Efficient and Effortful Theory of Mind Reasoning in the AToM Cognitive Model

Irina Rabkina (irabkina@oxy.edu)
Occidental College, Los Angeles, CA

Clifton McFate (cjm@ec.ai)
Elemental Cognition, New York, NY

Abstract

Apperly and Butterfill (2009) argue that adult theory of mind (ToM) requires two parallel systems. One system, efficient but inflexible, enables rapid judgements by operating without explicit modeling of beliefs, while a separate, effortful system, enables richer predictions over more complex belief encodings. Here, we agree with their qualitative distinction but propose a different model: a single process, but with effortful re-representation leading to two phases of ToM reasoning. Efficient reasoning, in our view, occurs over representations that include actions, but not necessarily explicit belief states. Effortful reasoning, then, involves re-representation of these initial encodings in order to handle errors, resolve real-world conflicts, and fully account for others' belief states. We present an implemented computational model, based in memory retrieval and structural alignment, that illustrates our approach.

Keywords: Analogy; Theory of Mind; Computational Modeling

Introduction

While the precise trajectory of human theory of mind (ToM) development continues to be debated (Onishi & Baillargeon, 2005; Wellman & Liu, 2004; de Villiers, 2021; Kovács, Téglás, & Csibra, 2021) it has been well established that young children often fail to take into account the mental states of others when predicting their actions. However, typically developing adults (and older children) are generally considered to be proficient ToM reasoners.

Yet, there is substantial evidence that adults do not always effectively utilize ToM, either—at least not automatically. Keysar, Barr, Balin, and Brauner (2000) demonstrated that adult participants in a diadic communication game often considered visual referents that their partner could not be aware of. These failures were further explored by Keysar, Lin, and Barr (2003), who required participants to give a ”director” an object from a table. The names of the objects on the table were polysemous (e.g., a roll of tape and a cassette tape), but, prior to the direction, the participant themselves hid one of the possible referents in a bag, leaving only one visible to the director. Even so, the participants sometimes gave the director the occluded object. This was the case even when they were told the director had a false belief about the contents of the bag that excluded the actual contents as a referent.

There is also evidence that ToM reasoning requires cognitive effort. Apperly, Riggs, Simpson, Chiavarino, and Samson (2006) found that adults were slower to answer questions about another person's false beliefs than about reality, but that this processing difference disappeared when the participants were instructed to track beliefs explicitly. Lin, Keysar, and Epley (2010) further found that working memory impairment degraded ToM reasoning, suggesting that humans are ”reflexively mind blind”, only explaining behavior with regard to mental states when cognitive resources allow.

These and other findings led Apperly and Butterfill (2009) to argue that adult humans have two systems for theory of mind. Per Apperly and Butterfill (2009), the first ToM system—efficient but inflexible—enables real time goal recognition but does not explicitly encode mental states. The second system does encode mental states and enables full ToM reasoning, but requires cognitive effort. They suggest that the first system is shared by young children and potentially non-human animals as well, while the latter develops with maturation, thus explaining ToM's developmental trajectory.

However, there is strong evidence that said trajectory is not strictly maturational. It has been found that even very young children can succeed at complex ToM tasks, given the right scaffolding(e.g. Hale & Tager-Flusberg, 2003; Hoyos, Horton, Simms, & Gentner, 2020). Taken together, these findings suggests that, instead of two separate ToM systems, ToM is better thought of as a continuum, with effort, experience, and cognitive control driving development.

We propose that both effortful and efficient ToM can be explained by building on the Rabkina, McFate, Forbus, and Hoyos (2017) Analogical Theory of Mind (AToM) framework. AToM conceptualizes ToM reasoning as retrieval and application of a person's prior experiences via analogical inference (Rabkina et al., 2017). We suggest that efficient ToM arises from a single retrieval and inference via AToM, while effortful ToM requires iterative re-representation of the situation being evaluated and re-retrieval of relevant analogical comparisons. The distinction between efficient and effortful ToM, then, is at the level of effort applied, rather than a qualitative difference in process.

