Operationalizing Theories of Theory of Mind: A Survey

Nikolos Gurney,[1] **Stacy Marsella,**[2,3]
David V. Pynadath,[1] **Volkan Ustun**[1]
[1]Institute for Creative Technologies, University of Southern California
[2]Northeastern University, [3]University of Glasgow
gurney@ict.usc.edu, marsella@northeastern.edu, pyndath@ict.usc.edu, ustun@ict.usc.edu

Abstract

Human social interaction hinges on the ability to interpret and predict the actions of others. The most valuable explanatory variable of these actions, more important than environmental or social factors, is the one that we do not have direct access to: the mind. This lack of access leaves us to impute the mental states—beliefs, desires, emotions, intentions, etc.—of others before we can explain their behaviors. Studying our ability to do so, our Theory of Mind, has long been the province of psychologists and philosophers. Computational scientists are joining this research space, however, as they strive to imbue artificial intelligences with human-like characteristics. We provide a high-level review of Theory of Mind research across several domains, with the goal of mapping between theory and recursive agent models. We illustrate this mapping using a specific recursive agent architecture, PsychSim, and discuss how it addresses many of the open issues in Theory of Mind research by enforcing a set of minimal requirements.

Our goal is to communicate to you a set of propositional attitudes (cognitive states such as beliefs or desires) related to developing an artificial intelligence helper agent (AI helper). Philosophers and psychologists have long posited that such communication requires knowing the mental states of both the listener *and* oneself. As convenient as it would be, we do not have direct access to your mental states nor our own. This means that if our goal is to communicate our position effectively (it is) we must impute these states. The capacity to do so is commonly known as *Theory of Mind* (ToM Premack and Woodruff 1978).[1]

Imagine that we were slightly more intrepid authors who thought it would be easier to enlist an AI helper to author our manuscript. The fundamental communication problem, that of understanding the mental states of its readers and the author team, would persist for the AI helper. Moreover, we believe that it is reasonable to assume that access to its own mental states would make the the AI helper better at its task (because then it could consider how its beliefs about you or us might factor into its work). When the helper explains a technical term, for example, it may want to assess the accuracy of its belief of your degree of belief that it has accurately depicted the term. In other words, the AI helper needs a ToM.

We are academics and as such it behooves us to share our opinions. It is only natural then that we would not just turn the AI helper loose and accept whatever it produced. We undoubtedly would find ways to insert ourselves into its process, to team with it as a sort of human-AI collective intelligence (Gupta and Woolley Forthcoming), in hopes of producing a high quality paper that reflects our positions. The success of our human-AI team largely hinges on trust (Bonaccio and Dalal 2006; Wang, Pynadath, and Hill 2016; Wang et al. 2018). For the human members of our team, trust is grounded in our assessment of the AI's abilities, critically, its level of intelligence (Glikson and Woolley 2020). Again, we believe that a robust theory of mind, in this instance for its human teammates, will facilitate the AI's success. Herein we review a subset of the vast literature on human social cognition, specifically focusing on ToM, and discuss implementation of myriad theories and models in such an AI helper using PsychSim, a recursive agent architecture. Critically, the PsychSim implementation enforces a set of minimal requirements that we believe reveals the strengths and weaknesses of different ToM theories.

PsychSim

This basic description of a PsychSim agent and its machinery will serve as a framework for comparing ToM theories. PsychSim is a social simulation platform with the capacity to implement psychologically valid theories of human behavior (Pynadath and Marsella 2005). It uses a recursive architecture, meaning that it applies the same rules repeatedly to generate outputs (Gmytrasiewicz and Durfee 1995). Each agent in a PsychSim simulation possesses a fully specified decision-theoretic model, i.e. model of choices based on a utility framework, of itself and the other agents in its environment. Most importantly, the platform readily facilitates modeling beliefs, including those related to ToM (Marsella, Pynadath, and Read 2004). A PsychSim-based AI helper thus has a model of itself, for each of its teammates, and potentially for the team holistically which it can use to sim-

[1] This name is somewhat fraught due to its implication that the development of an explicit theory is part of the underlying cognitive process. Adding to the confusion, cognitive and computational scientists often refer to how researchers have a theory of mind about how the mind works (Newell 1994). We adopt it, nevertheless, due to its universal recognition.

ulate different scenarios and base recommendations on just like our hypothetical paper mill agent.

The minimal requirements for creating an AI helper with ToM using a recursive agent architecture, such as PsychSim, are:

(i) framework for inferring beliefs, from observations, about others

(ii) means of translating these beliefs into predictions about behaviors

(iii) way of handling higher order reasoning (you believe that he believes that she believes...)

Partially observable Markov decision processes (POMDPs) are the backbone of PsychSim and fill each of these requirements. The POMDP framework posits that an agent assumes system variables, such as other agents, follow stochastic processes (MDPs) which it cannot fully observe. MDPs are characterized by a set of states, actions, probabilities of given actions for each state, and rewards associated with arriving in particular states. The task of an agent in an environment modeled by an MDP is simply observing the current state and potential rewards of each available action then selecting the best option given the data. States are not directly observable in the partially observable generalization of MDPs, so the agent must gather outcome information. The agent maintains models of itself and the other agents that it updates based on this information to overcome the observability obstacle. Similar to the use of recursive models in *interactive* POMDPs (I-POMDPs) (Gmytrasiewicz and Doshi 2005), these models take the form of probability distributions for its observations given the state of the system and models of the stochastic processes followed by the agents in the system. The POMDPs of an agent generally rely on utility-based functions to model behavior.

Theories about Theory of Mind

Humans gather, process, and create information about the actions of other agents in their environment—a set of information processing behaviors collectively called social cognition (Fiske and Taylor 2021). Like most domains of human cognition, that is where the consensus on social cognition ends and debate begins. In the case of ToM, which is widely considered a subdomain of social cognition, there are at least three theoretical explanations of human ToM: theory theory, simulation theory, and more recently, arguments for social cognition without ToM. Recursive agent models, including PsychSim, share many features with all of these explanations.

