Towards a Theory of Mind for Artificial Intelligence Agents

Jory Schossau[1,2] and Arend Hintze[2,3]

[1]Michigan State University, Department of Computer Science and Engineering
[2]Michigan State University, BEACON Center for the Study of Evolution in Action
[3]Dalarna University, Department of MicroData Analytics (ahz@du.se)

Abstract

In the growing fervor around artificial intelligence (A.I.) old questions have resurfaced regarding its potential to achieve human-like intelligence and consciousness. A proposed path toward human-level cognition involves the development of representations in A.I. systems. This paper focuses on establishing the methods and metrics necessary toward developing and studying an A.I. that can "impute the mental states of others" (Theory of Mind). Here we examine existing psychological and robotic research on this subject, then propose an information-theoretic metric to quantify the extent to which agents have a Theory of Mind. The metric is applied to agents trained using a genetic algorithm, demonstrating that an agent-specific Theory of Mind can be achieved without the need for a general Theory of Mind. This framework lays the operational groundwork for development toward more general Theory of Mind in artificial intelligence.

Introduction

Recent advances in Artificial Intelligence have shown both impressive growth in capabilities and public attention. Even researchers are asking if Large Language Model tools — such as chatGPT — would pass a Turing test (Elkins and Chun, 2020) or might even be conscious (Goldman, 2023). While these advancements are impressive, it quickly becomes clear that even our most advanced systems are still not advanced enough. The question then arises "How can we measure progress toward human-like intelligence, such as that required to pass a thorough Turing Test?" This requires operationalizing the steps toward human-like intelligence to indicate which milestones are conquered and which are next.

Obviously, this question has been discussed numerous times before, but here we will focus on an idea put forward by A. Hintze (Hintze, 2022), based on a panel discussion with C. Adami at the 2014 ALIFE conference, New York. In summary, the path toward human-level cognition follows four steps defined by the capability of a system to form representations. A representation in this context is the information a machine has about its environment or ultimately about itself.

Reactive Machines: The first milestone toward intelligence begins with purely reactive input/output (I/O) machines operating based solely on current input — such as an artificial neural network performing a single-shot classification. While prior training allows such an ANN to perform optimally under the right circumstances, the network itself does not retain information about earlier inputs, and is thus without an internal state or representations about their environment[1]. These machines might already be capable of performing complex tasks. Controlling an inverted pendulum (Barto et al., 1983) is such a situation. As long as weight of the dolly and pendulum, length and angle of the pendulum, and direction of the swing are known, the appropriate movement to keep the pendulum upright can be computed based on current inputs alone.

Limited Memory Machines In order to better respond to more complex and dynamic environments, machines require information about correlations in those environments. Here, machines belonging to this second category are different in this exact manner, as they possess information not only of the present but also of the past. More importantly, this past information is not provided at each instant from an exterior source, but retained within for future use. Every form of recurrent neural network or finite state machine can have this feature, and consequently these machines are already in use today. We can even quantify and pinpoint their internal representations — that is, the information these machines retain about their environment (Marstaller et al., 2013; Hintze et al., 2018; Bohm et al., 2022b; Hintze and Adami, 2023).

Theory of Mind Memory then provides the foundation for responding well in more complicated environments. These environments are made complicated because events have unseen causative factors for which only correlations between cues and events are observable across space and time. This makes such environments impossible for purely reactive machines to achieve good performance, if this kind of

[1]The features detected by convolutional kernels in this context, while sometimes referred to as representations, are not the same kind of representations mentioned here, see *Machines with limited memory*

prediction matters for their success. The most complicated environments contain events with many interacting unseen causative factors. We colloquially call such mechanisms "black boxes" due to the difficulty of inferring their internal processes. Anything with memory — such as an A.I. or a human — is a black box. When agents begin interacting with each other or humans, then they will be faced with the difficulty of dealing with inferring cause and effect relationships involving these black boxes. Obviously, information about the internal or mental states of others, as well as the whole idea that the environment can contain other agents with internal states, goes beyond simple environmental states and their internalized representations. In a sense, the ability to gain information about the internal or mental states of others helps to make these black boxes more transparent. After pure reactionary agents agents having a *Theory of Mind* thus seems to consequently be the next step toward human-level artificial intelligence. Here, we will investigate the question how agents that have Theory of Mind could be created, and how to quantify and pinpoint that information.