As a child matures, they gain executive control, a richer representational vocabulary, and a larger library of memories to draw from, all of which contribute to improved ToM reasoning. The quick and efficient reasoning based on a single retrieval, however, does not disappear. Thus, our account explains both developmental phenomena as well as the persistence of efficient, but incomplete, ToM into adulthood.

In the following sections, we begin by discussing compet-

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society*.

ing developmental accounts of theory of mind with regard to the phenomena described above. We then introduce our framework, the Analogical Theory of Mind, and demonstrate how we extend it to account for adult ToM failures and the efficiency trade-off proposed by Apperly and Butterfill (2009). We conclude with a motivating example of efficient vs effortful ToM reasoning and a corresponding simulation using the extended implemented AToM computational model.

Developmental Accounts of Theory of Mind

Developmental accounts of theory of mind reasoning, broadly speaking, fall under the umbrellas of ”Theory Theories”, ”Simulation Theories”, and ”Hybrid Theories”. Theory theories propose that ToM reasoning occurs with respect to a set of rules that can be applied to predict the beliefs, desires, and mental states of others (Gopnik & Wellman, 1994). Many Theory theories have been conceptualized as ”child scientist” theories. According to such a theory, a child might start with a simple rule (e.g., everyone knows what I know). As interactions with other individuals invalidate that hypothesis, the child generates alternatives and eventually settles on a mature rule-based model. Consistent with this account, Wellman and Liu (2004) found that ToM development proceeds with consistent phases across children which could correspond to discrete updated hypotheses.

However, such accounts struggle to explain findings like those of Keysar et al. (2000, 2003) which illustrate that even adults sometimes rely on a more primitive egocentric ToM model. If adults do develop abstract rules, they do not seem to apply them consistently.

Conversely, Simulation theories do not rely on rule-like models and instead argue that ToM requires simulation of another person’s actions from a first person perspective, colloquially, ”putting yourself in someone’s shoes” (Goldman et al., 2006). The first-person simulation account is consistent with findings of an egocentric bias in ToM reasoning failures (Goldman & Sebanz, 2005). Furthermore, simulation as a cognitive process flexibly incorporates efficiency trade-offs.

On the other hand, simulation accounts struggle to explain broad developmental shifts like those found by Wellman and Liu (2004) and are inconsistent with findings that adults seem to apply mental models of the intentions of others that they would not apply to themselves (Saxe, 2005).

Hybrid theories (see Bach, 2011) combine elements from both Theory theories and Simulation theories to explain both developmental phenomena and adult ToM errors. In the following section we present one such model, the Analogical Theory of Mind (AToM) model, and discuss how it can resolve the issues discussed above.

Analogical Theory of Mind

Analogical Theory of Mind (AToM; Rabkina et al., 2017) is a computational cognitive model of human ToM reasoning and development based on a theoretical model initially proposed by Bach (2011, 2014). According to the AToM model, theory of mind reasoning is the result of analogical inference from a retrieved structurally similar memory.

As a motivating example, consider a situation where a person sees a colleague, Sam, walk towards the office kitchen without a coffee cup. AToM claims that, in order to predict Sam’s beliefs, the person retrieves structural similar memories via analogical retrieval (Forbus, Gentner, & Law, 1995). For example, they may recall that they themselves had been surprised earlier to find no coffee cups in the kitchen, or a time when another teammate, Alex, had walked through with no identifiable purpose. These memories are then compared to the current situation using structural alignment, with the most similar memory being used to generate candidate analogical inferences (e.g., perhaps Sam, like the person, expects coffee cups in the kitchen).

Consistent with Simulation theories, a retrieved memory could be encoded from a first-person perspective (e.g., what would I do in a situation). However, memories may also be encoded in a third-person perspective (e.g., what did someone else do in a situation). Furthermore, they may not include explicit representations of internal beliefs at all if they weren’t relevant at the time of encoding. For example, Rabkina et al. (2017) trained AToM with stories using explicit representations of belief states to model how children learn ToM from hearing stories about others’ true and false beliefs. On the other hand, Rabkina, McFate, and Forbus (2018) modeled how children gain ToM from learning a complex grammatical structure; while the nested structure of representations played an important role in that model, belief states were not encoded at all.

