Leveraging first-person experience to infer third-person beliefs in a competitive gridworld task

Joel Michelson[1] , **Deepayan Sanyal**[1] , **Maithilee Kunda**[1,2]

[1]Dept. of Computer Science, Vanderbilt University, Nashville, United States
[2]School of Informatics, University of Edinburgh, Edinburgh, United Kingdom
{joel.p.michelson, deepayan.sanyal}@vanderbilt.edu, mkunda@ed.ac.uk

Abstract

A key challenge of theory of mind, or the ability to reason about others' mental states, is understanding the process by which others' perceptions influence their beliefs. While specific tasks benchmark human and animal abilities to infer beliefs, we know less about how such capabilities can be learned. In this work, we introduce a modular computational architecture for solving a competitive gridworld game in which two agents compete for treats. By systematically replacing components of a rule-based solution with neural networks, we identify whether different capabilities can be learned from narrow sets of experiences. We also implement and compare three novel strategies to improving generalization via explicit comparison of first- and third-person reasoning: parameter sharing based on first-person experience, parameter sharing across both first and third-person perspectives, and artificially inducing uncertainty to simulate varied belief formation.

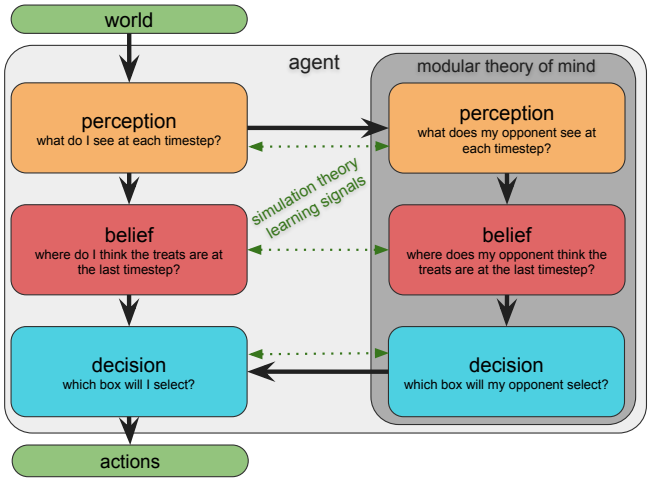

Figure 1: The general architecture used in this study to solve problems involving theory of mind. Inspired by simulation theory, the agent explicitly models the mental state of an opponent with a mirrored architecture for reasoning about treats' locations. To further leverage simulation theory, we may impose that modules' counterparts share parameters.

1 Introduction

Imagine you are at a wildlife preserve, watching a chimpanzee through a glass window. She was previously rewarded grapes for completing a psychology test, and is now carrying some away, perhaps to be shared among her community. Suddenly, she begins frantically searching below nearby bushes and trees. Without anyone telling you, you immediately understand her panic: she believes that she dropped a grape and it is now hidden nearby. How did you reach this conclusion? Is it because something similar has happened to you?

Theory of mind (ToM)—the ability to attribute mental states to others—emerged in early chimpanzee studies [Premack and Woodruff, 1978]. Since then, it has grown into a fundamental concept in cognitive science. Over the past five decades, developmental and comparative psychologists have proposed various theories to explain how humans and non-human animals acquire and use ToM skills. These theories sparked heated debates and led to numerous experimental paradigms, from false-belief tasks to nuanced measures of mental state attribution. One such theory, simulation theory, is a framework proposing that humans understand others' mental processes by using their own experienced mental processes as models, i.e., that we understand others by mentally putting ourselves in their shoes and simulating what they might think in their circumstances.

Recent advancements in computational modeling provide new tools for analyzing ToM processes in controlled, replicable environments. These simulated models offer unique insights into the cognitive mechanisms underlying ToM, complementing traditional behavioral studies. However, researchers continue to face challenges in creating computational models that generalize learned skills to novel scenarios, mirroring difficulties observed in non-human primate studies.

In this paper, we use a computational benchmark to investigate models of simulation theory. Specifically, we

- Present a modular architecture for decomposing competitive feeding ToM tasks into simpler, differentiable, and rule-based components, described in Section 3.
- Evaluate the necessity of each module in our framework towards robust ToM generalization, in Section 4.

- Analyze the effect of learning different sets of modules from narrow experiences on generalization, in Section 5, revealing a fundamental asymmetry in ToM learning.
- Implement and evaluate multiple computational analogs of simulation theory, and show that parameter sharing is insufficient for robust generalization, in Section 6.

2 Related Work

2.1 Computational Models of ToM

This work is heavily inspired by the ToMnet experiments of Rabinowitz et al. [Rabinowitz *et al.*, 2018]. In their study, they implement machine learning models with explicit ToM-like representations about agents' attributes and mental states, and are able to leverage the computational setting to probe those models for representations of those features.

