Standoff: benchmarking representation learning for nonverbal theory of mind tasks

Joel Michelson, Deepayan Sanyal, James Ainooson, Effat Farhana, Maithilee Kunda
Vanderbilt University, Nashville, United States
joel.p.michelson@vanderbilt.edu, deepayan.sanyal@vanderbilt.edu, james.ainooson@vanderbilt.edu,
effat.farhana@vanderbilt.edu, mkunda@vanderbilt.edu

***Abstract*—We present our design and implementation of Standoff, an innovative benchmark suite of computational theory of mind tasks, based on the competitive feeding paradigm from comparative psychology. We find that a small convolutional LSTM model without explicit theory of mind mechanisms can reach high levels of accuracy when exposed to the full variety of our task design during training. Such a model faces generalization challenges when exposed to narrower subsets of tasks. Finally, we discuss how this test may be used as a gateway for studying theory of mind skills beyond attribution of seeing and knowing.**

I. Introduction

Imagine you are a juvenile chimpanzee, in a deep local minimum within the social hierarchy of your tribe. A delicious fig falls to the ground in full view of your boss, but then it keeps rolling while out of his view. You have no chance in a head-to-head fig battle, but *you* know that *he* thinks the fig is somewhere other than where it *really* is, which means you can make a dash for it! This scenario presents a canonical example of theory of mind (ToM) reasoning—reasoning about the beliefs, desires, goals, and other invisible mental states of social agents in one's environment.

How do intelligent agents acquire ToM capabilities? In addition to rich webs of evolutionary, environmental, and sociocultural circumstances [1], agents must also learn from their experiences and, importantly, be able to generalize their knowledge appropriately to new situations. For example, the young chimpanzee ought to have learned, at some point, about lines of sight. However, our young chimp need not have seen all possible lines of sight, nor all possible permutations of events. More likely, some basic principles are learned that can then transfer to a wide variety of scenarios as needed.

Studying ToM, especially in non-human animals, is notoriously difficult because it is hard to discern whether an individual takes actions for ToM-related reasons or for other, non-ToM reasons [2]. Did you go for the fig because you knew what your boss didn't know? Or was it because you knew that it is okay if you cannot see his eyes? It is to solve this thorny problem of interpretation that comparative psychologists have designed elaborate, multi-stage ToM experiments. Using these setups, an individual animal's ToM capabilities can be deduced not necessarily from their performance on a single task, but from their performance across a structured set of tasks.

This work was supported in part by the Neurodiversity Inspired Science and Engineering (NISE) NSF NRT grant DGE 19-22697 and NSF BPE grant 22-17621 (K. Stassun, PI).

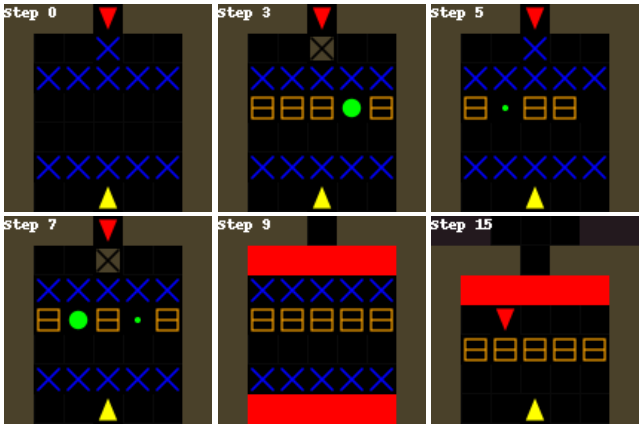

Fig. 1. Six steps from a challenging Standoff task. **(1)** The learning agent (yellow triangle) competes with an opponent (red triangle) for treats (green circles) hidden in boxes (orange squares), with the opponent having priority if they both choose the same one. **(2)** A large treat is placed in a box while the opponent's vision is obscured. **(3)** With the opponent's vision then unobscured, a small treat is placed in a different box. **(4)** Finally, the opponent's vision is obscured again while the two treats swap locations. **(5)** Once both agents are released (with red 'curtains' shielding their movements from each other), the learning agent should choose the smaller treat, because **(6)** it should know that the opponent will choose the true location of the larger treat, under a doubly false belief that that is where the smaller, only treat is.