AToM assumes that memories are retrieved, applied, and generalized within long term memory, allowing for the formation of rule-like schemas via analogical generalization (McLure, Friedman, & Forbus, 2015). Over time, frequently occurring ToM scenarios (e.g., object occlusion) become abstracted from individual objects or agents and function more like a rule (Gentner & Medina, 1998). Thus, AToM is capable of generating both simulation-like and rule-like judgements depending on retrieval.

Analogical Retrieval and Inference

Retrieval and alignment follow the principles of Structure Mapping Theory (SMT; Gentner, 1983). At a high level, SMT proposes that analogy, and comparison more broadly, involves a process of analogical alignment over structured representations, called a base and target. This alignment is governed by three hard constraints. The 1-1 constraint says that each item in the base can align to at most one item in the target. The parallel connectivity constraint requires that for any aligned relationship, its arguments are also aligned. Finally, the identicality constraint aligns only identical relationships unless their alignment is supported by participation in a larger structure. SMT also argues for a preference for aligned higher-order (e.g., causal) structure over shared lower-level features (e.g., size or shape). Given an alignment, parallel structure that is present in one case but absent in the other to

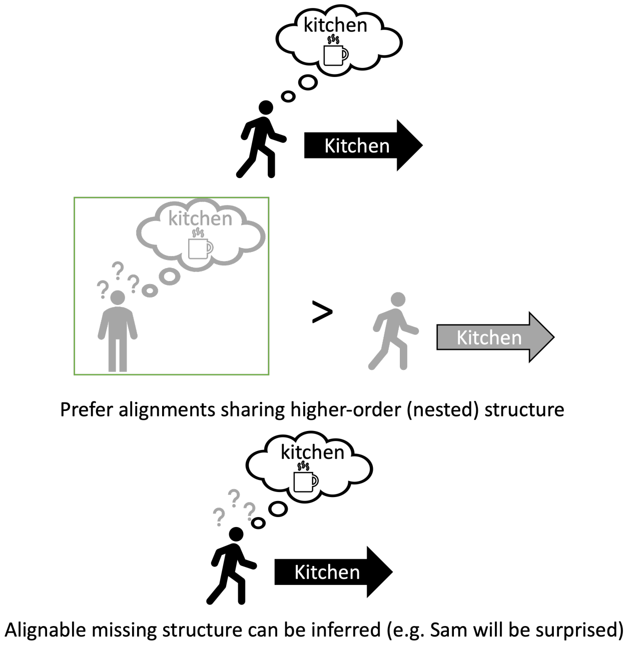

kitchen
Kitchen
kitchen
? ? ?
>
Kitchen
Prefer alignments sharing higher-order (nested) structure
kitchen
? ? ?
Kitchen
Alignable missing structure can be inferred (e.g. Sam will be surprised)

Figure 1: Analogical Alignment and Inference

be projected as an analogical inference.

As an example, we return to our colleague, Sam, walking to the kitchen (depicted in Figure 1). In the base case (top), Sam goes to the kitchen, and the retrieved memories (Self vs Alex) act as the targets. Formally, we could represent the self memory with the nested belief proposition `(believes self (locationOf cup kitchen))`[1] and their state of surprise, `(surprised self)`. The other contains the fact of Alex's walking to the kitchen `(walks Alex kitchen)`, but no representation of a belief or goal.

If the representation of Sam going to the kitchen contains both the walking and the assumption that there are cups, the first memory will align better based on the shared nested belief structure and should be retrieved. On the basis of the alignment, it can be inferred that Sam will also be surprised. On the other hand, if no belief representation is included, the second memory will be retrieved and no ToM expectations will be justified.

Different permutations of these facts in memories and scenario representations will lead to different retrievals, and therefore different reasoning outcomes. Note, too, that, while in this toy example, representations of the scenario and retrieved memories are exactly the same, such exact matching is not required.

Efficient and Effortful Reasoning in AToM

As discussed above, there is evidence that full ToM reasoning is not automatic and in fact requires substantial cognitive effort (Keysar et al., 2003; Lin et al., 2010; Apperly & Butterfill, 2009). AToM provides a mechanism by which ToM inference occurs, namely analogical retrieval of and mapping from episodic memories. Here we propose that this process underlies both efficient and effortful ToM reasoning, the latter being the result of an iterative sequence of retrieval and re-representation.