Consciousness or Self-Representations As outlined in (Hintze, 2022), we also assume that attaining human-level intelligence requires consciousness. However, consciousness is difficult to define, let alone quantify (Tononi and Koch, 2015). While this is a highly debated topic, some progress has been made toward measuring neuro-correlates as proxies for consciousness in computational agents (Edlund et al., 2011; Albantakis et al., 2014). One definition of consciousness is self-representation. Quantifying self-representation requires measuring a particular information — in this case not about the environment or other agents, but having information about your own (Theory of) Mind: *Cogito, ergo sum* (Descartes, 1901).

These four milestones outline a possible path for A.I. development, and they also anchor this advancement around the common theme of an A.I. agent with quantifiable representations. It starts with no representations, goes toward having representations about its environment, then about the internal states of others in the environment (Theory of Mind), to ultimately arrive at self-representations (a.k.a. consciousness). It remains open if the third step about having a Theory of Mind is a necessary stepping stone toward consciousness, or if consciousness might emerge before having a Theory of Mind about others. Observe, that our approach seeks to build and measure a computational brain that has Theory of Mind, and could be independent from mechanisms found in humans (Baker et al., 2009; Margolis et al., 2012).

If A.I. development were to follow these milestones, then the next research hurdle would seem to be the idea of agents having a Theory of Mind. Psychological research in the subject increased after Premack and Woodruff in 1978 asked if Chimpanzees have a Theory of Mind (Premack and Woodruff, 1978)? They defined Theory of Mind as having the ability to impute mental states of self or others. This is related to the Turing test, wherein the tester tries to impute whether an A.I. does or does not have the same internal state repertoire as that of a human consciousness (Horst, 2011). Since then, different psychological descriptions of the phenomenon have been proposed (Leslie et al., 2004; Ahmed and Stephen Miller, 2011; Singer and Tusche, 2014) (amongst others) that assume a range of psychological mechanisms based around quantification methods of brain imaging to illuminate the responsible areas (Gallagher and Frith, 2003). This includes different ways to test for the presence of Theory of Mind in test subjects, such as false-belief tasks (Happé, 1994), reading the mind in the eyes tests (Baron-Cohen et al., 2001), interpersonal perception tasks (Stone et al., 1998), or interviews and observations. Similarly, roboticists tried to build systems with a Theory of Mind (Scassellati, 2001; Kuniyoshi et al., 2004). Lastly, there are approaches using deep learning to optimize networks to infer internal states of agents (Rabinowitz et al., 2018). This deep learning can be extended to be symmetric when the agents can observe each other (Sclar et al., 2022) and are trained to form a Theory of Mind in that way. However, a *quantitative* and *mathematical* definition is still missing; a gap we begin to remedy here.

Some psychological tests result in a score and so they could be considered quantitative, but they can only be used on already conscious human beings and as such they heavily rely on introspection. These requirements make such psychological tests ill-suited for experiments on computational A.I. systems. Similarly, having an agent that indicates that another agent has Theory of Mind (Rabinowitz et al., 2018; Albrecht et al., 2020; Sclar et al., 2022) still does not allow us to measure the quantity of that information, and also does not allow precise location of Theory of Mind within the black box.

It would be presumptuous to assume that an A.I. possesses similar psychological or cognitive mechanisms to those of humans, although that would make measuring Theory of Mind far easier. Instead, the measurement and definition need to be purely functional and quantifiable. We begin by making a distinction between two types of information in Theory of Mind. The first information is that an agent is capable of being in one of a set of states. The other information is about which specific state. In other words, knowing that something has internal states is different from knowing its exact internal state. For example, understanding whether or not a tiger you just met in the wild is capable of being hungry and that its next actions would depend on that, is different from knowing that it is at present hungry and it is acting based on that. The former is a more generic insight that can be applied to almost any animate or inanimate object. The latter is knowledge about the other being in a specific state. Rabinowitz et al. called this general Theory of Mind, and agent-specific Theory of Mind, respectively (Ra-

binowitz et al., 2018). We believe that knowing that something *has* internal states is more closely related to animacy than knowing its *specific* internal state. Therefore, we focus on the agent-*specific* Theory of Mind as the ability to know the *specific* internal (mental) state of another agent[2].