A set of minimal requirements for ToM reasoning will help illustrate the strengths and weaknesses of each theoretical position as well as clarify the approaches that recursive agent models take. It is our argument that the three requirements of developing a recursive agent architecture can also serve as coarse but critical requirements for ToM theories.

Theory-Theory

Modern inquiry into how humans think about and represent the mental states of others is frequently traced back to Heider and Simmel's (Heider and Simmel 1944) famous geometric shapes experiment. The basic paradigm involves participants watching interacting geometric shapes and reporting what they saw. In a classic scene, three shapes move around the screen. One shape quickly moves towards another shape while a third shape later moves in between them. This ground breaking experiment revealed that participants, almost universally, ascribed agency to the shapes. For example, participants said that the above scenario depicted an aggressor, victim, and third party who intervened to stop the aggression.

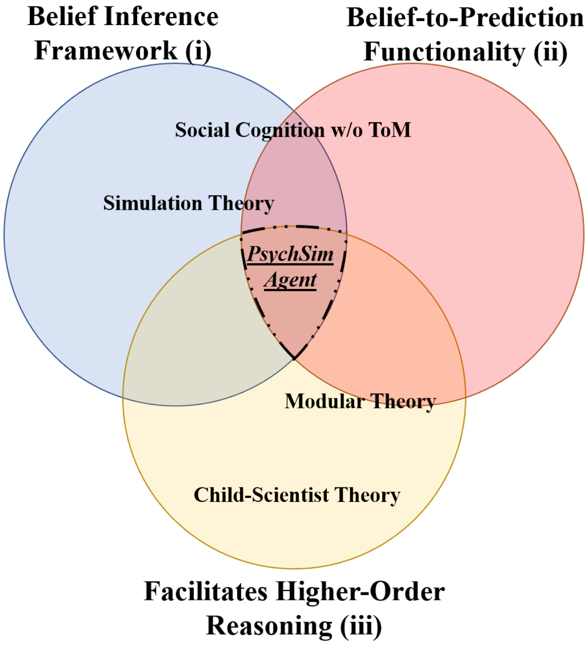
Belief Inference Framework (i)
Belief-to-Prediction Functionality (ii)
Social Cognition w/o ToM
Simulation Theory
PsychSim Agent
Modular Theory
Child-Scientist Theory
Facilitates Higher-Order Reasoning (iii)

Figure 1: The three minimal requirements of implementing ToM reasoning in a PsychSim agent and our subjective assessment of where major ToM theories fall in the space of these requirements. We argue that these requirements are also applicable to actual ToM reasoning and can serve as a means of evaluating theories and models.

Although Heider faltered in arguing that such abilities result from direct access to our own internal mental states (Heider 1958), his work with Simmel precipitated the first recognizable version of theory-theory from the philosopher Wilfred Sellars (1956). Sellars took the position that we develop, via experiences and enculturation, a naive theory to explain the mental states of our peers. In other words, we function as lay scientists developing and testing tacit theories about our social world based on input from the people around us, social observations, and even rudimentary experimentation. This position would eventually come to be known as theory-theory and finds most of its empirical support in adaptions of Wimmer and Perner's false belief task (Wimmer and Perner 1983).[2]

[2] The false belief task has proven extremely productive for the

The conceptual account of theory-theory maps relatively cleanly to a PsychSim agent. As will become clear, PsychSim offers a formalization of ToM requirements i and ii where theory-theorists do not. Using the lay scientist analogy, the PsychSim agent's research methods, or the ways in which it formalizes and answers hypotheses, are POMDPs. A PsychSim agent typically assumes that other agents in its environment are using POMDPs for their decision-making as well. Each time a PsychSim agent makes an observation, it updates it beliefs about how the agents and environment work. In the case of agents, it assumes that they are trying to maximize their utility. This renders the accuracy of its model, which determines its ability to predict how another agent will respond to a stimulus, contingent on its prior experiences. Importantly, any interpretation it makes of a person's behavior is thus circumscribed to be goal-driven via the maximization of some utility function. Even if the observed behavior is generated by an arbitrary stochastic process, the PsychSim agent will form posterior beliefs over a candidate set of POMDPs based on how well their corresponding behaviors would match the observed behavior.

Child-Scientists The child-scientist theory for ToM development is arguably the truest to Sellars' original vision and the most common among theory-theorist. Like many scientific perspectives, it started as an analogy to better understand how humans might develop the ability to represent others' minds. Its most ardent supporters, however, take it beyond the analogical insights and argue that philosophy of science's theory development processes are actual blueprints for how cognition transpires (Kuhn 1989; Gopnik 1996; Gopnik, Meltzoff, and Kuhl 1999; Perner and Lang 1999). Theories are viewed as systems that assign specific representations to inputs, similar to how the visual system assigns representations to input. These systems are not rigid, particularly in early development. Just like a young child refines her interpretation of visual input from a single representation for fuzzy, four-legged creatures (dog) to a multitude of representations (cat, horse, sheep, etc.), she refines her representations for the various processes that underpin the behavior of agents in her social world.

Proponents of the child-scientist view generally subscribe to one of two explanations for the brain's ToM mechanism: a general purpose (domain-general) or modular (domain-specific) learning mechanism. The child-scientist perspective adopts the former, i.e. there is a generalized psychological structure that supports learning across domains. This general purpose learning mechanism is supported by a suite of reflexes, or cognitive capabilities, that are present from birth (Gopnik and Meltzoff 1994). Together, the learning mechanism and reflexes exploit sensorimotor interactions with the environment to form theories, or models, of how the world works. These theories empower a child to not only understand her environment, but also solve increasingly complex prediction problems, including those that are classically social. The Bayesian flavor of all this is no fluke. Formalizations of domain-general mechanisms frequently take Bayesian forms (Gopnik et al. 2004; Lu et al. 2008; Frost et al. 2015).