Recently, Horschler et al. [Horschler *et al.*, 2023] used computational modeling to investigate ToM capabilities in non-human primates, focusing on visual perspective-taking tasks similar to the one investigated by this paper. They developed seven models of varying complexity to represent different theories of primates' social cognition, and parameterize the subjects' reliance on their ToM inferences to determine how well the theories explain primate behavior. [Quillien and Taylor-Davies, 2025] found that resource-limited agents more optimally track what others know rather than what they believe, successfully imitating patterns of test results seen in primates and young children.

Computational ToM skills have also been particularly well-studied recently in the context of large language models (LLMs). The ToMi dataset by Le et al. [Le *et al.*, 2019] consists of short, structured narratives based on the Sally-Anne false belief test. ToMi focuses primarily on first-order ToM reasoning about physical world states. More recently, Xu et al. developed OpenToM [Xu *et al.*, 2024] to benchmark ToM capabilities in large language models using longer, more natural narratives, covering both physical and psychological aspects of ToM. Despite recent advancements, LLMs continue to underperform humans on complex ToM tasks, highlighting the difficulty in acquiring robust ToM skills in machine learning models.

2.2 Simulation Theory

Developmental and comparative psychologists have produced a variety of theories about potential ToM mechanisms. When investigating a theory, we must note the context in which it was generated, which scientific problems it is meant to address, and how it contrasts with its counterparts.

Simulation theory emerged as an alternative to "theory theory" as a potential solution to the problem of how humans attribute mental states to others without directly accessing their minds. Under theory theory, mental state attribution is achieved by leveraging the human ability to do empirical science, where children refine their understanding of others' minds by testing internal models during social interactions [**?**]. In contrast, simulation theory involves using one's own mental states as a model for simulating those of others. It addresses a developmental problem with theory theory, as children demonstrate ToM abilities at ages when they haven't yet developed sophisticated conceptual abilities that theory theory seems to require. Simulation theory offers a more parsimonious explanation by suggesting that ToM skills could leverage existing mechanisms, rather than learning complex theoretical frameworks. Additionally, simulation theory could suggest a direct neural basis for ToM skills using mirror neurons, which are thought to enable the internal mirroring of others' actions and emotional states [Gallese and Goldman, 1998]. From a developmental perspective, simulation theory emphasizes the role of pretend play, which provides children with opportunities to practice mental simulation [Harris, 1992]. This contrasts with theory theory's emphasis on hypothesis testing and conceptual development.

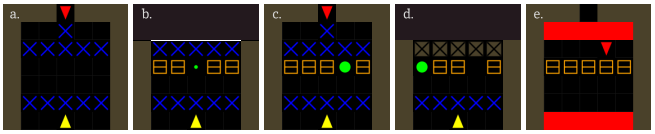

Figure 2: Sample trial from the Standoff environment. A) The subject (teal) and opponent (red) face each other, separated by transparent walls (blue). B) The small treat (green) is placed, along with four boxes. The opponent's vision is obscured during this step. C) The large treat (green) is placed. D) The large treat swaps locations to the leftmost position, while the opponent's vision is obscured again. E) The opponent selects the last location where it observed the large treat. The opponent is uninformed about the existence of the smaller treat, having never seen it, but is misinformed about the location of the larger treat, harboring a false belief. Because the opponent is misinformed, this trial is categorized as ToM-Complex, one of our test sets described in Section 3.

3 The Competitive Feeding Task

The competitive feeding paradigm is a test setup designed to distinguish whether a non-verbal subject will change its behavior to account for what it believes someone else (an "opponent") *knows*, based on evidence relating to what it can perceive that the opponent *sees* [Hare *et al.*, 2000].

In this paper, we use the Standoff environment [Michelson *et al.*, 2022; Michelson *et al.*, 2024], a gridworld setting that replicates the competitive feeding paradigm in the style of Penn and Povinelli [Penn and Povinelli, 2007]. In Standoff tasks, two treats of different sizes are visibly hidden in any of five boxes, which are then shuffled around. The player's challenge is to select the box containing the best possible treat.

This is made difficult by the presence of an opponent. The opponent follows simple rules: if it believes the larger treat is somewhere, it will claim that box, preventing the player from taking whatever treat is inside. Otherwise, the opponent will attempt to take the smaller treat, or will select a preferred box. These rules are obfuscated by the opponent's vision being obscured during the setup. The opponent might be unaware that either treat exists, or it might harbor a counterfactual belief about either of the treats' locations. Whenever the opponent's vision is obscured in the computational setting, it assumes nothing has changed in the environment. In real settings this assumption is made true by repeated trials. The player must either stay clear of the opponent or take advantage of the opponent's unawareness.