Computational cognitive modeling allows researchers to empirically analyze hypotheses about how ToM is represented, relying on controlled and varied tests for comparison. Many current computational ToM tests do not capture the variety of interactions needed to differentiate candidate explanations about embodied agents' reasoning in response to the changing beliefs of others. This lack of comprehensive testing scenarios makes it challenging for researchers to investigate how machine learning models may learn and apply ToM representations to unfamiliar interactions. In this paper, we:

1) Present our design of a benchmark suite of computational ToM tasks, based on the competitive feeding paradigm from comparative psychology (Sections III and IV).
2) Demonstrate that a convolutional LSTM model excels when fit to all regimes in our benchmark, but struggles to generalize when trained on subsets of the dataset (Sections V, VI, and VII).

3) Discuss how competitive feeding may be used for studying ToM skills, both at a finer-grained level and beyond attribution of seeing and knowing (Section VIII).

II. Related Work

A previous review focusing on the design of many different tests for ToM skills developed by comparative and developmental psychologists [3] justifies the use of a test paradigm called *competitive feeding* as a benchmark for studying the ToM skills of embodied computational agents. Competitive feeding is described in detail in Section III of this paper.

Our research is closely related to work that benchmarks the abilities of computational models that learn to predict the motivation and mental states of other agents from experience. Recently, two extensive benchmark tests, the Baby Intuitions Benchmark (BIB) [4] and AGENT [5] have been proposed to test the extent to which machines can reason about the intentions that drive agent actions. BIB is based on the Violation of Expectation (VOE) paradigm and contains several tasks, each of which has a familiarization phase and a test phase. During the familiarization phase, multiple trial videos show an agent performing some task, such as moving to a goal object. In the test phase, the model watches another video and determines whether the expected action of the agent is violated. BIB investigates whether AI agents can learn representations related to object preference and instrumentality of actions toward higher-order goals. The AGENT environment contains similar trials and a more diverse range of physical situations, such as ramps and bridges. The AGENT environment's evaluation protocol tests whether models can generalize from different kinds of learning trials, i.e. learn to perform well when the test trials belong to a different distribution than the training trials.

Several works explore the design of computational agents that can model the mental states and intentions of other agents. [6] developed the ToMNet architecture to model agents' intentions in order to predict their future actions. Its effectiveness is limited to cases where the test distribution is similar to the training distribution. More recently, [7] built reinforcement learning models whose policies are conditioned on beliefs about other agents. This kind of formulation leads to models that perform better than models without this conditioning. However, trained models perform much worse compared to when ground-truth beliefs are available to the learners, demonstrating a gap in modeling the beliefs of other agents.

Despite recent progress in building agents that demonstrate ToM, the nature of learning and representational mechanisms that lead to robust ToM skills remains an open problem. While the ToMNet architecture [6] demonstrates that it is possible to model agents' beliefs using specialized architecture, these architectural changes are insufficient for the model to generalize in cases where the test data is sampled from a different distribution [4], [5]. In contrast, [5] shows that a model that combines Bayesian Inverse Planning with strong built-in representations of object physics is able to effectively generalize to different evaluation settings in the AGENT benchmark.

ToM has been tested in other domains as well. For example, there are several tests for ToM in the natural language processing literature [8], [9]. With the recent surge of research surrounding large language models, there has also been a body of work that investigates whether large language models demonstrate theory of mind. Some research suggests that the ToM abilities of these models are sensitive to trivial alterations [10], indicating that learning robust theory of mind skills remains an open question for different ML domains.

III. Our solution: a computational benchmark

Our work is heavily inspired by the *competitive feeding paradigm*, an expressive task format in comparative psychology developed by Hare et al. [11] and substantially expanded and refined by Penn and Povinelli [12]. Competitive feeding focuses on nonverbal ToM skills related to perceiving and reasoning about an opponent's lines of sight, correct and incorrect beliefs about the world, and how the dynamics of these factors over a sequence of events.