Theory theorists can readily handle our requirement iii. A lay ToM scientist trying to workout higher-order beliefs simply needs to establish a hypothesis, test it, and update accordingly. The testing can even happen in situ by imagining various outcomes given a set of beliefs. Items i and ii present bigger challenges. The child-scientist approach is to assume that ToM is handled by a learning mechanism used in other types of learning, like the statistical learning mechanism proposed for vision and audition (Kirkham, Slemmer, and Johnson 2002). There is reasonable evidence that suggests domain-general mechanisms support ToM reasoning. For example, when a person is under cognitive load their performance in social reasoning tasks may decrease (McKinnon and Moscovitch 2007). This is, of course, only correlational and not causal.

The PsychSim agent uses maximum expected utility as its domain-independent theory of reasoning, one that also acts as a constraint on its theory of others. Within that constraint, it is free to choose an arbitrary set of (possibly recursive) POMDPs as its model of those others. Like a developing human, it is possible to specify a PsychSim agent that starts with a small set of simple (e.g., short horizon, few goals) POMDPs. These are analogous to the child's theories for how the world works. Experience allows it to continually expand that set when none of its current candidates satisfactorily explain its experiences in social interaction.

Modular Theory Cognitive modularity suggests that the brain has a fixed architecture. One perspective argues that this structure limits the flow and processing of information. The edges of modules, or regions, in the brain function as filters that can be uni- or bi-directional. That is, some of the information inside the module may not be available outside or vice versa (Fodor 1983). Although theorists conceptualize the modules as rigid units, they generally offer a bit of hand waving when pressed on the actual degree of modularity (Fodor 1983; Scholl and Leslie 1999). The alternative view of modularity is knowledge-centric, meaning that rather than the brain having distinct modules for processing information, they store it. This allows for flexible skill and belief systems to process across these core systems of object representation (Spelke and Kinzler 2007). Modular explanations for ToM generally adopt the former position and posit a distinct functional system in the brain which comes online during childhood development (Leslie 1994b,a; Scholl and Leslie 1999; Leslie, Friedman, and German 2004).

Alan Leslie's Theory of Mind Mechanism/Selection Processing is arguably the dominant modular explanation (Leslie 1987, 1994a, 2000; Scholl and Leslie 2001). Leslie and colleagues argue that ToM is innate and that there is a unique mechanism which yields representations solely for related reasoning. This mechanism, much like puberty, is genetically present (innate) and activated by environmental factors. They do not, however, argue that this mecha-

scientific community. Review of experimental methods is not in the scope of this paper, however it is worth noting that it and the empirical designs that followed in its footsteps likely have numerous flaws (Bloom and German 2000; Heyes 2014; Turner and Felisberti 2017; Quesque and Rossetti 2020).

nism is responsible for all ToM abilities—just that it has a specific innate basis and a function unique to ToM reasoning. The mechanism part of the theory stipulates that the ToM module automatically processes perceptual information about behaviors and computes what mental states may have produced them. Because this process is algorithmic and spontaneous, it is prone to errors, be they from biased learning or learned heuristics. This necessitates a supplemental, executive system that overrides the module's salient outputs when they are unwanted. In the case of a false belief task, the mechanism yields a true belief about the state of the world (the actual location of the hidden object).The selection processing overrides this belief to yield an accurate ToM for the target (who has a false belief about the object location).

Modular theorists attempt a more direct approach to meeting requirements ii and iii than their child-scientists counterparts, but stop short of pointing to an actual module in the brain that handles inference (i) or prediction (ii). The ToM module handles belief inference and the translation of these beliefs into predictions, both of which can be exported to other regions of the brain for various purposes. Unlike vision or audition researchers, however, modular theorists offer inconclusive evidence for where in the brain the module exists and about the algorithmic way in which perceptual information about behaviors is processed (Saxe and Wexler 2005).

Most recursive agent models are modular, in an information processing sense, by default. PsychSim encapsulate the models of different agents and at different recursive depths so that information cannot flow between the models except along edges in the corresponding influence diagram. This means that a unique model exists for processing information about each agent and these models output information that flows between them to account for their interactions. Furthermore, while both the agent's decision-making and its ToM models of others use the same decision-theoretic algorithms, there are typically computational shortcuts taken in the ToM models that are not used for the agent's own more thorough reasoning. For example, PsychSim ToM POMDPs are usually more abstract than the actual POMDPs used for behavior generation. This abstraction can be achieved by removing variables from the POMDP that have limited or no influence on the modeling agent's utility (Pynadath and Marsella 2007). Removing these variables, however, does introduce uncertainty. PsychSim usually handles this by implementing a softmax, instead of strict maximization function, which is more forgiving of errors that may result of the uncertainty. The resulting POMDP is smaller (and thus faster) than the original.

Simulation Theory

In its simplest form, simulation theory says that we understand and predict the mental states of others by trying to simulate them within ourselves (Gordon 1986). This yields a very straightforward means of achieving ToM: repurposing the psychological machinery used for our own cognition to gain insight into the cognition of others. Doing so implies a two part process: first, generate imaginary mental states that correspond to the target's mental states, and second, feed those mental states into the appropriate mechanism(s) to generate an output. The simulator is thus capable of producing feasible explanations of behavior whenever they make decisions in a manner roughly similar to the target (Goldman 2006). Simulation theorists argue that this yields a much more efficient ToM than any proposed by theory-theorists.

Theory-theory is constrained by high-level reasoning: the need to actively think about the cognitive state of another person to test and validate a naive theory about their every behavior is likely intractable for humans. Because of this, theory-theory is a poor candidate for explaining the seemingly automatic human capacity for detecting subtle social cues. If we had to stop to ponder on the expression of emotions, like a slight twitch in the face that is indicative of anger, we would frequently find ourselves in grave danger. Simulation theorist argue, however, that it is very capable of explaining low-level phenomena, like automatic detection of a fear response in another person. It is plausible that mirror neurons are the basis for this capability (Gallese and Goldman 1998), but the evidence for mirror neurons in humans is largely inferred rather than observed. Invasive procedures are necessary to observe mirror neurons, which is why the vast majority of studies reporting on them appear in the animal literature (Rizzolatti and Craighero 2004) save a single notable example in humans (Mukamel et al. 2010).