We even question if a general Theory of Mind mechanism is necessary for an agent-specific Theory of Mind. This will become clear as we investigate Theory of Mind generalization. We cannot be certain that we are capable of imputing internal states to every possible entity that has them. Consequently, our general Theory of Mind mechanism is likely limited to our horizon of experience, or the evolutionary niche within which we have adapted. If that is the case, then we would not possess a complete general Theory of Mind but have instead a Theory of Mind specific to some set of agents (plural). Furthermore, this Theory of Mind would be specific to what was evolutionarily useful to impute internal states about, not necessarily all possible entities and all possible states. Curiously, the phenomenon of pareidolia — seeing faces in everyday objects — also suggests that instead of having a properly functioning general Theory of Mind, we humans have over-generalized in this regard and falsely impute animacy to things that do not have internal states. For example, we can imagine happening upon and describing a "sad-looking car," even though a car is incapable of feeling sad.

Here we will use a genetic algorithm to create agents able to have internal states (representations) about the internal states of other agents. We further devise a simple information-theoretic metric to quantify how much agents know. This will also enable isolating which hidden nodes carry the information for Theory of Mind. We will then show that the Theory of Mind that evolved is agent-specific, and consequently does not require a general Theory of Mind. Finally, we will discuss how our simplifications agree or disagree with our intuitions, followed by possible improvements to the information-theoretic metric.

Environment Requiring Theory of Mind: The environment for this experiment requires agents to have internal states, communicate these states, act upon them, and act upon the states of the other agent (see Figure 1). There are five rooms arranged as on a compass rose: one room in each of the cardinal directions, and a center room connecting through them four doorways one to each of the cardinal rooms. Each room is of size 3×3. One agent is placed in the North and South rooms, both facing toward the middle. Agents can turn left, right, move forward, or do nothing. At the same time, they have two binary outputs to communicate, which can be imagined as beeping in three different frequencies, plus no beeping. Beeping can only be heard by agents when they are both in the central room at the same time. This 2-Dimensional environment is quantized into discrete tiles (grid locations). Agents receive information about the tile in front of them and also room-specific information while in the North or South rooms: which color lever they must later press (red=[0,1], green=[1,0], or blue=[1,1], encoded as 2 binary inputs). Once they leave their initial rooms, that color cue ceases. Once an agent passes through any door, that door closes until the session is reset. This prevents agents from observing the other agent's color cue. Or, if one agent walks over to the other, they would be locked in the same room, incapable of finishing the task. Agents can take up to 500 time steps to complete the task, which allows sufficient time in the central room to move and communicate with each other by beeping or observing each other's actions. Toward the end of the session, each agent should enter one of the side rooms, with the agent from North going into the West room, the other into the East. Here agents then find three levers they can choose, each colored (red, green, or blue signaled in the same way using 2 binary inputs as in the North or South room). Agents then should choose the lever with the color the *other* agent experienced in its own initial room. The agents also must beep the same color frequency they encountered in *their* own initial room.

Successful completion of this task requires that agents accomplish three things: Report something they encountered earlier (color of their own room), communicate this color to the other agent, report the color that the other agent communicated. In order to do this, information about the color of the initial room flows from the environment into each agent separately (see Figure 2 perception). Then agents can share the same space and communicate (see Figure 2 communication). We often assume that Theory of Mind is purely based on observation, such as in the Tiger example we conclude from its aggressive actions that the tiger may be hungry. Fortunately, this form of indirect communication is not necessary for Theory of Mind. A much more direct message of "I am hungry" could be sufficient. In our example both agents are allowed to directly state the color of their original room, though they could always make this communication more complicated. Once agents are in their final room they should report both internal states about rooms: their own room, and the room of the other agent (see Figure 2 reporting).