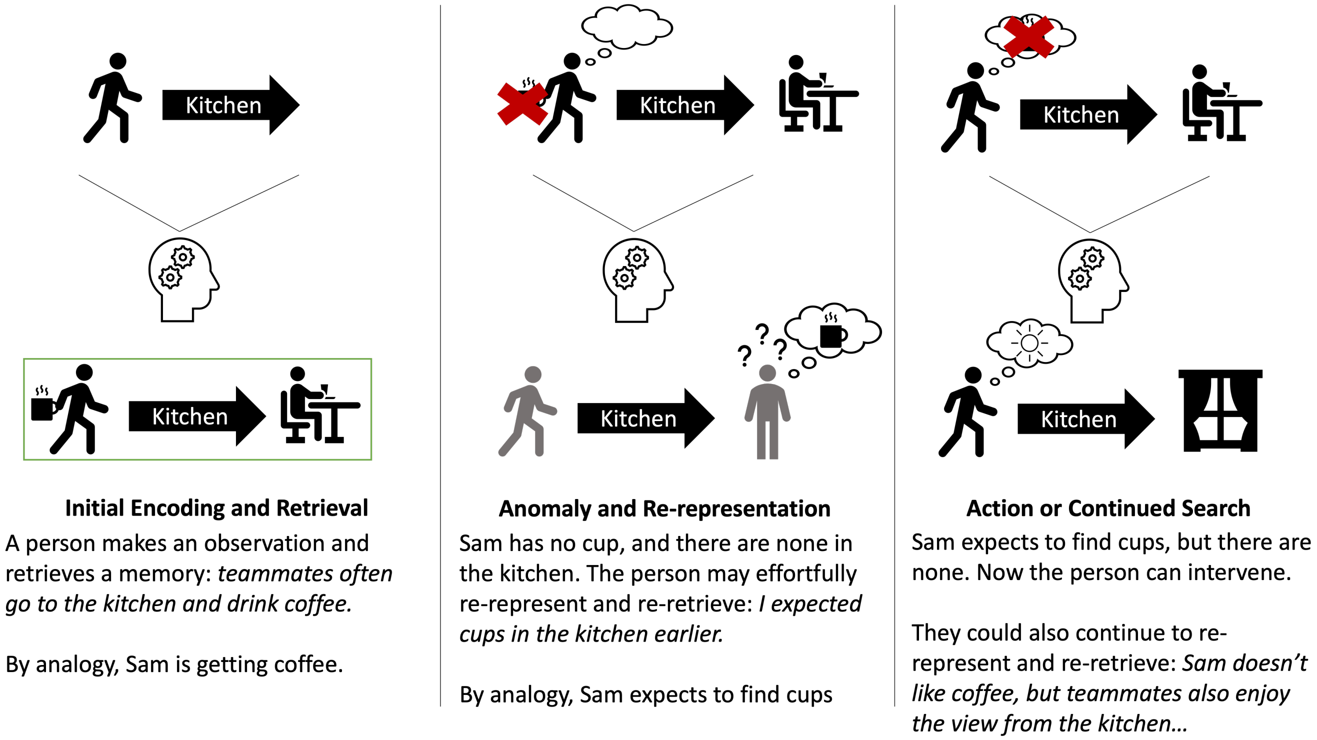
When observing a potential ToM reasoning scenario (e.g., Sam going to the kitchen), a person initially encodes their observation using sparse representations that do not include belief and knowledge states. They then retrieve an analogical experience from memory and infer a potential goal. As in Rabkina et al. (2017), this inference may be incompatible with the real world which triggers a search for explanation via further analogical retrieval.

In our proposed model, the incompatible inference leads to re-representation of the scenario given the false expectations generated by the alignment. This process of inference evaluation, re-representation, and retrieval requires additional cognitive effort and is subject to executive control.

Returning to our example, in Figure 2 the person observes Sam going to the kitchen. This prompts a search for explanation using the person's initial encoding of the situation. The person recalls that teammates often go to the kitchen and drink coffee. By aligning Sam to prior teammates, they can infer that Sam is likewise getting coffee.

