less likely to think about mental states during a social interaction than when alone, and furthermore were more likely to think of their own mental states than others. These results suggest that adults may find social interactions too cognitively taxing to employ ToM in considering others' mental states on the fly.

2.2 Modeling Human Theory of Mind

Computational modeling of ToM is often undertaken to develop and test cognitive theories of ToM as well as to allow us to build technologies that can interact more naturally and effectively with human users. Much research in this area has shown that Bayesian models can provide an impressive approximation of ToM (Baker et al. 2017; Csibra 2017; see also Yoshida et al. 2008 and Robalino and Robson 2012 for game theory and k-level thinking approaches that apply Bayesian reasoning).

This Bayesian inferencing is often performed through inverse reinforcement learning (IRL). As Jara-Ettinger (2019) describes, "Predicting other people's actions is achieved by simulating a RL model with the hypothesized beliefs and desires, while mental-state inference is achieved by inverting this model" (p. 105), and research has found that "In simple two-dimensional displays, IRL through Bayesian inference produces human-like judgments when inferring people's goals [Baker et al. 2009], beliefs [Baker et al. 2017], desires [Jern et al. 2017], and helpfulness [Ullman et al. 2009]" (p. 105). Inverting Bayesian reasoning, however,

requires strong priors to be successful (Baker et al. 2009). While humans likewise seem to employ strong priors, and while these priors may not always be well justified in novel situations, they are at least fairly transparent within models.

Partially observable Markov decision processes (POMDPs), as part of IRL, have been effectively used to model human ToM, where an agent's actions are observable within an environment but their beliefs and goals must be inferred through inverse planning with an assumption of approximate rationality. Such models, as in Fig. 2, have been shown to provide judgements comparable to those made by humans (Baker 2012). For example, in Baker (2012), human participants + POMDP models observed a simulated agent navigating a simple terrain with occlusion to select a food truck from which to purchase lunch, and they were then asked to provide judgments of the agent's goal (i.e., preferred food truck). In this context, the judgments from a model that allowed for changing goals provided results that matched human judgments closely and better than similar models that did not allow changing goals or that allowed goals to include subgoals. These results suggest that such a model can artificially approximate human judgments and may even be used by human reasoners.

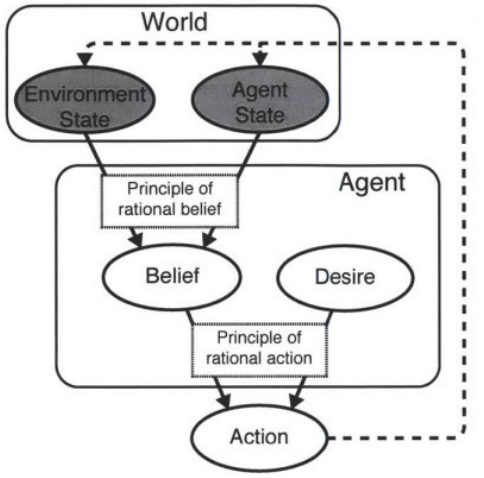

Fig. 2 Causal graph of ToM hypothesized in Baker (2012), where observed (gray) information influences an agent's (unobserved) beliefs, desires, and ultimately actions, mediated by rationality. (Image adapted with permission from Baker [2012].)

Similar work has explored more-complex reasoning environments, such as using a meta-Bayesian framework to model human ToM under conditions of varying trustworthiness (Diaconescu et al. 2014). Participants played an economic game in which they were given (veridical, nonsocial) probabilistic information to help them make a choice in a binary lottery, and they were also given (social) information from an advisor whose incentive for helping the player varied. Human results were best modeled by a hierarchical model that can assign different weights to social and

nonsocial information, and that allows inferences about an advisor's changing goals based on dynamic estimates from past performance.[*] While this and other (e.g., Meinhardt-Injac et al. 2018) work has explored human inference using cues from social versus nonsocial sources, research does not seem to have modeled ToM involving a wider variety of uncertain information sources and how humans prioritize and incorporate them in decision-making processes.

These types of models of ToM suggest that ablating parts of the model may be possible, with predictable results. For example, if humans use Bayesian priors when reasoning about others, and we update these priors based on experience, there may be individuals who are unable to effectively update priors. Indeed, this may be the case in schizophrenia, where people hold particularly negative and suspicious views of others that are not ameliorated through positive interactions. Autism, on the other hand, may represent a more general ablation of (access to) the entire ToM mechanism, as people with autism tend to perform poorly on tests of ToM (TEDx Talks 2014; Prevost et al. 2015).