To evaluate a variety of ToM skills in computational agents, we introduce Standoff, a novel computational benchmark. Standoff serves as a gridworld environment designed to gauge the extent to which models generalize abilities across a range of competitive feeding-inspired scenarios.[1] Standoff is built on minigrid [13] using Supersuit [14] and PettingZoo [15] for both supervised and reinforcement learning use. In this paper we evaluate only supervised learning models.

A. Competitive-feeding-inspired tasks

In Standoff tasks, the player's goal is to navigate an environment to reach a box containing the largest treat possible. Two differently sized treats are always available, and sometimes an opponent with the same goal is present. Treats are placed one-by-one, and then they might be shuffled around in specific patterns. Treats are briefly visible while placed or shuffled, but are otherwise hidden within one of five boxes. The opponent's vision is sometimes obscured by an opaque wall, causing it to be unaware of a placement or reshuffling. The challenge presented to the player is differentiating between the opponent's possible belief states in order to select which treat to approach: Is the opponent unaware of the big treat's location, or is the smaller treat the most desirable option?

In this setting, the opponent may harbor two kinds of unawareness: if the opponent never witnesses a treat of a given size being placed or swapped, it will be *uninformed*, or oblivious to the existence of such a treat. If the opponent witnesses a treat being placed or swapped, but then that treat is swapped again while the opponent's vision is obscured, then the opponent is *misinformed*, unaware of the treat's location. Note the difference in kind between these two sources of unawareness: while the former is a lack of knowledge, the latter involves a specific counterfactual belief, since the treat was previously observed at some location.

[1] Standoff can be accessed at https://github.com/aivaslab/standoff. Supplementary information about the hyperparameter tuning and training of our models can be found at: https://github.com/aivaslab/standoff/blob/main/icdl2024-standoff-appendix.md

attribute	group	range	condition	description
visible placements	main attributes	0-2		the number of visible placements
swaps		0-2		the number of swaps
visible swaps		0-swaps		the number of visible swaps
fsb	special cases	boolean	swaps > 0	first swap is btw. both treats
dsp		boolean	swaps > 0 & !fsb	delay 2nd placement after 1st swap
ssf		boolean	swaps == 2 & !dsp	2nd swap is to 1st swap location
first placement size	ordering attributes	boolean		which treat is placed first
first swap index		boolean	swaps > 0 & !fsb & !dsp	which treat is swapped first
uninf. placement		boolean	visible placements == 1	which placement is obscured
uninf. swap		boolean	swaps == 2 & visible swaps == 1	which swap is obscured

TABLE I
THE TEN ATTRIBUTES USED TO GENERATE STANDOFF TASKS. COMBINING THESE PRODUCES 296 UNIQUE EVENT ORDERINGS.

B. Measuring Generalization

Penn and Povinelli design competitive feeding as a curricular transfer learning task with three stages [12]. First, the subject is exposed to the idea that they may navigate to and select a treat from the two that are placed. In the second stage, a dominant opponent is introduced. This opponent always navigates to the larger treat, so the subject learns to maximize its reward by approaching the smaller treat. Once the subject has demonstrated success on the first two stages, it is evaluated on the third stage, which is split into eight or more tasks.

Stage-3 tasks are designed to control for superficial rules which might inform behavior. They specifically contrast scenarios, such as when the opponent has a counterfactual belief versus a true belief. In addition to varying visible placements and swaps, special cases allow for rich behavior comparison.

The validity of the test lies in the unlikelihood of solving all stage-3 tasks without 1) prior experience, and 2) some model of opponent beliefs. The opponent's behavior is determined by its belief state which, in turn, is shaped by numerous observable environment features, leading to a combinatorially explosive set of possibilities. Due to the set's variety, a high accuracy is incredibly unlikely unless the subject uses reasoning that generalizes well across opponent belief states.