For supervised learning, each datapoint is collected from a single trial, or a (5, 5, 7, 7)-sized video, of five timesteps, five channels (player and opponent, large treats, small treats, boxes, barriers), seven tiles in width and seven tiles in height. The target output to be learned is the *correct* box, meaning the player's best choice of the five boxes, given the opponent's selection.

In this paper, we categorize the environment trials into four datasets taken in different combinations for training and evaluation, patterned off of Penn and Povinelli's description of systematic competitive feeding. **Solo** has all tasks without an opponent present. **Informed** has all tasks with a fully-informed opponent. **ToM-Simple** has all trials where there are zero swaps visible to the opponent and either the opponent is only uninformed about the large treat and will choose the smaller treat or the opponent is uninformed about both treats and will default to a deterministic choice, box 2. Finally, **ToM-Complex** contains all other opponent mental states, including all examples of Misinformed-ness. ToM-Simple, a deviation from systematic competitive feeding, is included because a learning player exposed to Solo and Informed has no exposure to opponent behavior when it is less than fully informed.

3.1 A Modular Simulation Theory Architecture

Previous work found that various end-to-end (E2E) neural network architectures (e.g. MLPs, CNNs, LSTMs) trained on subsets of Standoff were able to learn tasks present in the training data to high accuracies (>95%), but struggled to generalize to novel settings [Michelson *et al.*, 2024]. The E2E training stymied an understanding of which aspects of the task were or were not being learned in a way that generalized. Certain aspects of the Standoff task, e.g. tracking whether a treat has been placed, are not intended to be difficult and should also generalize well to novel test cases. Others, e.g. tracking whether an opponent *observed* a treat being placed, are more relevant to ToM research. Additionally, testing specific ToM theories (like simulation theory) requires architectural control that E2E models cannot provide. How would one test parameter sharing between self and other reasoning in a black-box E2E model? To address this distinction, we present a modular architecture whose components may be altered independently, visualized in Figure 3. Our modular architecture implements simulation theory's premise that agents can use self-models to predict others' mental states by using identical modules for both self- and opponent-reasoning, showcased conceptually in Figure 1. Note that this architecture is one of many possible solutions to the task, selected for its relative simplicity for interpretability.

The architecture works as follows: First, information is extracted from perceptual signals by two pipelines: one for the player, and the other a simulation of the opponent. The extracted information is used to predict both the player and the opponent's beliefs. The simulated opponent, a greedy participant in this task (i.e., a 'dominant' in tests with chimpanzees), makes a decision based only on its beliefs; it infers the location of the best treat that it has seen. Finally, the player makes a decision given both its own beliefs and the opponent's predicted decision.

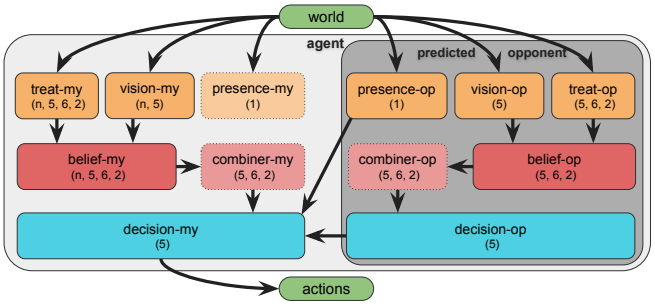

Figure 3: The specific model architecture used in this study. Inputs are processed to produce: tensors indicating the treats visible at each of 5 timesteps and 6 positions (including null) of both the large and small treat (**treat**), whether the opponent's gaze is obscured at each timestep (**vision**), and whether either player is present (**presence**). Only the opponent presence module is used prior to Section 6. **Belief** modules use the former two outputs to predict the treats' locations at the final timestep. **Combiner** modules combine n multiple uncertain beliefs into one. **Decision** modules use a belief vector and (only as the subordinate player) the opponent's decision to predict the location harboring the largest available treat.

In Table 1, we describe both **neural** and **rule-based** implementations of modules. Despite the modular construction, learning is end-to-end: All rule-based modules are differentiable functions. This way, we may replace any module with a neural network whose weights may be optimized with respect to the final behavioral output alone; we do not train learnable modules prior to the player's decision to predict specific values. This constraint enforces that our player only learns from its embodied observations and actions in first-person; it cannot use learning signals such as those implied by theory-theory to learn an independent ToM by observing opponent behavior.