2.3 Theory of Mind within Machines

Computational ToM is also used not to directly model human reasoning, but as a framework to allow agents to reason about other agents. Additionally, such agents may be more interpretable and lead to better human–agent interactions. ToM can allow an agent to hold appropriate priors about other agents even before they have been encountered, update beliefs about them, and recognize their false beliefs.

This is illustrated in Rabinowitz et al. (2018), where models learned to recognize different species of agents (e.g., one species tends to pursue a nearby object versus a faraway object) by predicting their future behavior based on their past behavior. Notably, this work included a false belief task, where an agent with limited vision observed its final goal object, which probabilistically changed position either within or outside of the agents view as it first pursued a subgoal object. The agent was observed to then pursue the final goal object in its original position more often when it had moved unobserved than when the move had been observed. This suggests that the model learned to represent its full knowledge of the environment separately from the agent's limited knowledge in a way that allowed it to recognize the agent's false beliefs.[†]

[*]This work supports results from fMRI (functional magnetic resonance imaging) and other modeling studies that find separate mechanisms for processing social and nonsocial information (Behrens et al. 2008).

[†]An important distinction between the work by Rabinowitz et al. (2018) and Baker (2012) is that the Rabinowitz models were learned whereas the Baker models were handcrafted a priori based on

2.4 Theory of Mind between Humans and Machines

When humans are able to make accurate inferences about machines' "states of minds," it likely improves trust and performance. This has become a motivation within the field of explainable AI for improving trust and performance (Akula et al. 2019). Similarly, as machines are more accurately able to infer human intentions, their utility increases and they further gain trust (Winfield 2018).

3. Metareasoning

Metareasoning is a general AI term for "thinking about thinking" within a computational system.[*] While reasoning algorithms are used to make decisions, a *meta*reasoning algorithm is used to control a reasoning algorithm or to select among a set of reasoning algorithms, determining which decision-making method should be used under different circumstances (Cox and Raja 2011). A classic example of metareasoning is to determine whether a reasoning algorithm should stop or continue in a given context (e.g., Carlin 2012).

Metareasoning can be described as in Fig. 3, where reasoning occurs at the Object Level based on observations at the Ground Level, and the decisions made at the Object Level are enacted at the Ground Level. For example, a sensored alarm might sound at the Ground Level when an algorithm at the Object Level determines from sensor input that an intruder was present (e.g., this algorithm may sound an alarm when two or more motion events are detected within a 10-s window). Metareasoning then occurs when information from the Object Level is observed and altered at the Meta-Level. In the previous example, an algorithm at the Meta-Level might adjust the sensitivity of the alarm if it is triggered too often, causing battery issues (e.g., this Meta-Level algorithm might impose a new algorithm at the Object Level that sounds the alarm only when *three* or more motion events are detected within a 10-s window).

cognitive theories. Thus, while Baker's method provides a way to test theories, Rabinowitz's method will likely scale better for real-world application.

[*]Metareasoning is a topic of interest in other fields as well, such as philosophy and psychology, where it is often referred to as "metacognition" (Cox 2005).

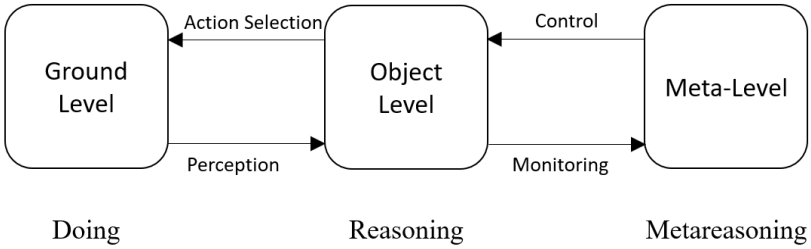

Fig. 3 Classic decision–action loop diagram of metareasoning, where reasoning happens at the object level to select the actions that will happen at the ground level, and metareasoning happens at the meta-level to control what occurs at the object level

Metareasoning can occur within a single agent, as depicted in Fig. 3, or it can occur within a multi-agent system (MAS) (Fig. 4). Metareasoning is often used in a multi-agent setting to optimize the performance of an entire system, and there are many options for how it is implemented with different consequences for resources such as time and compute power. For example, agents within an MAS may perform their metareasoning independently and communicate at the Object Level, which may be a good solution when communication is costly and coordination is a low priority. When coordination is more important, independently metareasoning agents may communicate at the Meta-Level to jointly determine how they will independently metareason (Langlois et al. 2020).