Agents are controlled by a Recurrent Neural Network (RNN) or by a Markov Brain (MB) (Hintze et al., 2017), which is a form of neural network similar to Cartesian genetic programming (Miller and Harding, 2008) but uses deterministic, probabilistic, and other mathematical operations for computation, while also allowing connections to be altered by mutations between generations. To evaluate the performance of each agent, every agent is cloned (an identical copy is made) such that one copy is the North agent and the other is the South agent. This pair of agents is refreshed (hidden states cleared) each time the session starts anew with the agents in their starting rooms. At every generation of evolution they are evaluated twice for each of the

[2]Knowing that something is alive (animacy) is different from knowing that something has intentions for example.

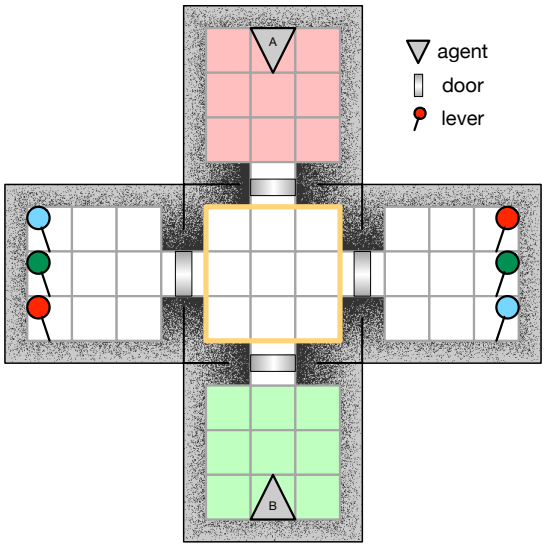

Figure 1: Illustration of the environment. Agents begin in the North and South rooms, in which they experience one of three colors (red, green, or blue). Once they pass through a door, that door closes, forcing information about the color to be carried by the agent, if at all. Once both agents are in the central room (orange) they can beep and hear the other agent beep, which is facilitated by two binary inputs and outputs. Agents can then progress to the side rooms, ideally North to West, and South to East. There they can choose one of three levers by colliding with it. The color they choose must correspond to the color the other agent experienced in its own initial room. While they choose levers, they must also beep the same color code (2 binary numbers) they experienced in their own initial room. The agent is removed from the environment once it selects a lever. Agents have 500 updates to complete their task.

9 possible color combinations. Choosing the lever indicating the color the other agent experienced in its original room correctly results in a reward of 0.4 points. While choosing any lever, correctly signaling the color for their own original room results in a reward of 0.1 points. Consequently, agents can maximally receive 18.0 points for perfect performance each generation. Fitness is then awarded by raising 2.0 to the power of that score turning the reward into an exponential fitness function. Based on this fitness score agents are optimized using a genetic algorithm with roulette wheel selection (Goldberg, 2013).

Quantifying Theory of Mind: We will quantify the amount of Theory of Mind information that one agent (source) has about the other (target) by measuring shared entropy between the brain states of both agents. Because both agents have states about their own goals, which are distinct from the states representing the other agent's goals, this metric becomes a directional asymmetric measure (see Figure 2). This can be done using Equation 1, where the joint entropy between brain states of the source agent (B_s) and

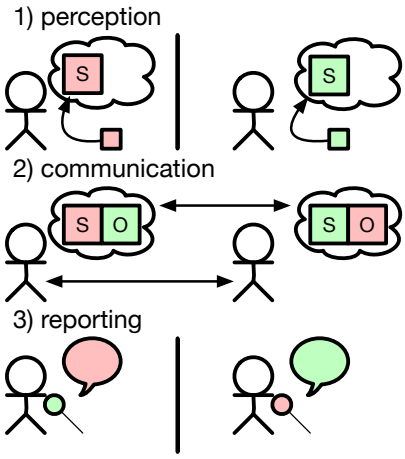

Figure 2: Illustration of the information flow between both agents. At first each agent must perceive (1) and form a mental representation (S, for self) of the room color they start with. Then, when in the central room, they should communicate (2) such that they can form a representation about the other agent's mental state (O). Finally, they should report (3) their mental state about the color of their own original room (S) and the color of the other agent's original room (O).