4 Analyzing Module Importance

We replace individual hardcoded modules with random functions to explore the effect of model failure. Most rule-based module outputs are (continuous but sharp) one-hot selections along some tensor dimension, so uniformly random outputs may be generated trivially for each. Only one, vision-op, is produced by random binary vectors. We test eight ablated architectures by replacing in each a single module of our initial **rule-based** architecture with a random output.

We hypothesize that the architecture requires all components. If any fails, performance on some test set must fall.

4.1 Results

Accuracy results are shown in Figure 4.

Our rule-based model achieves perfect accuracy (ceiling performance) on all datasets, establishing that the task is solvable with explicit reasoning, and our modular decomposition captures all necessary processes. Performance degradation in neural modules therefore reflects learning limitations, not architectural inadequacy.

The randomized my-decision model achieves the expected random chance performance at 20% (floor performance).

Table 1: Module descriptions and implementations

Treat Perception	**Function:** Processes raw visual input to identify treat locations at each timestep. **Rule-based:** Extracts treat positions from specific perceptual field channels, applies a sharp sigmoid to highlight likely locations, and uses softmax to create probability distributions across all possible positions including a "no treat" option. **Neural:** A feed-forward network transforms spatial features (5×5 inputs) into separate probability distributions for both large and small treats (5 boxes + null), with softmax normalization.
Vision Perception	**Function:** Determines when an agent's vision is obscured during the trial. **Rule-based:** Examines specific coordinates in the visual field associated with vision states, applies sigmoid activation to transform continuous values into binary vision states for each timestep. **Neural:** A linear network processes 5 timesteps of the vision channel to indicate whether vision is obscured at each timestep.
Presence Perception	**Function:** Detects whether an opponent is present in the environment. **Rule-based:** Directly extracts presence indicator value from the first timestep at specific coordinates, outputting a binary signal. **Neural:** A single linear layer transforms the extracted coordinate value with sigmoid activation to produce a binary presence indicator. The model learns to identify the specific input pattern associated with opponent presence.
Belief	**Function:** Integrates treat position observations and vision data over time to form beliefs about final treat locations. **Rule-based:** Uses exponentially-weighted temporal integration (e^{2t} weights) to emphasize recent observations. Computes both positional beliefs and a "never seen" probability for each treat. An uncertainty parameter simulates odds of treats changing during unseen timesteps (only applicable in the vision masking models of Section 6). This parameter is determined empirically; 0.3 for the player, and 0.0 for the opponent, who assumes explicit object permanence. **Neural:** A two-layer network (35→16→6) with softmax normalization processes treat positions and vision.
Combiner	**Function:** Reconciles multiple beliefs from different experiences into a single belief, only in masked-vision models in Section 6. **Rule-based:** Computes entropy-based confidence weights ($1 - \sum p \log p$) for each belief distribution, applying higher weights to more certain beliefs. Takes the maximum across vision scenarios and normalizes the result. **Neural:** Encodes beliefs into 12-dimensional latent vectors with ReLU activation, applies max-pooling across multiple belief instances, then decodes back to a unified belief distribution.
Decision	**Function:** Uses beliefs about treat locations and the predicted opponent presence and decision to select the optimal box. **Rule-based:** First attempts to claim the large treat if it is believed present and uncontested by an opponent; if unavailable, attempts to claim the small treat; if no treats are believed present, defaults to a predetermined position, location 2. Contested treats are detected continuously by multiplying the predicted large position and the predicted dominant decision. **Neural:** A two-layer network (18→16→5) processes the combination of beliefs, opponent's predicted choice, and opponent presence.

Implementation details, hyperparameters, and code available at: https://github.com/aivaslab/standoff

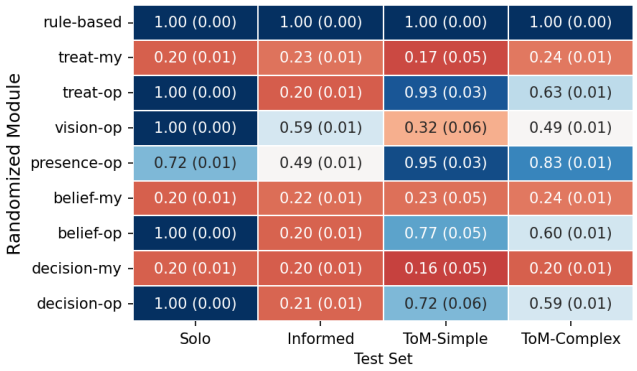

Figure 4: Mean and standard deviation accuracies of models with randomized modules on each of the four test sets. Each row describes which module has been replaced with random outputs (rule-based has none), and each column describes a test set. Standard deviations are aggregated over all trials among evaluated models.

