Anticipatory Thinking in Multi-Agent Environments: The Role of Theory of Mind

Irina Rabkina, Constantine Nakos & Kenneth D. Forbus

Qualitative Reasoning Group, Northwestern University

{irabkina, cnakos}@u.northwestern.edu; forbus@northwestern.edu

Abstract

Multi-agent environments pose unique challenges for the agents that interact with them. The complex behaviors of other agents can lead to novel modes of failure. Anticipating and mitigating failures due to other agents requires reasoning about their goals and beliefs, called theory of mind (ToM) reasoning. In this paper, we outline types of multi-agent scenarios that require ToM reasoning and discuss the complexity of mental models required to address them.

Introduction

Anticipatory thinking allows agents to manage risks posed by their environments (Amos-Binks & Dannenhauer, 2019). More complex environments pose a greater challenge for anticipatory thinking, since they make it more difficult to identify and mitigate potential risks. In particular, the presence of other agents can lead to unforeseen modes of failure. Agents may be capable of complex autonomous behavior, may pursue a variety of helpful or harmful objectives, and are likely to be motivated by uninspectable internal states. To adequately deal with the risks posed by other agents, an agent should have the ability to infer and reason about their beliefs, goals, and preferences.

In humans, reasoning about others' internal states is referred to as theory of mind (ToM; Premack & Woodruff, 1978). It is an integral part of social interaction ranging from everyday communication to complex team dynamics. ToM allows humans to anticipate others' reactions and behaviors, leading them to modify their own behaviors and plan for potential challenges. In short, humans use ToM for anticipatory thinking in social contexts.

Artificial agents in multi-agent environments would benefit from this capability, as well. For one agent to cooperate, compete, or coexist with another in a principled way, it should take ToM reasoning into account. In this paper, we outline the potential failure modes introduced by the presence of other agents and discuss how ToM can help an agent properly anticipate them. We also address the related question of how thorough a ToM model needs to be in order to sufficiently cope with these challenges. We end by proposing future directions for ToM reasoning in its application to anticipatory thinking.

Previous Work

Because the ability to anticipate the mental states of others affects most aspects of human interaction, ToM has been well-studied by psychologists (Wellman, 1992). Although several areas of prior research intersect with ToM, explicit ToM reasoning is a relatively new area for artificial intelligence research[1]. For example, Belief-Desire-Intention (BDI) frameworks (Bratman, 1987; Rao & Georgeff, 1991) provide a rich set of representations for reasoning about agents' internal states. However, BDI has largely been applied as a self-model, rather than being used as a formalism for reasoning about other agents. Other research has dealt with specific aspects of ToM such as those involved in collaborative planning and building up common ground (e.g., Allen et al., 1995; Rich & Sidner, 1998; Grosz & Kraus, 1999). These approaches account for task-related mental states of other agents, but do not attempt full ToM reasoning. Some recent work has also focused on task-specific implementations of ToM (e.g., Rabinowitz et al., 2018). The majority of recent work on ToM, however, has focused on modeling human ToM reasoning more generally.

The Bayesian Theory of Mind (BToM; Baker, Saxe & Tenenbaum, 2011) models ToM reasoning as inference over a partially observable Markov decision process (POMDP) with a stochastic policy. Given an agent's be-

[1] See Albrecht and Stone (2018) for a review of other approaches to agent modeling (e.g., plan recognition, type-based reasoning, etc.).

havior, BToM generates hypotheses about its beliefs and desires. The POMDP can be defined over any combination of action and state spaces, making BToM a domain-general model of ToM. Versions of BToM have been used to model children's ToM reasoning (Goodman et al., 2006) as well as adults' plan and intent recognition (Baker & Tenenbaum, 2014; Shum et al., 2019).

Other models of human ToM have been positioned within cognitive architectures. Hiatt and Trafton (2010) and Arslan, Taatgen, and Verbrugge (2013) have proposed models of children's ToM development in ACT-R. Although both models have been extended to encompass second-order ToM reasoning (Hiatt & Trafton, 2015; Arslan, Taatgen & Verbrugge, 2017), neither model has been applied to tasks outside of the cognitive modeling context. In other work, Hiatt, Harrison, and Trafton (2011) used a robot's ACT-R based self-model to reason about the decision making of its human teammates and found that humans preferred teaming with robots who performed such reasoning.