It is very plausible that ToM involves two (or more) distinct systems, possibly both theory-theory and simulation theory processes (Mitchell 2005; Perner and Kühberger 2005; Goldman 2006; Apperly and Butterfill 2009; Heyes and Frith 2014; Carruthers 2016; Low et al. 2016). Goldman (2006), for example, suggests that theorizing might play an important, even dominant, role in "high-level mind reading," which he defines as imaginative simulation that is conscious, actively controlled imputation of others' mental states. This supplements unconscious, simulation-based mind reading that handles simple mental states (e.g. detecting emotional states such as fear from facial expressions). Heyes and Frith (2014) adopt the terms implicit and explicit to describe neurocognitively inherited and enculturated skills respectively. The implicit mechanisms are present from birth and play a vital role in formulating accurate expectations about the behavior of agents. Further, they accept that the outputs of an implicit system may inform the explicit system by preprocessing observed behavior in a way that facilitates categorization. They make clear, however, that they do not believe this to be sufficient for the sort of complex ToM observed in mature humans.

Whereas theory-theory explanations are vague about the mechanism that handles inferences, requirement i, simulation theory is much more direct: the mechanisms that we use for our own behavior are repurposed for ToM. Predictions about behavior then become a simple task of running the belief outputs through the same system that determines our behavior. This account implies that we can, in a sense, re-enter the machinery we use for our own cognition. Before the discovery of mirror neurons this was pure speculation, and to a large extent it still is given the lack of human evidence related to mirror neurons.

Unlike a human, there is a clear way for a PsychSim agent to re-enter the machinery that it uses to understand its social world. PsychSim agents explicitly use their POMDP models of others to generate expectations of their behavior. In particular, they apply their own POMDP solution algorithms to the variables, graph structures, and parameters they (possibly incorrectly) ascribe to others' POMDPs. An agent will use this capability to simulate the outcomes of each of its alternate actions, evaluating those outcomes against their utility function, and choosing the behavior that maximizes that expectation. It will also use this simulation capability to evaluate alternate explanations of observed behaviors in terms of their likelihood, which in turn causes its update its belief in those explanations.

First-Person ToM and Introspection

Both theory-theory and simulation theory posit that first-person ToM plays an important role in our social capabilities (Gopnik 1993; Goldman 2006). There is division, however, over whether we have direct access to our internal states to make use of during first-person ToM reasoning. Gopnik (1993) argues that even though people may believe their first-person knowledge is derived from experience, it actually comes from the same theory of mind system that explains the behaviors of other agents. This is because, as she argues, we lack direct access to the psychological process underlying out own behavior. Goldman (2006), on the other hand, claims that we do have direct access to the psychological processes behind our behaviors. Moreover, it is his position that first-person theory of mind development precedes and is necessary for third-person abilities. An alternative simulation account is that ToM always has a "target" agent and the simulation is the same whether it is first or third person, implying that there is no direct access to internal mental states (Gordon 1995).

PsychSim is equipped to do all the above. First, it has two mechanisms for an agent to reason about its own behavior: via direct access to its "true" POMDP, or via a perceived POMDP. The latter may deviate from the former, just as its POMDP models of others can deviate from their true models. It is agnostic, however, about the timing of when a first-person ToM comes online. In theory, it could be inhibited or prioritized, but PsychSim implementations generally have everything happen simultaneously.

Social Cognition Without ToM

There are numerous, defensible explanations of how humans anticipate and respond to other agents that do not involve any of the classic ToM processes. Game Theory supplies many examples and one of the most recognizable: the tit-for-tat strategy (Axelrod and Hamilton 1981; Rabin 1993). An agent who uses this strategy in a social interaction first cooperates and then in every subsequent interaction simply replicates the behavior of the other agent(s). The tit-for-tat model of behavior easily explains complex human behavior, such as reciprocal altruism (Trivers 1971; Brosnan and De Waal 2002), without appealing to complex ToM processes. Economic theorists have posited rich social behaviors, including cooperation, emerge from learned and automated rules, i.e. social norm heuristics (Bowles and Gintis 2003; Chudek and Henrich 2011). Such heuristics may circumvent the need for deliberate self constraint—which would need to emerge from ToM reasoning—when an agent faces no risks from taking advantage of another agent (Gigerenzer and Gaissmaier 2011; Molnar and Loewenstein 2021).

Other behaviors are executed according to what psychologist call scripts (Bower, Black, and Turner 1979) and artificial intelligence researchers call frames (Minsky 2019). A script (frame) includes a set of elements, descriptions of what those elements can do, outcomes given combinations of the various elements, and is composed of scenes. Walking into a corner market, picking up a package of candy, and paying for it at the register can all happen without explicit ToM when a script is in place. You do not have to reason about the attendant wanting money from you in exchange for your snack—the script dictates all the necessary behaviors for the social interaction.

While POMDPs are the core part of PsychSim's base architecture, there are occasions where it has been convenient to bypass decision-theoretic reasoning altogether and directly encode an agent's policy. The form of this policy is a piecewise-linear decision tree. It can capture behaviors that are reminiscent of ToM, but do not require the complete machinery. Although these sort of policies are closer in spirit to frames, they are conceptually the same thing as a script and can be implemented in recursive agents when warranted (Pynadath and Marsella 2004).