Randomizing the treat perception module or belief-my modules lead to random or nearly random performance. Certain modules achieve perfect performance on Solo only, including vision-op, belief-op, and decision-op. Those modules' outputs are irrelevant to the Solo task in which no opponent is present. In Informed, presence-op is correct half the time, resulting in a policy that correctly selects one of the two treats in half of all trials. Vision-op performs slightly better, reflecting the likelihood that the opponent is predicted to be informed given 50% obscured timesteps. In ToM-Simple, where opponents are uninformed, vision-op performs poorly; the player tends to incorrectly assume the opponent is informed. In ToM-Complex, the only task including misinformedness, randomizing the prediction of the opponent's vision leads to the worst performance.

5 Learning from Limited Experience

Next, we examine the effect of learning to generalize capabilities from constrained datasets by substituting each of our random modules with neural networks, and train those networks on each of our datasets. We supplement the list of ablated architectures from Section 4 with two new architectures featuring multiple neural modules: **all-my** learns the full pipeline of the player's first-person information. **all-op** learns the full pipeline for third-person predictions of opponent information.

We hypothesize that, because each dataset is a superset of the previous, models will only improve on each evaluation dataset in sequence corresponding with stage. We also predict that perceptual modules will generalize well across all datasets since they learn basic environmental features, only if those modules are useful for the trained task. For example, we predict that treat-my should generalize well from narrow training data, since it is useful for all tasks, but treat-op will not generalize from Solo to other stages since it is irrelevant during training. We predict that opponent belief and decision modules will show poor generalization without explicit ToM training, even when trained on tasks involving opponent beliefs.

Table 2: Test datasets and cumulative training datasets

Test Dataset	
Solo	Trials without an opponent present
Informed	Trials with a fully-informed opponent
ToM-Simple	Trials with an opponent fully uninformed or uninformed about the large treat only (no observed swaps)
ToM-Complex	Trials with all other opponent mental states
Training Dataset	
Solo-Train	Contains only Solo trials
Informed-Train	Contains Solo + Informed trials
ToM-Simple-Train	Contains Solo + Informed + ToM-Simple trials
ToM-Complex-Train	Contains all trials

5.1 Training Methodology

We train each model for 4000 batches of size 1024. Because we are interested in capabilities, we train twenty models of each setting with randomized initialization and batches. Models are trained using the AdamW optimizer with beta values of 0.95 and 0.999 and a learning rate of 0.01. Validation sets are formed by a 90/10 split of the training distribution. Hyperparameters—including neural model architectures, weight initialization, learning rate, and momentum—were manually tuned to maximize consistent validation-set convergence of each module. The sharp sigmoid temperature of 90.0 was found by perturbing perceptual inputs with the hardcoded module. Using lower sigmoid temperatures was favorable for learning the earlier perceptual modules, but low temperatures enforced a hard maximum on model accuracy. Scheduling this parameter did not aid with convergence. We use feedforward networks for controlled comparison; architecture choices likely affect generalization beyond the scope of this paper.

For training, we use four datasets, each extending the prior with the proceeding test set from above, described in Table 2. This means that the training sets are shown in sequence, each a superset of the previous. Solo corresponds to Stage 1 of Penn and Povinelli's Competitive Feeding description, Informed-Train corresponds to Stage 2, and ToM-Complex-Train corresponds to Stage 3.

The intended difficulty of Competitive Feeding lies in generalization to Stage 3 tasks rather than convergence to training sets, so we select the three of our twenty trained models with the lowest validation loss after training for evaluation on our test sets. Each of our test sets probes specific generalization capabilities. Informed-test checks whether the player is able to correctly reason with an opponent present; when trained on Solo, where the optimal treat is always large, we expect floor performance for the decision module, but better performance for modules unrelated to opponents (e.g. treat-my, perception-my, belief-my). ToM-Simple-Test checks whether the player reasons about opponent awareness and unaware-

ness, and ToM-Complex-Test checks whether the player reasons correctly about misinformedness, including false beliefs.

5.2 Results

Figure 5 reveals striking patterns in how different ToM components can be learned from constrained experiences. Standard deviations are low across all conditions, confirming training stability. Training on ToM-Complex-Train, omitted for brevity, consistently yields high accuracy, confirming our models can learn the task without underfitting. As hypothesized, generalization tends to improve with exposure to more varied data, with notable exceptions. The treat-my and perception-my modules (first and fifth rows of each table) demonstrate strong generalization across all test sets, confirming our prediction that perceptual modules would generalize well from narrow experiences when relevant to all tasks. This suggests that the basic perceptual components of ToM can be learned robustly from limited experiences. Presence-op becomes learnable after Informed training as expected, since this stage introduces opponent-based policy changes. ToM-Simple training enables generalization to ToM-Complex for treat-op and belief-op modules. The vision-op module consistently demonstrates the poorest generalization (40-52% on ToM-Complex when trained on simpler datasets), confirming our prediction that third-person visual perspective-taking represents a core challenge for ToM reasoning. Belief-my generalizes surprisingly poorly (88%) despite the exposure to identical swap patterns in Solo trials, suggesting that incorporating differentiable rule-based computations with neural networks poses significant challenges for learning. This architectural constraint, also evident in decision-my performance, highlights the asymmetry between first-person and third-person module learning that we observe throughout this experiment.