However, in the retrieved memory, the teammate needed to bring a cup in order to drink coffee. Sam does not have a cup, triggering what, in analogy literature, is called an alignable difference (Markman & Gentner, 1996). This difference is re-represented into the scenario and the person again searches

[1] We use CycL-style representations (Lenat & Guha, 1991) with the NextKB knowledge base (Forbus & Hinrich, 2017) in this work.



Initial Encoding and Retrieval
A person makes an observation and retrieves a memory: *teammates often go to the kitchen and drink coffee.*

By analogy, Sam is getting coffee.

Anomaly and Re-representation
Sam has no cup, and there are none in the kitchen. The person may effortfully re-represent and re-retrieve: *I expected cups in the kitchen earlier.*

By analogy, Sam expects to find cups

Action or Continued Search
Sam expects to find cups, but there are none. Now the person can intervene.

They could also continue to re-represent and re-retrieve: *Sam doesn't like coffee, but teammates also enjoy the view from the kitchen…*

Figure 2: ToM through successive analogical retrieval and re-representation

for explanation. The person now retrieves the memory of themselves being surprised to find the kitchen did not have extra cups. So perhaps, by analogy, Sam is also expecting there to be cups in the kitchen.

Now primed with Sam's inferred beliefs, the person can continue to elaborate. Maybe this belief is inconsistent with reality (i.e., Sam knows there are no cups) or maybe the person remembers that Sam does not like coffee. They can continue to re-represent and re-retrieve explanations until satisfied or decide that further effort isn't worthwhile.

In Rabkina et al. (2017) the retrieved memories were encoded from a first person perspective (e.g., ”I once got coffee”; see Meltzoff, 2007 for role of such ”like me” encoding in reasoning). In the automatic representations in Rabkina et al. (2020), on the other hand, representations were allocentric (e.g., ”My teammate once got coffee”). We note here that, due to analogical alignment, ToM reasoning can arise from both egocentric and allocentric memories. The ToM target could align to oneself or one's mental model of another. Interestingly, this suggests that individual encoding biases as well as alignability between the self and the target may play a significant role in ToM prediction. If the target is easily aligned to the self, a person may be more likely to ascribe their own motivations and beliefs to the target. However if viewed as different, they may be more likely to apply the perceived mental model of a more similar analog or even fail to model ToM entirely.

Simulation

Here we provide a computational proof of concept for how the AToM model incorporates iterative re-representation to overcome initial shallow ToM judgements and describe the algorithmic changes made to the AToM model. Our simulation is implemented as hierarchical task network (HTN; Erol, 1995) plans executable in the Companion cognitive architecture (Forbus & Hinrich, 2017)[2]. For this simulation, we use the running example from above: I observe my colleague, Sam, walking to the kitchen.

Recall that the AToM model uses analogical retrieval and inference for reasoning. In the previous implementation (i.e., Rabkina et al., 2017, 2018), this was a single-shot process. That is, a single a memory was retrieved and used for reasoning. While re-retrieval was possible during learning, there was no re-representation or re-retrieval during reasoning. Here, we have extended the reasoning stage to allow for both re-representation and re-retrieval when indicated by an executive. For the purposes of this proof of concept, re-representation and re-retrieval occurred when a contradiction was observed between an inference and the real world, so long as a retrieval depth limit (corresponding to working memory limitations and intended effort level) had not been reached. This depth limit was set to allow up to 5 re-retrievals in our simulation.

Cases corresponding to Figure 2 were encoded in predicate

[2] Please contact the authors for original plans

calculus using using the NextKB knowledge base (Forbus & Hinrich, 2017). This included the initial observation of Sam walking to the kitchen, the allocentric memory of teammates often going to the kitchen to get coffee, and the egocentric memory of expecting to find coffee cups in the kitchen and not finding them there. We assumed that such appropriate memories would be available and did not include distractor memories. Note that distant distractors would not be retrieved by analogical retrieval, while close distractors may lead to different ToM conclusions (e.g., if a memory of wanting tea was retrieved, the model would infer that Sam wants tea, rather than coffee).