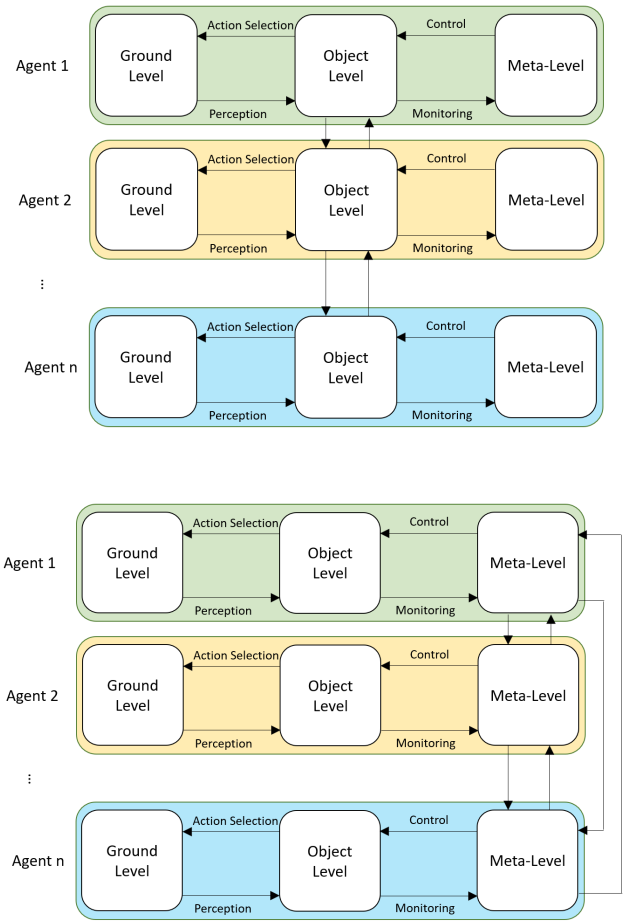

Fig. 4 (top) An MAS system where metareasoning occurs independently for each agent. (bottom) An MAS where each agent's metareasoning communicates and coordinates with each other agent's metareasoning. (Diagrams from Langlois et al. [2020] with permission.)

Metareasoning can also be performed in a more centralized fashion by separate metareasoning agents (Fig. 5, top). As communication resources allow, the best coordination and metareasoning is expected to come from a single centralized metareasoning agent (Fig. 5, bottom) (Langlois et al. 2020).

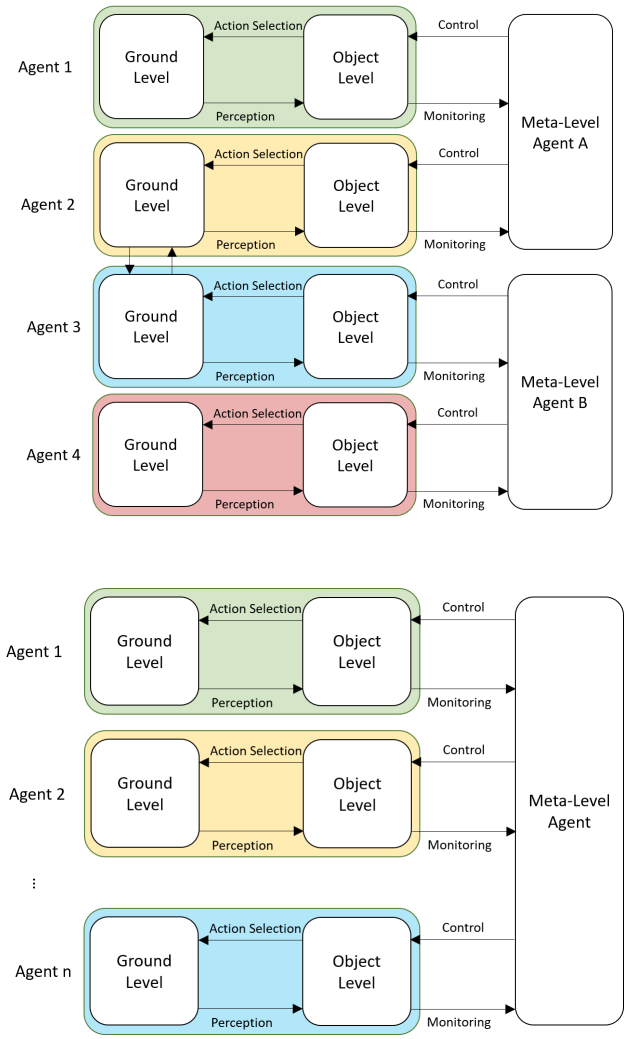

Fig. 5 (top) An MAS with multiple separate metareasoning agents. (bottom) An MAS with a single centralized metareasoning agent. (Diagrams from Langlois et al. [2020] with permission.)