Modeling Theory of Mind

A complete model of ToM reasoning will have a framework for inferring beliefs about others, a way to translate those beliefs into predictions, and the capacity to handle higher order reasoning. Researchers have formalized the proposed theoretical structures of ToM in a diversity of ways. Many models are grounded in paradigms with rich histories in computational research including reinforcement learning (RL), partially observable Markov decision processes (POMDPs), utility maximization, and Bayesian inference. Additionally, modelers take both-model based and model-free approaches (Friston et al. 2016; Gershman et al. 2016; Jara-Ettinger 2019). Model-based approaches are prospective, meaning that they assume a goal and active cognition, often referred to as system II thinking (Kahneman 2011). Model-free approaches, on the other hand, are retrospective and capable of capturing habitual cognition, i.e. Kahneman's system I thinking. Each modeling approach comes packaged with a suite of benefits and laundry list of short comings that effect its ability to capture the nuances of ToM reasoning. Moreover, there are costs and benefits when implementing the myriad models in an AI helper. The shortlist of exemplars included here is far from exhaustive, but we believe each merits consideration when designing AI helpers for human-machine teams.

Bayesian Inference

Bayesian approaches to ToM often conceptualize the observing agent as forming a hypothesis about the target's behavior. This hypothesis is evaluated given observable data and under the constraints of an underlying theory of behavior in a very theory-theory fashion (Gopnik et al. 2004; Baker, Tenenbaum, and Saxe 2006; Tenenbaum, Griffiths, and Kemp 2006; Baker, Saxe, and Tenenbaum 2009). This allows for ToM to be cast as an inverse planning and inference task. When a person observes another's actions, she answers the ToM problem (implicitly) by assuming that the other person made the decision based on data, that likely includes beliefs, and according to some model of how to act in the world which approximates rationality. Next she attempts to invert this model of how to act by applying Bayesian inference: integrate the likelihood of her observations with a prior over mental states. The output becomes her ToM for the other's behavior. In one computational example of this, a target agent's plans and inferences are formalized as POMDPs that capture propositional attitudes (e.g. desires and beliefs) via utility functions and probability distributions respectively (Baker, Tenenbaum, and Saxe 2006). The target is assumed to be approximately rational, i.e. the target is utility maximizing. Inverting the target's forward model using Bayesian inference yields the observer's ToM model of the target.

Bayesian models of ToM not only offer a handy way of operationalizing an opaque set of cognitive processes, they also facilitate other important capabilities for adaptive agents. For example, if an observer has reason to believe that a target is knowledgeable, then she can adapt the Bayesian process used for ToM reasoning to learn how to act or refine her beliefs (Shafto, Goodman, and Frank 2012). Or, if an observer is observing a group of targets, rather than attempting to compute a ToM for each target, she can adopt a model that represents the "average" member of the group and rely on it to make inferences and predictions about how the individuals within the group may act (Khalvati et al. 2019). Importantly, the same Bayesian machinery used to model basic social cognitive processes can be used in modeling affective cognition (D'Mello, Kappas, and Gratch 2018; Ong, Zaki, and Goodman 2019).

POMDPs are one of the most common means of capturing the prior-observation-update-posterior belief pipeline of reasoning depicted by Bayesian ToM theories. One of the earliest examples of a Bayesian theory of mind also used a POMDP-based architecture (Si, Marsella, and Pynadath 2010). And, as we have explictly noted, this is what PsychSim implements for its agent models.

Game Theory and Economics

Economists use game theory to model how economic agents think about and respond to the mental states of others (Camerer 2011). For a given model, the "game" is captured by a mathematical description of the strategies and associated payoffs available to each agent. Games often have multiple stages during which agents choose actions to execute from a limited set, can be competitive or cooperative, and range in length from a single shot to (theoretically) infinite number of rounds. Every agent is assumed to hold beliefs, which are captured as probability distributions, about the available actions, progress of the game, and even beliefs of other agents. Importantly, games are structured such that predictions about a player's behavior can be derived without any observations.

In game-theoretic approaches to modeling ToM, each agent generally has a policy over strategies that dictate how it will behave given a set of conditions—including inferences of other agents' policies and observed behaviors. This policy is subject to a state-dependent value that the agent is attempting to optimize for in a particular game setting. Each agent has a level of *sophistication* that describes the degree to which it considers the depth of other agents' models of it (Yoshida, Dolan, and Friston 2008; Camerer 2011).

Rousseau's stag hunt problem is a classic example of a social dilemma easily captured by game theory. Two hunters must independently decide whether to hunt a stag or hare. Hunting a stag successfully requires input form both hunters and results in each hunter garnering a greater reward (more meat). Hunting a hare can be accomplished without coordination, but also has a lower payoff. There is a risk to choosing to hunt a stag in that if the other hunter pursues hare, you will go hungry. This simplified game has two strategies (hunting a hare or stag) and the value function hinges on the amount of meat from each strategy. Each hunter has a model of the other hunter's likelihood of selecting stag that includes the other's beliefs about herself. The concept of sophistication captures how many recursions an agent considers in her model. In line with Herbert Simon's famous work on the bounds of human rationality, people generally do not go much beyond two-step logic (Camerer 2003; Simon 1997).

Game theoretic models can capture a number of interesting social behaviors, but they fall short of explaining the rich set of mental gymnastics that comprises social cognition. That does not mean that these models are out of place in recursive agent architectures. Situations arise when the decision-theoretic models are overly cumbersome and a simple heuristic, like tit-for-tat, is warranted. In practice, these are implemented as explicit policies (Pynadath and Marsella 2004).

Reinforcement Learning

The basic concept of model-based RL is to combine a world model and reward function to produce a policy. The world model is a learned, simplified model of an agent's environment and used to make predictions about future states of the world. Reward functions can take many forms, but frequently a cost minimization or benefit maximization function of the world model's accuracy is implemented. In essence, these are models of what an agent "ought" to do. If a food item tastes good (bad), it ought (not) to eat it and (or) be rewarded with the good (bad) flavor. The policy is the sequence of actions that an agent uses in pursuit of a goal and generated, or learned, from the repeated combination of the world model and reward function. If we assume that a mind functions via model-based RL, then predicting mental states from observable behavior can take the same form as *inverse* RL.