The initial retrieval returned the memory of teammates getting coffee in the kitchen. By analogical inference, AToM predicted that Sam had the goal of getting coffee and that they should be carrying a cup as a pre-requisite for this goal. AToM then compared that prediction to the real world (in this case, another predicate calculus representation that included the fact that Sam is not carrying a cup), and found a contradiction. Given this contradiction, AToM re-represented the observation of Sam to include the fact that they are not carrying a cup. Using the updated observation, it retrieved the memory of expecting to find cups in the kitchen. This generated three analogical inferences:

1. Sam is walking to the kitchen to get coffee.
2. Sam is not carrying a cup because they believe there are cups in the kitchen.
3. Sam will be surprised that there are no cups.

Because the depth limit was not yet reached, AToM checked these inferences against the real world for contradictions and, because no contradictions were found, accepted the inferences. Thus, the ToM model inferred Sam's goal, knowledge, and future emotional state, allowing for the executive to act upon these inferences.

When the depth limit was lowered to 1, allowing only a single retrieval and indicating disinterest in Sam's activities or limited working memory capacity (e.g., due to concentrating on another task), only the memory of teammates drinking coffee was retrieved. In this case, AToM also inferred Sam's goal (drinking coffee), but did not recognize that Sam had a mistaken belief about the location of cups. Thus, the executive did not receive this information and would not have been able to act upon it. However, efficient ToM reasoning—known to be error-prone (e.g., Keysar et al., 2000, 2003)—was achieved.

Discussion

In our proof of concept model, we extended the AToM computational model to iteratively retrieve memories, generate ToM inferences, and refine its representation in order to resolve inconsistencies. In our proof of concept example, AToM successfully generated the expected ToM inferences, and when constrained to be less effortful produced shallow albeit incomplete inferences —consistent with the phenomena found in Apperly and Butterfill (2009).

In our model, the level of effort used corresponds to AToM's depth limit which would be controlled externally as a part of a broader cognitive architecture based on attention and available resources. Here we have demonstrated performance at the shallowest level (depth 1) and with significant effort (depth 5), but appropriate settings are likely situation-specific and the upper limit remains an empirical question.

Aside from allowing efficient and effortful reasoning in a common framework, our model makes testable predictions about the development of theory of mind. In AToM, ToM performance is dependent on available memories. As such, we predict that childhood (and adult) ToM failures will not be uniform and, in fact, that young children may be able to demonstrate higher-order theory of mind reasoning given a situation that closely aligns to their experience. Similarly, both children and adults should make slower and less accurate predictions when the situation or model of the person under consideration have structural differences from their experience. That is, when reasoning about the mental state of a person in a situation that is structurally very different from anything we have experienced or observed, several rounds of re-retrieval and re-representations will be necessary in order to make accurate predictions. For example, we predict that it is harder for a high school student to predict what their teacher is thinking when handing back a paper (because they have never taught) than what an employee is thinking when receiving their annual review.

Related Work

To the best of our knowledge, no other computational cognitive models of ToM reasoning have attempted to model the two-systems (Apperly & Butterfill, 2009) account of ToM. However, it is possible that some could be extended to account for this phenomenon.

For example, Hiatt and Trafton (2010) model ToM as a two-step process that first generates several hypotheses and then uses inhibition to select the appropriate choice. A possible extension that would allow them to model the distinction between efficient and effortful ToM might be a change to the inhibition function: perhaps initial efficient judgements are less inhibited than later effortful ones. Much like our model, this would suggest that a single process can account for both types of ToM reasoning. Unlike our model, however, this would imply that the effort falls on inhibition of possible inferences, rather than on their generation.

Similarly, the Bayesian Theory of Mind (Baker, Saxe, & Tenenbaum, 2011), which models ToM reasoning as Bayesian inference over a joint distribution of possible beliefs/desires integrated with a prior distribution of mental states, could be adapted to model the two-systems account by modifying the prior or abstracting the system's representations. The distinction, then, would come from the types of assumptions made about people's mental states (i.e., via the prior) or from the generality of the hypothesis space, and not from the process itself. It is important to note, however, that as a computational level model (cf. algorithmic level; Marr, 1982), the Bayesian Theory of Mind is intended to model behavior; differences in the model's processes may not be indicative of differences in the processes of human reasoning.