the brain states of the target agent (B_t) are quantified as:

$$I(B_s, B_t) = H(B_s) + H(B_t) - H(B_s, B_t) \qquad (1)$$

Here, one of three possible colors for the room each agent starts in, as well as one of three possible colors the other agent started in should at some point be known to the agent. Colors are binary encoded: "no color" being two zeros ([0,0]), "red" being a [0,1], "green" a [1,0], and "blue" a [1,1]. A human engineer might dedicate four of the binary hidden states (in the case of a Markov Brain) to storing this information. In an RNN, one might use a similar arrangement even though states are continuous. However, these agents are optimized using a genetic algorithm, which is opportunistic in the sense that any mutation that provides computational advantage will likely be retained in a sort of cumulative discovery. Consequently, the resulting computational machinery will be less organized than that of a human engineer, and in all likelihood will be an epistemically opaque solution (Marstaller et al., 2013). Furthermore, we would assume that the fully connective nature of an RNN operating with aggregation and threshold functions would result in equally opaque solutions. This is supported by other findings showing that RNNs tend to result in wider distributed representations (Hintze et al., 2018; Hintze and Adami, 2022; Bohm et al., 2022b). Consequently, our metric must account for multiple and widely distributed representations in both agents.

Another problem comes from spurious hidden states between the agents. Imagine that both agents have a hidden state that just toggles back and forth between 0 and 1, chang-

ing with each computational time step. If we would measure the information between those mental states (see Figure 3 blue area), we would find a high shared entropy, which would be purely coincidental. Thus, applying Equation 1 on the shared brain states of both agents would likely result in an incorrect measurement — it would only be about joint brain states, not about mental states of the target agent pertaining to actual world states. In other words, it is the *grounding*[3] with the environment that makes a state meaningful, and the measure must include this information lest it measure spurious correlations with no bearing on the agents' reality. Consequently, the brain states that the source brain has about the target's states need to be conditioned on actual environmental states. In other words, Theory of Mind of the source agent is information not only about mental states of the target agent. Instead, it is about specific mental states of the target agent — in this case the color of the first room. This can be expressed as an information-theoretic Venn diagram (see Figure 3) where the brain states of the agent (B_s and B_t) are shown together with the environmental states that pertain to the target agent (E_t).

The Theory of Mind information ($I(B_s; B_t; E_t)$) can now be calculated as the joint entropy between all three of these random variables (see Equation 2):

$$\begin{aligned} I(B_s; B_t; E_t) &= H(B_s) + H(B_t) + H(E_t) \\ &- H(B_s, B_t) - H(B_s B_t) \\ &- H(B_t, E_t) + H(B_s, B_t, E_t) \end{aligned} \quad (2)$$

When dealing with perfectly-trained agents in this environment, the mental states that lead an agent to utter the right color at the end are identical to the states of the environment (the agent utters the color it saw at the start). Thus, the information an agent has about the environment that isn't shared with the other agent (indigo section in Figure 3) is zero. This allows simplifying the Theory of Mind information to computing the joint entropy between the environmental states of the target agent (E_t) and the brain states of the source agent (R_s) using equation 1. This joint entropy is what we call the Theory of Mind information — or I_{ToM}.

If the environment is more complex or agents are erroneous about environmental states, a more complex approach to identify the information each agent has about the environment (Marstaller et al., 2013) must be taken first. Secondly, information that both agents can obtain from the environment, without the need of communication (indigo section in Figure 3) also must be identified properly. Thus, our simplification only works under these specific experimental conditions.