C. The gridworld environment dynamics

Standoff tasks exist in an eight-by-eight gridworld setting. The player character is represented by an agent object which may be controlled by a computed policy or by a human player. The opponent character's behavior dynamically adjusts to different scenarios to navigate to the best treat reward that it has seen. Player observations are seven-by-seven birds'-eye views of the agent's surroundings, many-channel arrays that describe the properties of surrounding tiles: opacity, solidity, visibility, opponent presence, and the presence of either treat. Opaque tiles are used to block opponent vision as well as the player's vision of the opponent.

The player and opponent agents begin on opposing sides of the grid, facing inward. Between them lie five boxes. For several timesteps, both players are stationary while they observe a series of events: Two treats of different value are placed one-by-one in the boxes. Swaps might occur, during which the contents of two boxes discontinuously swap locations. The order of events varies. During each event, the contents of the relevant boxes are briefly visible. The opponent's vision of the boxes might be temporarily obscured by a wall, which is visible to the player. After these events take place, both characters are released. The boxes are positioned closer to the opponent than the player, indicating that the opponent will win, should the two compete for one box.

For supervised learning, the five box locations correspond to a five-category classification problem. The inputs are five-timestep sequences of seven-by-seven multi-channel observations of a stationary player, up until the moment of release. The correct output is the box containing the highest-value available (not being targeted by the opponent) treat.

IV. METHODS

A. Dataset generation

Rather than using a minimal number of test scenarios like real life competitive feeding trials, we enumerate a broad range of possibilities under the test's logic. Penn and Povinelli's tasks have counterparts which function as controls. E.g. the *removed uninformed* task can be directly compared with *removed informed* to see the effect of the opponent's informedness. We expect generalizing to a broader set of novel scenarios to be increasingly difficult, at least without explicit belief representation.

1) Events: Each task is a unique sequence of events, of four types: **placements** have new treats being placed in the environment. The treat is visible for 1 timestep before being hidden in a box. A treat is always placed in a random empty box, except for the delay second placement (*dsp*) condition, during which the second treat is placed in the previous treat's former location. There are always 2 placement events, guaranteeing a treat for the player. During a **swap** event, a box containing a treat is emptied and that treat is shown in a new location before it is hidden in a box. Normally, an empty location is selected, except for first swap both (*fsb*) condition, which swaps the two treats, and second swap to first (*ssf*) condition, which places the second swapped treat in the newly emptied box provided by the first swap. The other two events instantly **obscure** or **reveal** the room's contents to the opponent via placing or removing opaque tiles.

2) Parameters: The 3 major parameters, *visible placements, swaps,* and *visible swaps* form 18 sets of tasks. By adding the three special case boolean parameters, *fsb*, *dsp*, and *ssf*, we have 66 unique tasks. By adding up to four event parameters (first treat size, first swap index, visible placement, visible swap), we are left with 296 unique event orderings. Finally, each task may or may not have an opponent present.

3) Permutations: Since these event orderings require different amounts of timesteps before the release event, we lengthen the placement and swap events such that all sequences are equal length, five timesteps. Accounting for all valid (totalling five) event lengths, we are left with 880 lengthened event orderings. Finally, we must account for the varied locations at which treats are placed during placement and swap events; this combination leaves us with 47,040 unique *trials*, which we shall use as our datapoints.

	Tt	Tf	Tn	Ft	Ff	Fn	Nt	Nf	Nn
full-absent	a	a	a	a	a	a	a	a	a
full-present	p	p	p	p	p	p	p	p	p
full-both	a+p	a+p	a+p	a+p	a+p	a+p	a+p	a+p	a+p
single-Tt-a	a	-	-	-	-	-	-	-	-
single-Tt-p	p	-	-	-	-	-	-	-	-
single-Tf-a	-	a	-	-	-	-	-	-	-
single-Tf-p	-	p	-	-	-	-	-	-	-
...	...	...	...	...	...	...	...	...	...
contrast-Tt	a+p	a	a	a	a	a	a	a	a
contrast-Tf	a	a+p	a	a	a	a	a	a	a
...	...	...	...	...	...	...	...	...	...
homogeneous	a+p	-	-	-	a+p	-	-	-	a+p

TABLE II

THE SPECIFIC INFORMEDNESS REGIMES THAT COMPRISE EACH OF THE DATASETS USED IN THIS PAPER. EACH ROW REPRESENTS ONE DATASET AND EACH COLUMN REPRESENTS ONE INFORMEDNESS REGIME. A CHARACTER IN A CELL INDICATES REGIME INCLUSION WITH AN ABSENT OPPONENT (A), A PRESENT OPPONENT (P), OR BOTH THOSE CASES (A+P).