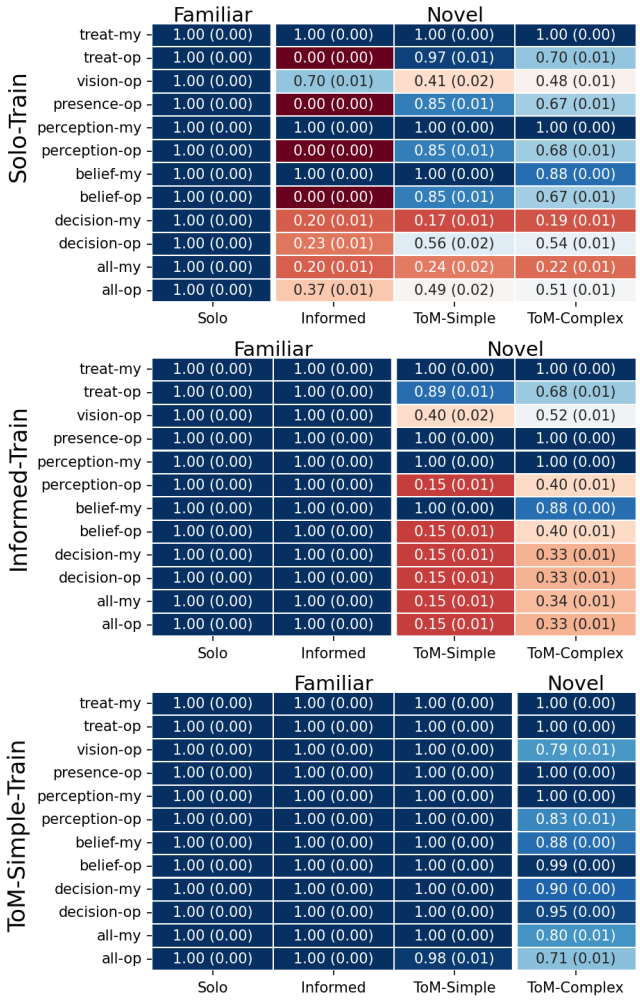

Figure 5: Learning module experiment accuracies across different training sets and learned modules. Top table: Training on Solo data. Middle table: Training on Informed data (Solo + opponent present). Bottom table: Training on ToM-Simple data (previous + basic opponent unawareness). Each row is a different model configuration, and each column is a different test dataset. Each cell shows mean and standard deviation of accuracy among our top three models of twenty training runs. Results from ToM-Complex-Train are omitted for brevity; all achieve near perfect (>98%) validation accuracy with low variance, excepting perception-op which failed to produce any model that fully converged to all four columns during training. Generally, modules are able to converge to training data, but few generalize to novel tasks. With more varied training data, more modules are able to generalize to ToM-Complex trials.

6 Parameter Sharing

In the previous experiment, we found that first-person modules tend to generalize to novel experiences to different degrees than their third-person counterparts. A successful model of ToM should model opponent reasoning in ways that generalize beyond training data, so explaining this performance gap is of critical importance for understanding ToM learning. In particular, we aim to discern whether architectural strategies like sharing self- and opponent-modules' parameters are sufficient for better generalizing reasoning in this task, or if instead other changes to the setup—e.g. curricular training, modified learning signals, or more task-specific module architectures like LSTMs—might be necessary.

In this experiment, we explore strategies for learning both self- and opponent-modules simultaneously. For comparison, we refer to learning both modules (e.g. both my beliefs and opponent beliefs) simultaneously as **split** models. To make our player learn to simulate an opponent's reasoning by practicing its own reasoning skills, we shall impose that first- and third-person modules be functionally identical. They might be learned from both the player and the simulated opponent's experiences (**shared**), or they might only be learned from the player's experiences (**detached**).

Note the asymmetry in the parameter sharing task: while the opponent experiences many different cases of uncertainty, the subject only experiences *certain* beliefs. Because of this asymmetry, there is no first-person experience of how different belief states are formed from uncertain observations. To address this asymmetry, We induce uncertainty by randomly masking vision for the player at each timestep. We produce these masked perceptions multiple times, resulting in the player having a *set* of beliefs, which may contradict each other. The sets of beliefs are resolved back into one belief by the combiner module. In this experiment, we examine the effect of masking vision on shared and detached models using sets of five vision masks. The obscuring probability begins at 0% when training begins and increases with cosine interpolation to 50% halfway through training.