Others have modeled aspects of the two-systems theory. Nakos, Rabkina, Hill, and Forbus (2020) model a study by Epley, Keysar, Van Boven, and Gilovich (2004), which showed that children have an initial egocentric bias in a perspective taking task. Nakos et al. modeled the perspective taking as reference resolution (the process of identifying which entities a speaker is referring to) via analogy, and assumed that a ToM rule would be available to drive corrections in representations. That is, their model initially retrieved a referent without taking the interlocuter's knowledge into account, then corrected its representation upon identifying an error.

This is similar to our approach, both in the use of analogy and in correction upon finding an error (in our case, a contradiction between an inference and the real world). However, whereas Nakos et al. used representations of verbal referents (e.g., "the big truck") to determine which object was being referenced, we use observations of a person's actions to determine their goals, desires, and beliefs. Furthermore, we retrieve against a library of memories and make inferences based on those retrievals. Nakos et al., on other hand, retrieve against a library representing real-world objects and do not make additional references. Despite the differences in the two models, the corrective behavior that they are modeling is similar. Both also must account for another individual's mental states, if at different levels. As suggested by Nakos et al.'s use of ToM rules, it is likely that the processes modeled by the two models are related.

Another computational approach similar to the model proposed here is the Refinement via Analogy for Goal Recognition (RAGeR; Rabkina, Kantharaju, Wilson, Roberts, & Hiatt, 2022) algorithm. RAGeR is not a cognitive model, but it is also based on the Analogical Theory of Mind (Rabkina et al., 2017) and also uses re-retrieval to update an observation in order to make predictions about an observed agent's goals. Re-representations in RAGeR, however, are effectively walking up a hierarchical task network (Erol, 1995). It iteratively recognizes tasks in the initial observation and replaces their components (subtasks and actions) with the recognized task. On the other hand, the model presented in this work re-represents to improve its representation of the current observation and re-retrieves to improve its inferences about a compatriot's mental states. It is a computational cognitive model that makes predictions about the processes involved in people's ToM reasoning.

Conclusion

While typical adults are capable of impressive theory of mind reasoning, they fail to reliably apply said reasoning in everyday situations. These and other findings led Apperly and Butterfill (2009) to propose a two-system model of theory of mind. One system, efficient but inflexible, arises in early childhood and does not require explicitly represented beliefs. The other, which emerges with maturation, allows deep theory of mind reasoning but requires considerable cognitive effort.

In this paper, we have instead argued that efficient and effortful theory of mind could be the result of a single iterative process, with effort corresponding to re-representation of the situation under consideration and re-retrieval of relevant analogical comparisons to generate ToM inferences. We present an extension to the computational implementation of the Analogical Theory of Mind model, and demonstrate how AToM is capable of producing the intended behavior (Rabkina et al., 2017).

In future work, we plan to simulate additional behavioral experiments and investigate how autonomous agents can use such reasoning to interact in real-time environments (e.g., Rabkina et al., 2020). Finally, to date, AToM has relied on manually constructed episodic memories or domain-specific training data. We are also interested in examining how other models of long term memory, including large generative neural models, may be used to simulate episodic memory. As an example, Mostafazadeh et al. (2020) collected a large corpus of semi-structured natural language causal explanations which they used to train neural models for causal prediction. Such models may be able to generate plausible beliefs and causal chains as a stand-in for real-life experiences.

References

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological review*, *116*(4), 953.

Apperly, I. A., Riggs, K. J., Simpson, A., Chiavarino, C., & Samson, D. (2006). Is belief reasoning automatic? *Psychological Science*, *17*(10), 841–844.

Bach, T. (2011). Structure-mapping: Directions from simulation to theory. *Philosophical Psychology*, *24*(1), 23–51.

Bach, T. (2014). A unified account of general learning mechanisms and theory-of-mind development. *Mind & Language*, *29*(3), 351–381.

Baker, C., Saxe, R., & Tenenbaum, J. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 33).

de Villiers, J. G. (2021). The role (s) of language in theory of mind. In *The neural basis of mentalizing* (pp. 423–448). Springer.

Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of personality and social psychology*, *87*(3), 327.