B. Informedness regimes

Competitive feeding assumes that animal subjects possess certain core knowledge priors, including object permanence, navigation skills, social hierarchy, preference for larger treats, and gaze comprehension. While stages 1 and 2 expose the subject to the task setup, they are not intended to *teach* these priors; instead they verify that the priors are used correctly.

To expose our players to all necessary priors during training without also letting them memorize situational strategies, we categorize our tasks by opponent presence and informedness. Our opponents may be informed about either of two treats, so we label these individually: an uppercase letter indicates informedness about the larger treat (T for true belief, F for false belief, or N for no belief), lowercase for the smaller (t, f, or n), and one character indicates the opponent's absence or presence (a or p). We classify tasks with absent opponents by a hypothetical informedness state, as though an opponent were present, to for ease of comparison. This classification ultimately produces 18 regimes, each labeled by a unique string, e.g. "Tf-p" refers to all tasks in which a present opponent is aware of the large treat's location but harbors a counterfactual belief about the small treat's location.

C. Our learning target: optimal treat selections

If no opponent is present, the optimal policy is to select the larger treat. Otherwise, the correct selection depends on the opponent's beliefs about treat locations. If the opponent is correctly informed about the larger treat's location, the player should go for the smaller treat. Otherwise, the optimal policy depends on the opponent's beliefs. Sometimes, the opponent arrives at a counterfactual belief about one treat location coinciding with the true location of the other. In such an event, the player's optimal choice could be the smaller treat even if the opponent is unaware of the large treat's location. If an opponent is fully uninformed about both treats, we define its policy to be selecting the closest box. This choice ensures that the opponent's policy is universally predictable to an observer. Due to special cases like these, most informedness regimes do not have a single optimal treat size.

V. EXPERIMENT 1: FULL-SCOPE TRAINING

Before evaluating generalization from limited sets of tasks, we will first establish that our selected model, a convolutional LSTM architecture without any explicit ToM-oriented architectures or training, is capable of succeeding at the overall challenge presented by Standoff. In this experiment, we train models on three datasets, containing all tasks with an opponent absent (*full-absent*), an opponent present (*full-present*), and the union of those two conditions (*full-both*). *Full-absent* will serve as a baseline for comparison with future models: this dataset exposes a model to all relevant elements of the environment except for the opponent. The next dataset, *full-present*, imitates stage 3 of competitive feeding by containing the full variety of scenarios with an opponent present. Models fit to this dataset may find success by memorizing situational rules as opposed to generalizing strategies. The third dataset, *full-both*, includes all trials from both the previous two. As with *full-present*, we have no reason to expect models trained on *full-both* to learn generalizing rules, but the dataset is useful for ensuring that models do not underfit.

A. Methods

We train our convolutional LSTM models with supervised learning. The specific architecture and hyperparameter tuning process are detailed in the online supplementary material. All reported data on model results in this paper use mean results from three models, separately trained with random initializations and batches for 5000 batches of 256 datapoints each. We train for a constant number of batches so as to ensure fair comparisons between datasets of different sizes, though we do not find a strong connection between dataset size and accuracy in single-regime training sessions.

B. Metrics

1) Accuracy: The key metric for performance is accuracy, i.e. whether the rational option that accounts for the presence of a dominant opponent's beliefs is selected. To delve deeper into a model's capabilities, we may slice accuracy by factors like the opponent informedness, *fsb*, *ssf*, or *dsb*.

C. Results

All models consistently reach high accuracies ($\geq$97%) on their own training regimes, with low inter-model standard deviations ($\leq$2%). When any opponent-present (_-p) and opponent-absent (_-a) models are tested on novel regimes, they score well when the optimal policies of those regimes match their absent/present counterparts, and poorly otherwise.