[3]Grounding is "...the processes by which an agent relates beliefs to external physical objects. Agents use grounding processes to construct models of, predict, and react to, their external environment." Roy (2005)

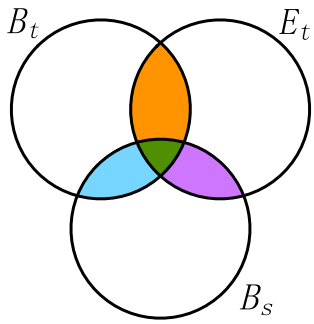

Figure 3: Information-Theoretic Venn Diagram. Each circle represents a random variable, with B_s being the one for the brain states of the source agent, B_t are the brain states of the target agent, and E_t being the environmental states that the target agent may have knowledge about. E_t are also the states that the source agent could impute about the target agent. The joint orange and green area is the amount of information the target agent has about the environment. Consequently, the green area is the Theory of Mind information that the source agent has about the target agent that is also grounded in the target agent's mental representations of the environment (see Equation 2). The blue area is the set of representations shared between both agents that do not pertain to the environment. Ostensibly, these states matter for communication, but because they are not grounded in the environment then they may simply be nonexistent spurious correlations (see Discussion). Lastly, the indigo area must be empty with no information because the source agent has no ability to assess the world states experienced by the target agent in this experiment. The green area is calculated by measuring the joint entropy between the environmental states relevant to the target agent E_t and the mental states of the source agent B_s (see Equation 1.

In other words: By experimental design, the entropy of the indigo section in Figure 3 is 0.0. Thus the joint entropy between E_t and B_s (green and indigo section in Figure 3) is identical to $I_{ToM} = I(B_s; B_t; E_t)$ (only green section).

A peculiar thing happens when the random variable E_t is not specific to the target agent, but instead represents general environmental states (E). In such a scenario, any information about the environment that both agents perceive would factor into the measure, leading to a circular phenomenon when both agents have knowledge about each other's mental states. This becomes problematic since we would be measuring the information between the source agent's mental states and the target agent's mental states. To avoid this, we need to condition the information based on the relevant environmental states associated with the target agent (E_t). An alternative could be a conditioning on the source agent's representations, resulting in higher dimensional Venn diagrams. Those could be resolved in the future, allowing more complex situations to be quantified appropriately.

The simplification used here will not work for those more

complicated cases and serves only to illustrate an initial step toward properly quantifying Theory of Mind. This approach will fail catastrophically for a situation in which, for example, one agent makes up a random number, communicates that, and then the other agent has that information. This situation fails because no environmental state exists. To detour this problem we must have well-defined environmental states that can also work with an agent-based behavior.

Identifying Theory of Mind Representation Location: Computing the information I_{ToM} shows whether or not the information exists and by how much, but we could additionally identify *where* the information exists among the agent's hidden nodes. This is similar to using fMRI to locate in a human brain where representations are forming based on correlates for neural activity. Here we leverage a simple information-theoretic principle for an algorithm that discovers the localization of information. We consider all hidden nodes to be a set of nodes for which I_{ToM} can be computed (similar to the SHAP method, which does this for feature correlations (Lundberg and Lee, 2017)). We determine the amount of information that each node contributes to the overall I_{ToM} measure. By removing the nodes in order of increasing information loss, we can determine which nodes are crucial to the transmission of information between the agents. This method has been shown to successfully identify highly-informative sets of nodes (Hintze and Adami, 2023).

Verification of Identified Mental States: The previous method identifies a sequence of nodes of increasing I_{ToM} loss. However, it may not work as intended in that the identified nodes may not carry information about the other agent's original room color. To test this, we perform a perturbation experiment. We intercede for the perturbation in the moment after communication, before the agent must act appropriately. For each agent, the sequence of identified nodes determines a sequence of node sets starting with the complete set and ending with an empty set. Every time an agent enters the last room and the door closes behind it, we then replace the values for nodes specified in each set with random values from a uniform distribution ($[-1.0, 1.0]$).