We hypothesize that split architectures will perform similarly to the third-person models in Section 5, since they must learn the same non-generalizing modules under similar circumstances, and that parameter sharing between player and opponent modules will improve generalization by leveraging the symmetry of perspective-taking in ToM reasoning. While detached modules are a stronger constraint on learning—disallowing the player from learning from predicted opponent experiences—we anticipate that they will improve upon split architectures whenever the player and opponent tasks are symmetric. When asymmetric, we predict that the introduction of first-person uncertainty to the player will cause shared and detached architectures to learn to perform similarly.

6.1 Results

Figure 6 demonstrates the limitations of our simulation theory implementation using parameter sharing. For treat and belief models, shared and detached training does not improve generalization over split training. For decision and all models, shared training shows only marginal improvements in generalization over split performance. No models consistently generalize well to ToM-Complex, and performance is not substantially improved over single-module training approaches. Vision masking (marked -mv in Figure 6) sometimes results in imperfect convergence to the training data across nearly all configurations, as expected when introducing ambiguity into perception. While vision masking notably improved the prediction of treats over the non-masked shared models, it did not consistently aid with generalization for other models, including those with split treat modules. These results challenge our hypotheses about parameter sharing and suggest that while architectural symmetry may provide modest benefits for higher-level reasoning components, more sophisticated approaches are needed to enable robust theory of mind generalization across diverse scenarios.

7 Discussion and Limitations

Our randomized module ablation experiment in Section 4 demonstrates that all components of our modular architecture are necessary for robust performance. The learning experiments reveal a clear asymmetry between first-person and third-person module learning. Modules related to the agent's own perceptions and beliefs (treat-my, perception-my) generalize well from limited experience, maintaining near-perfect

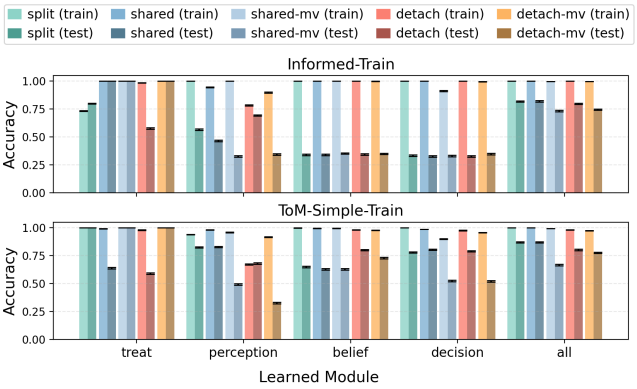

Figure 6: Accuracy results from the parameter sharing experiment plotted as bars for module sets trained on Informed-Train and ToM-Simple-Train. These are aggregated into two groups: train (familiar tasks) and test (novel tasks). All strategies for parameter sharing are compared: **split** do not share parameters. **Shared** and **detach** do, where detach only learns from first-person experience. "**-mv**" refers to masked vision variants; vision is not masked at test time. Error bars depict standard deviations across all trials. Generalization is not consistently improved by any of our parameter sharing strategies, highlighting the challenges of learning robust ToM capabilities.

accuracy across test conditions. A consistent generalization gap between first-person and third-person modules led us to investigate whether parameter sharing could leverage successful mechanisms to improve opponent modeling. Our parameter sharing experiments reveal that shared architectures do not consistently improve generalization over split architectures. These results demonstrate that simple parameter sharing may be insufficient to capture the full complexity of ToM reasoning in this setting, and more sophisticated mechanisms may be needed to capture the full complexity of human-like ToM capabilities.

Generalization is difficult to achieve for isolated modules learning to perform simple tasks. For a module to converge but not overfit, it must have not only sufficient learning signals, but also dynamics that tend to approximate desirable functions *outside* its experience. In this paper, although we experimented with different simulation theory-inspired learning signals on well-specified training sets, we did not examine the effect of intra-module architecture. It could be the case that given a more ideal learning signal (e.g., someone saying, "Your prediction was wrong, I believed *X*"), a learned module might still be unable to generalize to novel trials. Our supervised learning approach forces modules to optimize for task performance, where the only belief that matters is the one at the last timestep, rather than developing flexible reasoning mechanisms. Given our module-specific findings, a clear next step is to experiment without any imposed embodiment or simulation theory constraints. Different architectures such as recurrent networks or transformers are likely better suited for generalizing reasoning about sequential information. By isolating the problem of intra-module generalization, we may better understand the effect of learning signals and explicit knowledge representation.