We also examine the qualitative effect of individual tasks and task parameters. The *full-present* models perform best overall on tasks with zero visible swaps, and poorly with two visible swaps. Tasks with zero visible swaps include those with zero total swaps, so it is unsurprising that the *full-absent* models follow a similar pattern, assuming swaps

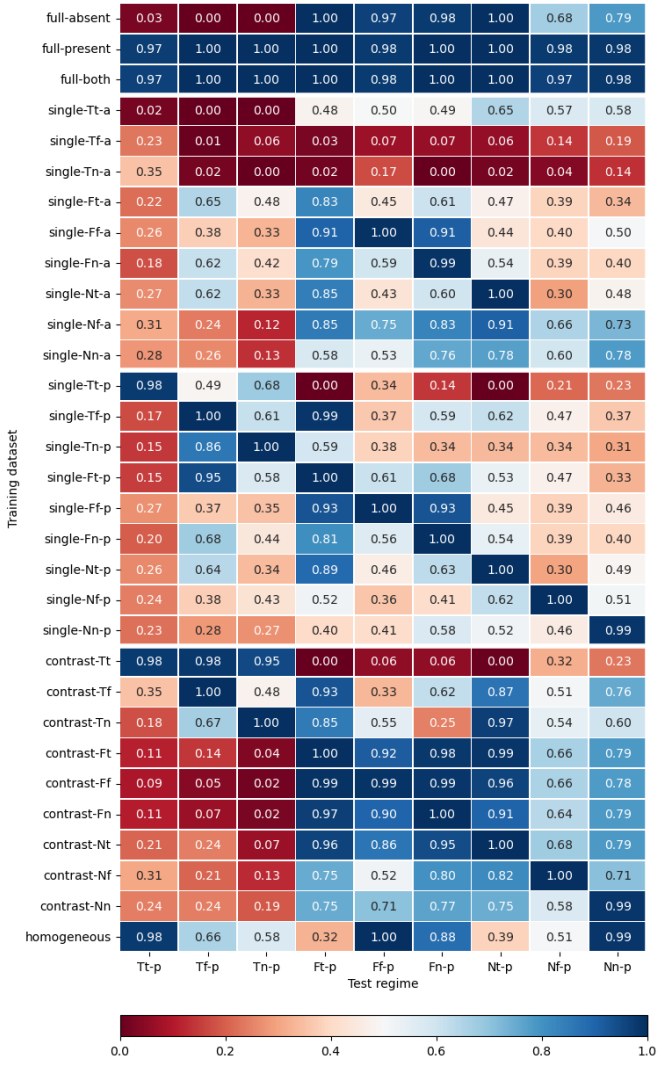

Fig. 2. Mean accuracy results of convolutional LSTM models trained on each of our datasets (rows) and tested on the nine informedness regimes with an opponent present (columns). Each reported value is the mean accuracy of three models trained with different random initializations and minibatches. The top three rows display the results of Experiment 1, the next eighteen display Experiment 2, and the next ten display Experiment 3.

are universal sources of confusion. Models trained on *full-both*, whose highest-scoring tasks are those with zero swaps, score highly on tasks with the *dsp* special case, in which the second placement takes place after the first swap. Nine of the ten poorest-scoring individual tasks of models trained on *full-absent* include the special case *fsb*, where the first swap is between both treats. That pattern is repeated for both the *full-present* and *full-both* models, although their weakest tasks simultaneously include *ssf*; tasks with both *fsb* and *ssf* have the two treats swap locations with each other twice, returning to their initial positions.

VI. Experiment 2: Single-regime training

Next, we examine how models tend to generalize when trained on narrow subsets of the dataset. In particular, we are interested in the extent to which models' performance transfers from any one informedness regime to the next.

A. Methods

We train models on each of our eighteen informedness regimes individually. In addition to the performance metrics established in Experiment 1, we note that transfer across different informedness regimes could be asymmetric. The asymmetric transfer between two regimes is calculated by measuring the difference in mean accuracy when training on one and testing on the other, and vice versa.