Systems also vary in the object of their metareasoning. Single agent metareasoning is often used, as described in the opening of this section, to control algorithm halting or switching and applied to a wide variety of fields, including scheduling and planning (e.g., Lin et al. 2015), heuristic search (Gu 2021), and object detection (e.g., Parashar and Goel 2021). Within MASs, metareasoning is often used to control communication and resources within the systems, including controlling communication frequency or content, or assigning tasks (Herrmann 2020).

An additional concern in metareasoning is how much learning or metareasoning should happen online versus offline. Because online metareasoning can be costly in terms of time and computation, offline policies are often maximized to the extent that they do not unduly impair system accuracy (e.g., Carrillo et al. 2020).

Broadly speaking, ToM is a form of metareasoning, or "thinking about thinking." As described in Fig. 3, however, metareasoning is performed through monitoring and controlling the Object Level, whereas ToM involves making inferences from what is happening at the Ground Level without direct access to the Object Level (e.g., an agent's beliefs).

4. Conclusion

While metareasoning is already widely used in single- and multi-agent systems to improve performance, ToM approaches arguably have not been explored as deeply as a method to improve performance of an artificially intelligent agent. This is almost certainly in part because ToM is more closely tied to human cognition, which places strong restrictions on plausible ToM models and biases research toward human applications. Additionally, ToM itself is still somewhat controversial (e.g., Who has it? When is it acquired? Under what conditions is it exercised?) but it holds promise for creating more-transparent (if not authentically human) systems, especially systems reasoning with multiple sources of information and with differing provenance and certainty. In particular, recent computational ToM approaches, which use simpler, heuristic definitions of ToM (e.g., Rabinowitz et al. 2018), may be the best source of innovation in this field.

5. References

Akula AR, Liu C, Saba-Sadiya S, Lu H, Todorovic S, Chai JY, Zhu S. X-ToM: explaining with theory-of-mind for gaining justified human trust. arXiv; 2019. http://arxiv.org/abs/1909.06907.

Baillargeon R, Scott RM, He Z. False-belief understanding in infants. Trends in Cognitive Sciences. 2010;14(3):110–118.

Baker CL. Bayesian theory of mind: modeling human reasoning about beliefs, desires, goals and social relations [dissertation]. Massachusetts Institute of Technology; 2012.

Baker CL, Saxe R, Tenenbaum JB. Action understanding as inverse planning. Cognition. 2009;113:329–349.

Baker C, Jara-Ettinger J, Saxe R, Tenenbaum JB. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. Nat Hum Behav. 2017;1:0064.

Baron-Cohen S, Leslie AM, Frith U. Does the autistic child have a "theory of mind"? Cognition. 1985;21(1):37–46.

Behrens TEJ, Hunt LT, Woolrich MW, Rushworth MFS. Associative learning of social value. Nature. 2008;456(7219):245–249.

Blaha LM, Abrams M, Bibyk SA, Bonial C, Hartzler BM, Hsu CD, Khemlani S, King J, St Amant R, Trafton JG, Wong R. Understanding is a process. Frontiers in Systems Neuroscience. 2022;16.

Bryant L, Coffey A, Povinelli DJ, Pruett John RJ. Theory of mind experience sampling in typical adults. Conscious Cogn. 2013;22:697–707. doi: 10.1016/j.concog.2013.04.005

Carlin AS. Decision-theoretic meta-reasoning in partially observable and decentralized settings [dissertation]. University of Massachusetts Amherst; 2012.

Carrillo E, Yeotikar S, Nayak S, Jaffar MKM, Azarm S, Otte M, Xu H. Communication-aware multi-agent metareasoning for decentralized task allocation. University of Maryland; 2020.

Cox MT. Metacognition in computation: a selected research review. Artificial Intelligence. 2005;169(2):104–141.

Cox MT, Raja A. Metareasoning: an introduction. In: Metareasoning: thinking about thinking. MIT Press; 2011. p. 3–14.

Csibra G. Cognitive science: modelling theory of mind. Nat Hum Behav. 2017;1. Article 0066.

Diaconescu AO, Mathys C, Weber LAE, Daunizeau J, Kasper L, Lomakina EI, Fehr E, Stephan KE. Inferring on the intentions of others by hierarchical Bayesian learning. PLoS Comput Biol. 2014;10(9):e1003810.