Our training focuses on supervised learning from correct

action selection, which differs significantly from humans' development involving interactive experience. This embodiment restriction ignores how humans learn ToM through rich social interaction, potentially explaining why our simulation theory implementations underperform. While our architecture is inspired by psychological theories, it abstracts away many details of human cognitive processes. This approach allows for controlled comparison but it might also limit the models' ability to discover novel strategies or representations. These may be better captured using different training methods such as curricular learning or auxiliary loss signals.

Our training and testing datasets could also be improved for further insight into ToM capability learning. Future work shall distinguish Gettier cases [Gettier, 1963] from informed beliefs, as they produce different developmental patterns in children compared to false belief reasoning [Fabricius *et al.*, 2010; Oktay-Gür and Rakoczy, 2017]. Our environment contains multiple types of opponent errors that could be analyzed for their generalization difficulty and training value.

Generally, we have found that ToM reasoning can be decomposed into sufficient, separable components, but these components have very different learning requirements and generalization capabilities. While this paper focuses on simulation-theory-inspired mechanisms, the method that we showcase allows us to analyze completely different hypotheses about learned ToM mechanisms. Theory-theory, to which we have repeatedly compared simulation theory, might be simulated by providing explicit learning signals to models that predict observed opponent behavior. Alternatives include mentalizing/systemizing theory, which has been used to explain differences in the ToM skills of individuals on the Autism spectrum by differentiating between reasoning about social versus non-social rules [Baron-Cohen, 2000]. Future work could leverage our modular setup to compare such approaches to investigate their effect on generalization, and perhaps more importantly, combine multiple approaches to varying degrees to better understand the importance of structured belief representations in human ToM development.

Acknowledgements

This research was supported in part by the US NSF (NRT #1922697).

References

[Baron-Cohen, 2000] Simon Baron-Cohen. Theory of mind and autism: A review. *International review of research in mental retardation*, 23:169–184, 2000.

[Fabricius *et al.*, 2010] William V Fabricius, Ty W Boyer, Amy A Weimer, and Kathleen Carroll. True or false: Do 5-year-olds understand belief? *Developmental Psychology*, 46(6):1402, 2010.

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

[Gettier, 1963] Edmund L Gettier. Is justified true belief knowledge? *analysis*, 23(6):121–123, 1963.

[Hare *et al.*, 2000] Brian Hare, Josep Call, Bryan Agnetta, and Michael Tomasello. Chimpanzees know what conspecifics do and do not see. *Animal Behaviour*, 59(4):771–785, 2000.

[Harris, 1992] Paul L Harris. From simulation to folk psychology: the case for development. *Mind & Language*, 1992.

[Horschler *et al.*, 2023] Daniel J Horschler, Marlene D Berke, Laurie R Santos, and Julian Jara-Ettinger. Differences between human and non-human primate theory of mind: Evidence from computational modeling. *bioRxiv*, pages 2023–08, 2023.

[Le *et al.*, 2019] Matthew Le, Y-Lan Boureau, and Maximilian Nickel. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5872–5877, 2019.

[Michelson *et al.*, 2022] J Michelson, D Sanyal, J Ainooson, Y Yang, and M Kunda. Experimental design and facets of evidence for computational theory of mind. In *Proceedings of the 8th International Workshop on Artificial Intelligence and Cognition (AIC)*, 2022.

[Michelson *et al.*, 2024] Joel Michelson, Deepayan Sanyal, James Ainooson, Effat Farhana, and Maithilee Kunda. Standoff: benchmarking representation learning for nonverbal theory of mind tasks. In *2024 IEEE International Conference on Development and Learning (ICDL)*, pages 1–6. IEEE, 2024.

[Oktay-Gür and Rakoczy, 2017] Nese Oktay-Gür and Hannes Rakoczy. Children's difficulty with true belief tasks: Competence deficit or performance problem? *Cognition*, 166:28–41, 2017.

[Penn and Povinelli, 2007] Derek C Penn and Daniel J Povinelli. On the lack of evidence that non-human animals possess anything remotely resembling a 'theory of mind'. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1480):731–744, 2007.

[Premack and Woodruff, 1978] David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526, 1978.

[Quillien and Taylor-Davies, 2025] Tadeg Quillien and Max Taylor-Davies. Factive mindreading reflects the optimal use of limited cognitive resources. *structure*, 1:7, 2025.

[Rabinowitz *et al.*, 2018] Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, SM Ali Eslami, and Matthew Botvinick. Machine theory of mind. In *International conference on machine learning*, pages 4218–4227. PMLR, 2018.

[Xu *et al.*, 2024] Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. Opentom: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. *arXiv preprint arXiv:2402.06044*, 2024.