B. Results

All models converge to high accuracies on the regimes on which they were trained. Surprisingly, most models trained with an absent opponent do not perform well on other regimes with an absent opponent. Across all evaluation regimes, mean inter-model standard deviation of is 2.65%, and the standard deviation of inter-model standard deviations is 4.00%. The models trained on *single-Tt-p* display an outlying standard deviation on the Tf-p test regime at 24%.

Chimpanzees given the similar competitive feeding test [16] produce results indicating that they are able to distinguish between informed, uninformed, and misinformed opponents, but only regarding one treat. When an opponent is informed about one treat but uninformed or misinformed about the other, the subjects become unable to distinguish informedness states.

If our own models were to follow a similar pattern, we might observe differences in their ability to transfer knowledge across regimes of different informedness homogeneity, meaning whether the opponent's informedness states about the two treats are the same or different. Specifically, we posit that models trained on heterogeneous regimes (Tf, Tn, Ft, Fn, Nt, Nf) exhibit superior performance when evaluated on homogeneous regimes (Tt, Ff, Nn), compared to the reverse. We do not find such a pattern in our models' performance. When transferring from homogeneous to heterogeneous regimes with an opponent present, our models are overall *more* accurate by a mean of 10.2% over all inter-homogeneity regime pairs, with considerable variation (standard deviation 21.3%).

The most notable asymmetric transfer of individual regime pairs emerges from the Nf-a regime, which exhibits a distinct pattern of strong performance when models trained under it are transferred to other regimes, particularly Ft-a, Fn-a, Ft-p, and Fn-p. This regime has 1740 unique trials, which is lower than the mean of 2613, but its tasks tend to have more swaps than average. Nf-p has an identical set of swaps, so it is unclear why its transfer to other regimes differs so markedly.

VII. Experiment 3: Mixed-regime training

Our ultimate goal is to train models which display generalization on tasks with opponents with various mental states, so we must minimize the amount of exposure our models have to those tasks. For a *tabula rasa* player to understand the competitive feeding format, however, it must be exposed to trials with an opponent present. In this experiment, we

examine the effect of exposure to mixtures of regimes with present and absent opponents.

A. Methods

In this experiment, we use nine *contrastive* datasets. These datasets each include all tasks with an opponent absent, plus tasks from exactly one of the nine informedness regimes with an opponent present. Each contrastive dataset is labeled by the opponent-bearing informedness regime it contains; e.g. the contrastive dataset containing all of *full-absent* plus Tt-p is labeled *contrast-Tt*. Since the *full-absent* regimes are similar to stage 1 of competitive feeding, and stage 2 features a fully informed opponent, *contrast-Tt* is analogous to the competitive feeding exposure phase. Additionally, because such datasets include examples of only one type of opponent informedness, we introduce the *homogeneous* dataset, which includes all homogeneous regimes, with an opponent both present and absent. This dataset allows us to more directly test for generalization across informedness homogeneity.

B. Results

Similar to the results from Experiment 2, models trained on each contrastive dataset—and *homogeneous*—perform well on regimes included in their training. Accuracy generally follows the pattern of near perfection across regimes with similar optimal treat sizes, and is low otherwise. Models trained on the Tt regime, which we point out for its similarity to stages 1 and 2 of standard competitive feeding, do not generalize well to any regimes featuring opponents with uninformed or misinformed belief states. The Nf regime is a notable outlier, as all models trained on other regimes struggle to generalize effectively to it. In this regime, the opponent selects a box which does not contain the smaller treat, but might contain the larger *by accident*. The opponent harbors similar beliefs in the Fn regime, but in those trials that fact is irrelevant to the player's optimal strategy: to always select the larger treat. The amount of accidentally-best selections in Nf totals 34.5% of all datapoints, compared to 20% or less for all other regimes.

VIII. Discussion and future work

Our convolutional LSTM model's inability to generalize suggests that it tends to learn shallow mechanical patterns rather than how to make or use accurate inferences. The asymmetry in transfer between different regimes indicates the potential for building and leveraging training curricula. Finally, these baseline experiments present us with a set of especially challenging test conditions to highlight for rapid evaluation of ToM skills, where we find that the most difficult tests involve reasoning about uncertain mental states.