IRL involves an observer agent that tries to learn a target's utility function given repeated decision observations. In its simplest form, this requires a state space, an action space, and transition function which are modeled as a Markov decision process. Although powerful, this is computationally expensive—vast numbers of labeled training examples are required in order to infer a reward function from a policy (set of observed states and actions) and transition function, which is frequently a researcher degree of freedom (Lake et al. 2017). A human infant with an IRL-based ToM would need hundreds of thousands of labeled training examples a day (Jara-Ettinger 2019).

Discussion

Evaluating, especially comparing, theory about and modeling approaches to ToM reasoning is challenging. We believe that this challenge arises because theorists and modelers approach ToM without a unifying set of requirements for ToM reasoning. This leads to the value of their unique perspectives being lost to the variance in their interpretation of the problem. As we have illustrated, such a set of requirements is indispensable when comparing different approaches and perspectives. Theories that only focus on how humans handle higher order reasoning, for example, are hard to compare to those that are primarily concerned with a framework for inferring beliefs. Developing and deploying artificial agent systems, such as PsychSim, forces not only acknowledging the need to take a stance on the minimal requirements of ToM reasoning, but implementing and validating them as well.

Theory-theory approaches that do not specify a belief inference framework or its origin are incomplete. Without a framework, it is not possible to falsify a theory because simple adjustments to the mechanism that produces inputs, i.e. beliefs, that support ToM reasoning alter the theory's accuracy. This is particularly important for accounts that claim ToM is acquired rather than innate. ToM input-output data are scarce. This leaves researchers in the position of being able to select from a vast array of candidate functions the one that best fits the data and their theory. Validating their selection from the legions of alternatives is not possible, however, given the data paucity. Thus their claim becomes this is the acquired function because it fits the data and our model.

If ToM is innate, then the fundamental difference between the theory and simulation perspective is simply the inference mechanism. All that the theory-theorists are saying is that we do not know the mechanism, however it is not *that* mechanism. This again leaves them in a position to select whatever mechanism works with the data and their theory. Meanwhile, the simulation account of a repurposed mechanism leads to a paradox. If it is the same mechanism that we use for our own behavior, it must be one which can be reentered or that supports recursion. This is necessary for the third item in our list, higher order reasoning, but not necessarily for behavior. This suggests that the mechanism was not repurposed, but designed/evolved with the capability for ToM.

A theory without a belief translation mechanism is also hard to falsify. This is because, like not having a belief inference framework, all it takes is a convenient function to make your theory valid. And again, the lack of data makes it challenging, if not impossible, to validate whether a given functional form is correct. This makes it impossible to validate the entire pipeline. A model that lacks the first or second item and only specifies a way of handling higher order reasoning, like many perspectives in the theory-theory camp do, lacks the needed structure to test its validity.

Explanations for social cognition without the higher order reasoning that typifies ToM face the challenge of exploding complexity and generalization. Scripts and frames can account for social interactions and do not need higher order reasoning, but scripting every class of social interaction would quickly become intractable for humans as the number agents and levels of reasoning increase.

All of these challenges to existing modeling and theoretical approaches to ToM reasoning point to the need for a more holistic account. Recursive agent models, like PsychSim, force researchers into taking a position on each requirement. PsychSim assumes that ToM reasoning follows a POMDP framework. The same framework is repurposed for learning about the environment, making predictions about behaviors, and higher order reasoning. Game Theory and reinforcement learning offer framework alternatives, but we believe each has a fundamental flaw. If an agent implemented a pure reinforcement learning approach, like direct policy search, it would need a way to down select to a good set of candidate policies from the vast set of possible policies. This adds a new requirement for ToM reasoning and it is possibly intractable as each new variable for consideration increases the complexity of the policy selection task. A purely game-theoretic approach would mean specifying the details for each game that an agent may enter and knowing when to use each unique game, i.e. belief inference framework. Again, the complexity of a system based entirely on this approach quickly grows intractable.

Lastly, PsychSim is more than just an approach to modeling ToM. Because it is modular, adding additional capabilities becomes trivial, which renders it a general mechanism for artificial cognition. This is already demonstrated in the literature. The PsychSim approach, for example, can implicitly generate the appraisals found in appraisal theories of emotion (Si, Marsella, and Pynadath 2010). Also, its decision theoretic approach to ToM constrained mental models of others into exhibiting preference ordering (Pynadath and Marsella 2005).

Conclusion

There are numerous theories and modeling approaches that attempt to capture the essence of human theory of mind. The development of an agent capable of a similar level of ToM reasoning reveals where each theory and approach may falter. The requirements of creating such an agent, we believe, are also minimal requirements for actual ToM reasoning. Combining these requirements with lessons learned from attempts to explain and model ToM holds the potential of producing a more complete, viable theory.

Acknowledgements

Part of the effort depicted is sponsored by the U.S. Defense Advanced Research Projects Agency (DARPA) under contract number W911NF2010011, and that the content of the information does not necessarily reflect the position or the policy of the Government or the Defense Advanced Research Projects Agency, and no official endorsements should be inferred.

References

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

Axelrod, R.; and Hamilton, W. D. 1981. The evolution of cooperation. *science* 211(4489): 1390–1396.

Baker, C. L.; Saxe, R.; and Tenenbaum, J. B. 2009. Action understanding as inverse planning. *Cognition* 113(3): 329–349.

Baker, C. L.; Tenenbaum, J. B.; and Saxe, R. 2006. Bayesian models of human action understanding. *Advances in neural information processing systems* 18: 99.

Bloom, P.; and German, T. P. 2000. Two reasons to abandon the false belief task as a test of theory of mind. *Cognition* 77(1): B25–B31.

Bonaccio, S.; and Dalal, R. S. 2006. Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational behavior and human decision processes* 101(2): 127–151.