The Analogical Theory of Mind (AToM; Rabkina et al., 2017), built within the Companion cognitive architecture (Forbus & Hinrichs, 2017), models children's ToM reasoning and development. It treats ToM as an analogical process, with most reasoning occurring from inferences based on analogical comparison. Because these inferences depend on the cases in the Companion's memory, AToM is domain general. As with BToM, AToM has been used to model intent recognition and action prediction (Rabkina & Forbus, 2019).

Although BToM and AToM have demonstrated some success outside of the specific modeling contexts in which they were developed, neither has been applied to situations in which an agent must recognize and mitigate potential failures (i.e., in which anticipatory thinking is required). On the other hand, Hiatt et al.'s (2011) findings point to the promise of ToM in practical applications but could benefit from a more complete model of ToM. More work is needed to develop a complete computational model of ToM and situate it in a framework where it can be used to guide an agent's reasoning.

Anticipation in Multi-Agent Environments

To successfully navigate a multi-agent environment, an agent should be able to anticipate complications caused by the presence of others. We identify three broad types of multi-agent interaction: competition, cooperation, and incidental interaction. For each, we describe potential failure modes and the role of ToM reasoning in identifying and mitigating them. We use the open-world game Minecraft as a running example.

Competition

In a competitive multi-agent scenario, agents may have competing goals, not all of which are known a priori. To properly cope with competition, an agent should be able to identify its competitors, their goals, and the actions they intend to take to fulfill them.

To identify the role of ToM reasoning in competitive scenarios, it is useful to distinguish between two types of competition: 1) resource scarcity and 2) intentional interference. *Resource scarcity* occurs when two or more agents engage in a zero-sum competition for the same limited resource. Consider a Minecraft agent whose goal is to arm itself with a diamond sword, the strongest weapon available, in case a battle breaks out. Other players may have the same goal. Because diamonds are a scarce resource on most Minecraft maps, a competitor mining the diamonds first would hinder a player's ability to reach its objective. Upon recognizing competition, the agent should evaluate its options: it may need to choose another objective or find a way to deal with the threat. Successfully anticipating competing goals requires consideration of other agents' internal states in addition to their observed actions.

Intentional interference occurs when a competitor actively attempts to prevent an agent from achieving its goals. This may take the form of *deception*, affecting the agent's knowledge about the world, or *sabotage*, affecting the agent's ability to act on its knowledge. In both cases, the actions of the competitor impair the agent's ability to achieve its goals as planned. But where sabotage only requires reasoning about the competitor's intentions, deception involves more sophisticated ToM reasoning.

For example, when an agent is attempting to craft a diamond sword, a competitor may sabotage its attempt by killing it. Upon respawning, the agent would no longer have the tools necessary to mine diamonds, a setback to achieving its goal. As in the case of resource scarcity, ToM reasoning is necessary to recognize hostile intention and plan accordingly.

On the other hand, identifying deception requires an additional level of ToM reasoning. Deception exploits a victim's mental states, causing them to be inconsistent with the real world. Identifying deception therefore requires not only recognizing the intent to deceive, but also the nature of the deception. This is an example of second order ToM reasoning, where the agent must accommodate the competitor's ToM capabilities when making decisions.

Often, deception occurs through communication. For example, a competitor may send a message telling an agent that there are diamonds in a location where there is actually lava, playing on the agent's credulity. However, deception by action is also possible. Consider a Minecraft player helping an agent build a fortress. This may be a well-intentioned attempt to help the agent survive, a scheme to

gain the agent's trust for later betrayal, or cover for an immediately harmful action, such as laying a trap. In the latter two cases, recognizing that deception is occurring is crucial for the agent's survival. Moreover, identifying the goal of the deception allows the agent to form appropriate mitigation strategies. For example, allowing a competitor to keep building the fortress may be advantageous in the short term if the competitor is planning a later betrayal, but may prove deadly if the competitor is laying a trap.

Cooperation

Even when an agent has teammates genuinely working toward the same goal, the presence of other agents can lead to new forms of failure. Unlike competitive failure modes, cooperative failure modes are not intentionally harmful to the agent, but rather are due to a misunderstanding or other inconsistency between teammates' internal states and the real world.