In future experiments, the variety provided by the competitive feeding test will allow for a much more fine grained exploration of models' capabilities. One benefit of studying machine learning models, as explored by [6], is the ability to directly probe their internal activations for representations of beliefs. Additionally, we might consider a metric which quantifies internal activation or external policy changes of a model with respect to changes in opponent belief states.Varying certain environment parameters allows for expression of a wider range of ToM and ToM-related skills [3]. In this paper, we examined an agent's ability to reason about another agent's beliefs only as they pertained to its goal-oriented behavior. While competitive feeding tests for attribution of gaze and beliefs, modifications might include shared rewards, different preferences, or sources of environmental and social uncertainty. By reversing the roles of player and opponent, we could test for models' ability to extrapolate knowledge from first- to third-person experiences.

References

[1] M. Tomasello, *The cultural origins of human cognition*. Harvard university press, 2009.

[2] C. Heyes, "Animal mindreading: what's the problem?" *Psychonomic bulletin & review*, vol. 22, pp. 313–327, 2015.

[3] 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.

[4] K. Gandhi, G. Stojnic, B. M. Lake, and M. R. Dillon, "Baby intuitions benchmark (bib): Discerning the goals, preferences, and actions of others," *Advances in neural information processing systems*, vol. 34, pp. 9963–9976, 2021.

[5] T. Shu, A. Bhandwaldar, C. Gan, K. Smith, S. Liu, D. Gutfreund, E. Spelke, J. Tenenbaum, and T. Ullman, "Agent: A benchmark for core psychological reasoning," in *International Conference on Machine Learning*. PMLR, 2021, pp. 9614–9625.

[6] N. Rabinowitz, F. Perbet, F. Song, C. Zhang, S. A. Eslami, and M. Botvinick, "Machine theory of mind," in *International conference on machine learning*. PMLR, 2018, pp. 4218–4227.

[7] M. Sclar, G. Neubig, and Y. Bisk, "Symmetric machine theory of mind," in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 17–23 Jul 2022, pp. 19 450–19 466. [Online]. Available: https://proceedings.mlr.press/v162/sclar22a.html

[8] E. Grant, A. Nematzadeh, and T. L. Griffiths, "How can memory-augmented neural networks pass a false-belief task?" in *CogSci*, 2017.

[9] M. Le, Y.-L. Boureau, and M. 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)*, 2019, pp. 5872–5877.

[10] T. Ullman, "Large language models fail on trivial alterations to theory-of-mind tasks," *arXiv preprint arXiv:2302.08399*, 2023.

[11] B. Hare, J. Call, B. Agnetta, and M. Tomasello, "Chimpanzees know what conspecifics do and do not see," *Animal Behaviour*, vol. 59, no. 4, pp. 771–785, 2000.

[12] D. C. Penn and D. 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*, vol. 362, no. 1480, pp. 731–744, 2007.

[13] M. Chevalier-Boisvert, B. Dai, M. Towers, R. de Lazcano, L. Willems, S. Lahlou, S. Pal, P. S. Castro, and J. Terry, "Minigrid & miniworld: Modular & customizable reinforcement learning environments for goal-oriented tasks," *CoRR*, vol. abs/2306.13831, 2023.

[14] J. K. Terry, B. Black, and A. Hari, "Supersuit: Simple microwrappers for reinforcement learning environments," *arXiv preprint arXiv:2008.08932*, 2020.

[15] J. Terry, B. Black, N. Grammel, M. Jayakumar, A. Hari, R. Sullivan, L. S. Santos, C. Dieffendahl, C. Horsch, R. Perez-Vicente *et al.*, "Pettingzoo: Gym for multi-agent reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 15 032–15 043, 2021.

[16] B. Hare, J. Call, and M. Tomasello, "Do chimpanzees know what conspecifics know?" *Animal behaviour*, vol. 61, no. 1, pp. 139–151, 2001.