Bower, G. H.; Black, J. B.; and Turner, T. J. 1979. Scripts in memory for text. *Cognitive psychology* 11(2): 177–220.

Bowles, S.; and Gintis, H. 2003. The Origins of Human Cooperation, Peter Hammerstein (ed.) The Genetic and Cultural Origins of Cooperation.

Brosnan, S. F.; and De Waal, F. B. 2002. A proximate perspective on reciprocal altruism. *Human Nature* 13(1): 129–152.

Camerer, C. F. 2003. Behavioural studies of strategic thinking in games. *Trends in cognitive sciences* 7(5): 225–231.

Camerer, C. F. 2011. *Behavioral game theory: Experiments in strategic interaction*. Princeton university press.

Carruthers, P. 2016. Two systems for mindreading? *Review of Philosophy and Psychology* 7(1): 141–162.

Chudek, M.; and Henrich, J. 2011. Culture–gene coevolution, norm-psychology and the emergence of human prosociality. *Trends in cognitive sciences* 15(5): 218–226.

D'Mello, S.; Kappas, A.; and Gratch, J. 2018. The affective computing approach to affect measurement. *Emotion Review* 10(2): 174–183.

Fiske, S. T.; and Taylor, S. E. 2021. *Social cognition: From brains to culture*. Sage.

Fodor, J. A. 1983. *The modularity of mind*. MIT press.

Friston, K.; FitzGerald, T.; Rigoli, F.; Schwartenbeck, P.; Pezzulo, G.; et al. 2016. Active inference and learning. *Neuroscience & Biobehavioral Reviews* 68: 862–879.

Frost, R.; Armstrong, B. C.; Siegelman, N.; and Christiansen, M. H. 2015. Domain generality versus modality specificity: the paradox of statistical learning. *Trends in cognitive sciences* 19(3): 117–125.

Gallese, V.; and Goldman, A. 1998. Mirror neurons and the simulation theory of mind-reading. *Trends in cognitive sciences* 2(12): 493–501.

Gershman, S. J.; Gerstenberg, T.; Baker, C. L.; and Cushman, F. A. 2016. Plans, habits, and theory of mind. *PloS one* 11(9): e0162246.

Gigerenzer, G.; and Gaissmaier, W. 2011. Heuristic decision making. *Annual review of psychology* 62: 451–482.

Glikson, E.; and Woolley, A. W. 2020. Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals* 14(2): 627–660.

Gmytrasiewicz, P. J.; and Doshi, P. 2005. A framework for sequential planning in multi-agent settings. *Journal of Artificial Intelligence Research* 24: 49–79.

Gmytrasiewicz, P. J.; and Durfee, E. H. 1995. A Rigorous, Operational Formalization of Recursive Modeling. In *ICMAS*, 125–132.

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

Gopnik, A. 1993. How we know our minds: The illusion of first-person knowledge of intentionality. *Behavioral and Brain sciences* 16(1): 1–14.

Gopnik, A. 1996. The scientist as child. *Philosophy of science* 63(4): 485–514.

Gopnik, A.; Glymour, C.; Sobel, D. M.; Schulz, L. E.; Kushnir, T.; and Danks, D. 2004. A theory of causal learning in children: causal maps and Bayes nets. *Psychological review* 111(1): 3.

Gopnik, A.; and Meltzoff, A. N. 1994. *Minds, bodies and persons: Young children's understanding of the self and others as reflected in imitation and theory of mind research*, 166–186. Cambridge University Press. doi:10.1017/CBO9780511565526.012.

Gopnik, A.; Meltzoff, A. N.; and Kuhl, P. K. 1999. *The scientist in the crib: Minds, brains, and how children learn.* William Morrow & Co.

Gordon, R. M. 1986. Folk psychology as simulation. *Mind & language* 1(2): 158–171.

Gordon, R. M. 1995. Simulation Without Introspection or Inference From Me to You. In Davies, M.; and Stone, T., eds., *Mental Simulation*. Blackwell.

Gupta, P.; and Woolley, A. W. Forthcoming. Articulating the Role of Artificial Intelligence in Collective Intelligence: A Transactive Systems Framework. *Proceedings of the Human Factors and Ergonomics Society* .

Heider, F. 1958. *The psychology of interpersonal relations*. Lawrence Erlbaum Associates, Inc.

Heider, F.; and Simmel, M. 1944. An experimental study of apparent behavior. *The American journal of psychology* 57(2): 243–259.

Heyes, C. 2014. Submentalizing: I am not really reading your mind. *Perspectives on Psychological Science* 9(2): 131–143.

Heyes, C. M.; and Frith, C. D. 2014. The cultural evolution of mind reading. *Science* 344(6190).

Jara-Ettinger, J. 2019. Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences* 29: 105–110.

Kahneman, D. 2011. *Thinking, fast and slow*. Macmillan.

Khalvati, K.; Park, S. A.; Mirbagheri, S.; Philippe, R.; Sestito, M.; Dreher, J.-C.; and Rao, R. P. 2019. Modeling other minds: Bayesian inference explains human choices in group decision-making. *Science advances* 5(11): eaax8783.

Kirkham, N. Z.; Slemmer, J. A.; and Johnson, S. P. 2002. Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition* 83(2): B35–B42.

Kuhn, D. 1989. Children and adults as intuitive scientists. *Psychological review* 96(4): 674.

Lake, B. M.; Ullman, T. D.; Tenenbaum, J. B.; and Gershman, S. J. 2017. Building machines that learn and think like people. *Behavioral and brain sciences* 40.

Leslie, A. M. 1987. Pretense and representation: The origins of" theory of mind.". *Psychological review* 94(4): 412.

Leslie, A. M. 1994a. Pretending and believing: Issues in the theory of ToMM. *Cognition* 50(1-3): 211–238.

Leslie, A. M. 1994b. ToMM, ToBy, and Agency: Core architecture and domain specificity. *Mapping the mind: Domain specificity in cognition and culture* 29: 119–48.