Consider a Minecraft scenario in which players are trying to maximize the harvest from a farm. Crops are worth different food points and can be combined into recipes, usually with increased value. Maximizing the farm's output requires growing the optimal combination of crops, given seed availability. Due to the types of work involved, this is a natural opportunity for cooperation.

This domain provides examples of three key types of cooperative failure which can be mitigated by ToM reasoning. The first two involve correcting a teammate's beliefs, while the third involves repairing a teammate's plan. We describe each in turn.

First, a teammate may not have access to the information it needs to complete its task. For example, it may not know where the seeds are for a high-value crop, and so choose to pursue a low-value crop instead. In the absence of a fully-inspectable teammate, it is necessary to infer the teammate's lack of knowledge in order to address the problem, either by providing missing information or delegating around it.

Similarly, a teammate may have incorrect or outdated information about the world which interferes with proper planning. Such errors are particularly difficult to recognize and properly address, as the degree of misconception can be arbitrarily large. In a simple case, however, the approach should be similar to dealing with missing information. If a teammate is acting on a mistaken belief (e.g., that seeds needed for a given recipe are available, when in reality those seeds have been used elsewhere), correcting the misconception should be sufficient for correcting the teammate's behavior. In a broader sense, recognizing not only what the teammate is doing, but also what beliefs could be driving the behavior, will lead to improved outcomes in cooperative scenarios.

Finally, a teammate's incorrect actions might be caused by a faulty plan, rather than incomplete or mistaken beliefs about the world. In this case, it is important to recognize an agent's plan well enough to anticipate potential failures. For example, in Minecraft, the recipe for cake requires milk, sugar, an egg, and wheat. To correct a teammate that is trying to bake a cake with only wheat, milk, and sugar, an agent must recognize the intended goal, infer the missing step, and correct the teammate's model. Note that this is a correction to procedural knowledge, rather than semantic, and may require alternate mitigation techniques.

Incidental Interaction

It is important to note that the presence of other agents in an environment can cause failures even if there is no explicit competition or collaboration involved. One clear example of this is changes to the environment. Simply by interacting with the environment, an agent may change its structure or resource availability.

In Minecraft, this may take the form of using a resource or creating a structure in an unexpected location. These changes may help or hinder an agent's goals. However, such effects are a byproduct of the other player's goals, undertaken without consideration of their effects on the agent. In order to navigate a dynamic multi-agent environment, an agent would benefit from considering the internal states of others and anticipating incidental, in addition to intentional, acts of cooperation or competition.

Levels of Complexity in ToM Modeling

We have presented three classes of problems that can arise in multi-agent scenarios and which can be mitigated with the help of ToM reasoning. However, not all of these problems require a complete model of other agents. We propose a spectrum of ToM models, ordered by complexity, that can be used in various multi-agent scenarios (Figure 1).

The broadest level consists of causal reasoning, where agents are treated as factors in the environment. No special care is taken to differentiate agents from other entities in the world, and agent behavior is predicted without reference to their internal states (i.e., ToM reasoning). This may be sufficient to capture simple agent behaviors but is not enough to handle complex situations like those described above.

For example, mobs in Minecraft are computer-controlled agents with simple behaviors that populate the game world. A zombie mob will always attack the nearest player, while a sheep will run from the player if attacked. Knowledge of these behaviors is sufficient for any and all reasoning about mobs, without resorting to more complicated models.

In the middle of the spectrum is generic or stereotyped reasoning. This involves reasoning about agents whose