Gu T. Metareasoning for heuristic search using uncertainty [dissertation]. University of New Hampshire; 2021.

Hampton RR. Designer receptor inhibition suggests mechanism for monkey theory of mind. Learning & Behavior. 2021;49(2):171–172.

Herrmann J. Data-driven metareasoning for collaborative autonomous systems. Institute of Systems Research, University of Maryland; 2020.

Horschler DJ, MacLean EL, Santos LR. Do non-human primates really represent others' beliefs? Trends in Cognitive Sciences. 2020;24(8):594–605.

Jara-Ettinger J. Theory of mind as inverse reinforcement learning. Current Opinion in Behavioral Sciences. 2019;29:105–110. ISSN 2352-1546. https://doi.org /10.1016/j.cobeha.2019.04.010.

Jern A, Lucas CG, Kemp C. People learn other people's preferences through inverse decision-making. Cognition. 2017;168:46–64.

Keysar B, Lin S, Barr DJ. Limits on theory of mind use in adults. Cognition. 2003;89:25–41. doi: 10.1016/S0010-0277(03)00064-7.

Langlois ST, Akoroda O, Carrillo E, Herrmann JW, Azarm S, Xu H, Otte M. Metareasoning structures, problems, and modes for multiagent systems: a survey. IEEE Access 8. 2020. p. 183080–183089. doi: 10.1109/ACCESS .2020.3028751.

Lin CH, Kolobov A, Kamar E, Horvitz E. Metareasoning for planning under uncertainty. Proceedings of the 24th International Conference on Artificial Intelligence; 2015. p. 1601–1609.

Meinhardt-Injac B, Daum MM, Meinhardt G, Persike M. The two-systems account of theory of mind: testing the links to social-perceptual and cognitive abilities. Frontiers in Human Neuroscience. 2018;12.

Parashar P, Goel AK. Meta-reasoning in assembly robots. In: Lawless, WF, Mittu R, Sofge DA, Shortell T, McDermott TA, editors. Systems engineering and artificial intelligence. Springer; 2021. p. 425–449.

Premack D, Woodruff G. Does the chimpanzee have a theory of mind? Behavioral and Brain Sciences. 1978;1(4):515–526.

Prevost M, Brodeur M, Onishi KH, Lepage M, Gold I. Judging strangers' trustworthiness is associated with theory of mind skills. Frontiers in Psychiatry. 2015;6.

Rabinowitz N, Perbet F, Song F, Zhang C, Eslami SA, Botvinick M. Machine theory of mind. Proceedings of the International Conference on Machine Learning. Proceedings of Machine Learning Research; 2018. p. 4218–4227.

Robalino N, Robson A. The economic approach to 'theory of mind.' Philosophical Transactions of the Royal Society B: Biological Sciences. 2012;367(1599):2224–2233.

Rusch T, Steixner-Kumar S, Doshi P, Spezio M, Gläscher J. Theory of mind and decision science: towards a typology of tasks and computational models. Neuropsychologia. 2020;146:107488.

TEDx Talks. Theory of mind through the lens of algorithms: Andreea Diaconescu. [Video]. 2014 Nov 20. https://www.youtube.com/watch?v=a19lSbFPRqY.

Ullman T, Baker C, Macindoe O, Evans O, Goodman N, Tenenbaum J. Help or hinder: Bayesian models of social goal inference. Advances in Neural Information Processing Systems. 2009;22.

Wellman HM, Cross D, Watson J. Meta-analysis of theory-of-mind development: the truth about false belief. Child Development. 2001;72(3):655–684.

Wimmer H, Perner J. Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. Cognition. 1983;13(1):103–128.

Winfield AF. Experiments in artificial theory of mind: from safety to story-telling. Frontiers in Robotics and AI. 2018;5:75.

Yoshida W, Dolan RJ, Friston KJ. Game theory of mind. PLoS Computational Biology. 2008;4(12):e1000254.

List of Symbols, Abbreviations, and Acronyms

AI	artificial intelligence
fMRI	functional magnetic resonance imaging
IRL	inverse reinforcement learning
MAS	multi-agent system
POMDP	partially observable Markov decision process
ToM	Theory of Mind

1 (PDF)	DEFENSE TECHNICAL INFORMATION CTR DTIC OCA
1 (PDF)	DEVCOM ARL FCDD RLD DCI TECH LIB
1 (PDF)	DEVCOM ARL FCDD RLC IT E ZAROUKIAN