Leslie, A. M. 2000. How to acquire a 'representational theory of mind'. *Metarepresentations: A multidisciplinary perspective* 197–223.

Leslie, A. M.; Friedman, O.; and German, T. P. 2004. Core mechanisms in 'theory of mind'. *Trends in cognitive sciences* 8(12): 528–533.

Low, J.; Apperly, I. A.; Butterfill, S. A.; and Rakoczy, H. 2016. Cognitive architecture of belief reasoning in children and adults: A primer on the two-systems account. *Child Development Perspectives* 10(3): 184–189.

Lu, H.; Yuille, A. L.; Liljeholm, M.; Cheng, P. W.; and Holyoak, K. J. 2008. Bayesian generic priors for causal learning. *Psychological review* 115(4): 955.

Marsella, S. C.; Pynadath, D. V.; and Read, S. J. 2004. PsychSim: Agent-based modeling of social interactions and influence. In *Proceedings of the international conference on cognitive modeling*, volume 36, 243–248.

McKinnon, M. C.; and Moscovitch, M. 2007. Domain-general contributions to social reasoning: Theory of mind and deontic reasoning re-explored. *Cognition* 102(2): 179–218.

Minsky, M. 2019. *A framework for representing knowledge*. de Gruyter.

Mitchell, J. P. 2005. The false dichotomy between simulation and theory-theory: the argument's error. *Trends in cognitive sciences* 9(8): 363–364.

Molnar, A.; and Loewenstein, G. 2021. Thoughts and Players: An Introduction to Old and New Economic Perspectives on Beliefs. *The Science of Beliefs: A multidisciplinary Approach (provisional title, to be published in October 2021). Cambridge University Press. Edited by Julien Musolino, Joseph Sommer, and Pernille Hemmer* .

Mukamel, R.; Ekstrom, A. D.; Kaplan, J.; Iacoboni, M.; and Fried, I. 2010. Single-neuron responses in humans during execution and observation of actions. *Current biology* 20(8): 750–756.

Newell, A. 1994. *Unified theories of cognition*. Harvard University Press.

Ong, D. C.; Zaki, J.; and Goodman, N. D. 2019. Computational models of emotion inference in theory of mind: A review and roadmap. *Topics in cognitive science* 11(2): 338–357.

Perner, J.; and Kühberger, A. 2005. Mental simulation. In *Other minds: How humans bridge the divide between self and others*, 174–189. The Guilfod Press New York.

Perner, J.; and Lang, B. 1999. Development of theory of mind and executive control. *Trends in cognitive sciences* 3(9): 337–344.

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

Pynadath, D. V.; and Marsella, S. 2007. Minimal mental models. In *AAAI*, 1038–1044.

Pynadath, D. V.; and Marsella, S. C. 2004. Fitting and compilation of multiagent models through piecewise linear functions. In *International Conference on Autonomous Agents: Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems-*, volume 3, 1197–1204.

Pynadath, D. V.; and Marsella, S. C. 2005. PsychSim: Modeling theory of mind with decision-theoretic agents. In *IJCAI*, volume 5, 1181–1186.

Quesque, F.; and Rossetti, Y. 2020. What do theory-of-mind tasks actually measure? Theory and practice. *Perspectives on Psychological Science* 15(2): 384–396.

Rabin, M. 1993. Incorporating fairness into game theory and economics. *The American economic review* 1281–1302.

Rizzolatti, G.; and Craighero, L. 2004. The mirror-neuron system. *Annu. Rev. Neurosci.* 27: 169–192.

Saxe, R.; and Wexler, A. 2005. Making sense of another mind: the role of the right temporo-parietal junction. *Neuropsychologia* 43(10): 1391–1399.

Scholl, B. J.; and Leslie, A. M. 1999. Modularity, development and 'theory of mind'. *Mind & language* 14(1): 131–153.

Scholl, B. J.; and Leslie, A. M. 2001. Minds, modules, and meta-analysis. *Child development* 72(3): 696–701.

Sellars, W.; et al. 1956. Empiricism and the Philosophy of Mind. *Minnesota studies in the philosophy of science* 1(19): 253–329.

Shafto, P.; Goodman, N. D.; and Frank, M. C. 2012. Learning from others: The consequences of psychological reasoning for human learning. *Perspectives on Psychological Science* 7(4): 341–351.

Si, M.; Marsella, S. C.; and Pynadath, D. V. 2010. Modeling appraisal in theory of mind reasoning. *Autonomous Agents and Multi-Agent Systems* 20(1): 14–31.

Simon, H. A. 1997. *Models of bounded rationality: Empirically grounded economic reason*, volume 3. MIT press.

Spelke, E. S.; and Kinzler, K. D. 2007. Core knowledge. *Developmental science* 10(1): 89–96.

Tenenbaum, J. B.; Griffiths, T. L.; and Kemp, C. 2006. Theory-based Bayesian models of inductive learning and reasoning. *Trends in cognitive sciences* 10(7): 309–318.

Trivers, R. L. 1971. The evolution of reciprocal altruism. *The Quarterly review of biology* 46(1): 35–57.

Turner, R.; and Felisberti, F. M. 2017. Measuring mindreading: a review of behavioral approaches to testing cognitive and affective mental state attribution in neurologically typical adults. *Frontiers in psychology* 8: 47.

Wang, N.; Pynadath, D. V.; and Hill, S. G. 2016. The Impact of POMDP-Generated Explanations on Trust and Performance in Human-Robot Teams. In *AAMAS*, 997–1005.

Wang, N.; Pynadath, D. V.; Rovira, E.; Barnes, M. J.; and Hill, S. G. 2018. Is it my looks? or something i said? the impact of explanations, embodiment, and expectations on trust and performance in human-robot teams. In *International Conference on Persuasive Technology*, 56–69. Springer.

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

Yoshida, W.; Dolan, R. J.; and Friston, K. J. 2008. Game theory of mind. *PLoS computational biology* 4(12): e1000254.