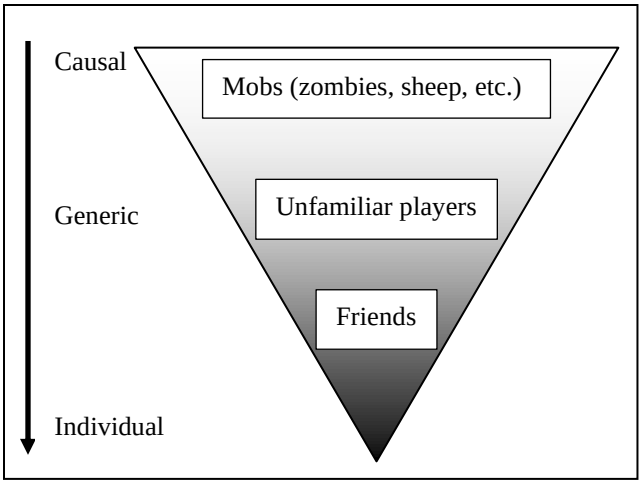

Figure 1. The amount of ToM reasoning necessary in multi-agent situations can be expressed as a spectrum, ranging from causal reasoning (e.g., about the behavior of a Minecraft mob) to generic or stereotyped reasoning (e.g., about the behavior of unknown players) to detailed models of individuals (e.g., close friends).

behavior is driven by unobservable internal states, but individual models of whom have not been built up. Generalizations may be formed about such agents based on trends in past experience and used for ToM reasoning.

In Minecraft, unfamiliar players (i.e., those who have not been encountered in the past) can be modeled at the generic level. Their behaviors can be predicted based on stereotypes: new players may need to be taught certain aspects of the game, while experts may be useful sources of knowledge. As a player becomes familiar, the agent may build up an individual model of that player and rely less on the stereotype.

The narrowest type of ToM reasoning is at the individual level and consists of models about specific players' goals, behaviors, and internal states. This level of reasoning has the greatest potential to accurately capture the factors determining an agent's behavior but comes at the cost of more complex modeling and the need to update information over time.

The formulation of individual models can also play a role in the formation of broader categories of ToM reasoning. For example, if all the new Minecraft players an agent has interacted with were unaware that an iron pickaxe is necessary for mining diamonds, this fact may be assimilated into the agent's generic model of new players. It will then assume that every new player it meets does not know about the iron pickaxe requirement and can plan accordingly.

Ultimately, the level of ToM needed for any given agent depends on the task it needs to perform. ToM reasoning encompasses a variety of potential approaches to reasoning about others. Choosing the right level of specificity entails a tradeoff between simpler, broader models that may not explain the full range of agent behavior and more detailed, specific models that require more complex representations and reasoning.

Conclusions & Future Directions

In this paper, we have outlined some of the scenarios in which agents would benefit from ToM reasoning for mitigating risks posed by other agents. Without the ability to model other agents' internal states, an agent cannot properly anticipate their behavior and adjust its own accordingly. This plays a role in competitive scenarios, where another agent is actively trying to interfere with the agent's goals, as well as in cooperative scenarios, where a teammate's behavior may be driven by flawed semantic or procedural knowledge. Further, ToM reasoning is helpful for anticipating environmental changes caused by other agents, even when those agents are not directly competing or cooperating.

For designers of cognitive systems, determining the complexity of ToM model necessary for a given task also poses a challenge. We have proposed a spectrum of ToM models that enable a designer to tailor an agent's ToM capabilities, given the needs of the environment and the data available. Implementing these models remains an area for further research.

Although the simplest level of ToM reasoning (i.e., causal views of agents) is often implicit in existing multi-agent models, the more complex levels require more sophisticated representation and reasoning techniques. Analogical Theory of Mind (Rabkina et al., 2017) and Bayesian Theory of Mind (Baker et al., 2011) have taken steps towards functional implementations of ToM reasoning but are not yet sufficiently developed for complex anticipatory thinking. Developing these and other models to handle the full spectrum of ToM will lead to agents that are better capable of navigating multi-agent environments. Moreover, designing agents from the ground up with ToM capabilities will lead to a better understanding of both ToM reasoning and anticipatory thinking in multi-agent contexts.

References

Albrecht, S. V., Stone, P. 2018. Autonomous Agents Modelling Other Agents: A Comprehensive Survey and Open Problems. Artificial Intelligence. 258:66-95.

Allen, J. F.; Schubert, L. K.; Ferguson, G.; Heeman, P.; Hwang, C. H.; Kato, T., ... and Traum, D. R. 1995. The TRAINS project: A case study in building a conversational planning agent. *Journal of Experimental & Theoretical Artificial Intelligence*, 7(1), 7-48.