Erol, K. (1995). *Hierarchical task network planning: formalization, analysis, and implementation.* Unpublished doctoral dissertation, University of Maryland, College Park.

Forbus, K. D., Gentner, D., & Law, K. (1995). Mac/fac: A model of similarity-based retrieval. *Cognitive science*, *19*(2), 141–205.

Forbus, K. D., & Hinrich, T. (2017). Analogy and relational representations in the companion cognitive architecture. *AI Magazine*, *38*(4), 34–42.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive science*, *7*(2), 155–170.

Gentner, D., & Medina, J. (1998). Similarity and the development of rules. *Cognition*, *65*(2-3), 263–297.

Goldman, A. I., et al. (2006). *Simulating minds: The philosophy, psychology, and neuroscience of mindreading*. Oxford University Press on Demand.

Goldman, A. I., & Sebanz, N. (2005). Simulation, mirroring, and a different argument from error. *Trends in cognitive sciences*, *9*(7), 320.

Gopnik, A., & Wellman, H. M. (1994). The theory theory. In *An earlier version of this chapter was presented at the society for research in child development meeting, 1991.*

Hale, C. M., & Tager-Flusberg, H. (2003). The influence of language on theory of mind: A training study. *Developmental science*, *6*(3), 346–359.

Hiatt, L. M., & Trafton, J. G. (2010). A cognitive model of theory of mind. In *Proceedings of the 10th international conference on cognitive modeling* (pp. 91–96).

Hoyos, C., Horton, W. S., Simms, N. K., & Gentner, D. (2020). Analogical comparison promotes theory-of-mind development. *Cognitive Science*, *44*(9), e12891.

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, *11*(1), 32–38.

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition*, *89*(1), 25–41.

Kovács, Á. M., Téglás, E., & Csibra, G. (2021). Can infants adopt underspecified contents into attributed beliefs? representational prerequisites of theory of mind. *Cognition*, 104640.

Lenat, D. B., & Guha, R. V. (1991). The evolution of cycl, the cyc representation language. *ACM SIGART Bulletin*, *2*(3), 84–87.

Lin, S., Keysar, B., & Epley, N. (2010). Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology*, *46*(3), 551–556.

Markman, A. B., & Gentner, D. (1996). Commonalities and differences in similarity comparisons. *Memory & cognition*, *24*(2), 235–249.

Marr, D. (1982). Vision.

McLure, M., Friedman, S., & Forbus, K. (2015). Extending analogical generalization with near-misses. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 29).

Meltzoff, A. N. (2007). 'like me': a foundation for social cognition. *Developmental science*, *10*(1), 126–134.

Mostafazadeh, N., Kalyanpur, A., Moon, L., Buchanan, D., Berkowitz, L., Biran, O., & Chu-Carroll, J. (2020, November). GLUCOSE: GeneraLized and COntextualized story explanations. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 4569–4586). Online: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2020.emnlp-main.370` doi: 10.18653/v1/2020.emnlp-main.370

Nakos, C., Rabkina, I., Hill, S., & Forbus, K. D. (2020). Corrective processes in modeling reference resolution. In *Cogsci*.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *science*, *308*(5719), 255–258.

Rabkina, I., Kantharaju, P., Roberts, M., Wilson, J., Forbus, K., & Hiatt, L. M. (2020). Recognizing the goals of uninspectable agents. *Advances in Cognitive Systems*.

Rabkina, I., Kantharaju, P., Wilson, J. R., Roberts, M., & Hiatt, L. M. (2022). Evaluation of goal recognition systems on unreliable data and uninspectable agents. *Frontiers in Artificial Intelligence*, 211.

Rabkina, I., McFate, C., Forbus, K. D., & Hoyos, C. (2017). Towards a computational analogical theory of mind. In *Proceedings of the 39th annual meeting of the cognitive science society.*

Rabkina, I., McFate, C. J., & Forbus, K. D. (2018). Bootstrapping from language in the analogical theory of mind model. In *Cogsci*.

Saxe, R. (2005). Against simulation: the argument from error. *Trends in cognitive sciences*, *9*(4), 174–179.

Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child development*, *75*(2), 523–541.