Amos-Binks, A., Dannenhauer, D. 2019. Anticipatory Thinking: A Metacognitive Capability. arXiv preprint. arXiv: 1906.12249

Arslan, B.; Taatgen, N.; and Verbrugge, R. 2013. Modeling Developmental Transitions in Reasoning About False Beliefs of Others. In Proceedings of the 12th International Conference on Cognitive Modeling, Ottawa: Carleton University.

Arslan, B.; Taatgen, N. A.; and Verbrugge, R. 2017. Five-year-olds' Systematic Errors in Second-Order False Belief Tasks are due to First-Order Theory of Mind Strategy Selection: a Computational Modeling Study. *Frontiers in Psychology*, 8:275. https://doi.org/10.3389/fpsyg.2017.00275

Baker, C.; Saxe, R.; and Tenenbaum, J. 2011. Bayesian Theory of Mind: Modeling Joint Belief-Desire Attribution. In *Proceedings of the Annual Meeting of the Cognitive Science Society*. Boston, MA: Cognitive Science Society.

Baker, C. L., Tenenbaum, J. B. 2014. Modeling Human Plan Recognition Using Bayesian Theory of Mind. *Plan, Activity, and intent Recognition: Theory and Practice* edited by Sukthankar, G., Geib, C., Bui, H. H., Pynadath, D., & Goldman, R. P. 177-204. Newnes.

Bratman, Michael. 1987. *Intention, Plans, and Practical Reason*. Vol. 10. Cambridge, MA: Harvard University Press.

Forbus, K.D. & Hinrichs, T. 2017. Analogy and Relational Representations in the Companion Cognitive Architecture. *AI Magazine*.

Goodman, N. D.; Baker, C. L.; Bonawitz, E. B.; Mansinghka, V. K.; Gopnik, A.; Wellman, H.; ... and Tenenbaum, J. B. 2006. Intuitive Theories of Mind: A Rational Approach to False Belief. In *Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society*. Vancouver, Canada: Cognitive Science Society.

Grosz, B. J., & Kraus, S. 1999. The evolution of SharedPlans. In *Foundations of Rational Agency* (pp. 227-262). Springer, Dordrecht.

Hiatt, L. M., Trafton, J. G. 2010, August. A Cognitive Model of Theory of Mind. In *Proceedings of the 10th International Conference on Cognitive Modeling*. Philadelphia, PA: Drexel University.

Hiatt, L. M.; Harrison, A. M.; and Trafton, J. G. 2011. Accommodating Human Variability in Human-Robot Teams Through Theory of Mind. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*. Barcelona, Spain: AAAI Press.

Hiatt, L. M., Trafton, J. G. 2015. Understanding Second-Order Theory of Mind. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*: ACM.

Premack, D., Woodruff, G. 1978. Does the Chimpanzee Have a Theory of Mind? *Behavior & Brain Sciences*. 4:515–526. doi: 10.1017/S0140525X00076512

Rabinowitz, N.; Perbet, F.; Song, F.; Zhang, C.; Eslami, S. A.; and Botvinick, M. 2018. Machine Theory of Mind. In *International Conference on Machine Learning*. Jinan, China: ACM.

Rabkina, I.; McFate, C. J.; Forbus, K. D.; and Hoyos, C. 2017. Towards a Computational Analogical Theory of Mind. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*. London, England: Cognitive Science Society.

Rabkina, I., Forbus, K.D. 2019. Analogical Reasoning for Intent Recognition and Action Prediction in Multi-Agent Systems. In *Proceedings of the Seventh Annual Conference on Advances in Cognitive Systems*. Cambridge, MA: Cognitive Systems Foundation.

Rao, A. S., Georgeff, M. P. 1991. Modeling Rational Agents within a BDI-Architecture. *KR*, 91, 473-484.

Rich, C., Sidner, C. L. 1998. COLLAGEN: A collaboration manager for software interface agents. In *Computational Models of Mixed-Initiative Interaction* (pp. 149-184). Springer, Dordrecht.

Shum, M.; Kleiman-Weiner, M.; Littman, M. L.; and Tenenbaum, J. B. 2019. Theory of Minds: Understanding Behavior in Groups Through Inverse Planning. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Honolulu, HI: AAAI Press.

Wellman, H. M. 1992. *The Child's Theory of Mind*. The MIT Press.