Suppose we have a small agent with only three hidden nodes numbered 0, 1, and 2. By applying the previously described method we determine that the nodes are arranged in the order of increasing importance as 2, 1, and 0. This ordering gives us a sequence of node sets, i.e., $\{2, 1, 0\}$, $\{1, 0\}$, and $\{0\}$. Based on this ordering, we assume that node 0 contains the most information about the Theory of Mind state of the other agent. If we noise the complete set of nodes $\{2, 1, 0\}$, then we expect the agent's performance to suffer significantly. It may not even be able to recall the original color of its own room, let alone that of the other agent. On the other hand, assuming the final set $\{0\}$ perfectly contains all the Theory of Mind information, then adding noise to only $\{0\}$ ideally destroys the Theory of Mind information while leaving the other information intact. In this way, we can apply noise to all the node sets between $\{2, 1, 0\}$ and $\{0\}$ to arrive at identifying the nodes that are critical for destroying Theory of Mind while still allowing other knowledge to remain. This test was performed on all optimal agents controlled by a Markov Brain, and all perfect or near-perfect RNN-controlled agents. Nodes for each set were noised testing for each possible color combination twice.

Results

The optimization of agents using a genetic algorithm has proven to be challenging, in this case for agents controlled by a Markov Brain or controlled by a Recurrent Neural Network. We conducted 500 replicate experiments, allowing them to run for the maximum time available to us of 7 days. Out of these experiments, we observed that only 31 Markov Brains evolved to become perfect-performing agents requiring 1 million generations, whereas only one RNN reached perfect performance after 0.5 million generations. Additionally, we observed that eight RNNs performed near-perfectly [4] (see Figure 4).

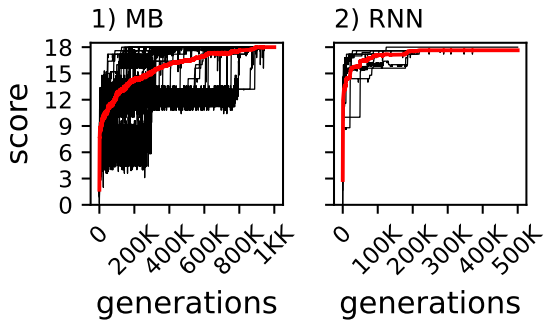

Figure 4: Evolution of fitness for perfect or near-perfect agents. 500 independent replicate experiments were started and terminated upon reaching the walltime of 7 days. This resulted in 31 optimal performing agents controlled by a Markov Brain, their fitness is shown in the top panel 1) as black lines, and in red the average fitness. RNNs are computationally more expensive to evaluate, so experiments only ran up to 500K generations. One RNN reached perfect performance, and 8 others reached near-perfect performance. Their performance is shown as black lines in the bottom panel 2), and in red the average performance of those 9.

To illustrate that the greedy algorithm is always removing the least information-carrying node, we computed the results of the greedy algorithm and compared it to the results for all possible combinations of sets for a test case (see Figure 5). We can confirm that it indeed succeeds in finding the most information-carrying set for every given

[4] All but one of the actions were wrong within all 9 color combinations with 4 actions per combination performed correctly

size, without enumerating all sets. Doing this computation is extremely computational expensive, because instead of performing $N(N-1)$ computations, it requires $(2^N)-1$ computations, which is the reason we did not test all possible networks, but we assume that even if sets are not perfectly identified, errors would be negligible (see Supplementary Information from Hintze and Adami 2023 (Hintze and Adami, 2023)).

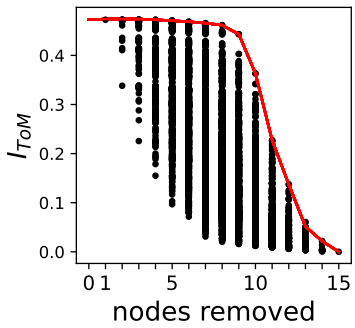

Figure 5: The I_{ToM} (y-axis) for each tested set given a particular set size (x-axis) for all possible sets shown as a black dot. In red the sets identified by the greedy algorithm.

The algorithm determines in which order nodes must be removed to minimize loss of I_{ToM}. This algorithm is "greedy" in the sense that it iteratively takes the best choice at every elimination without evaluating all possible combinations of options. While removing nodes in this order, loss of I_{ToM} must eventually happen (see Figure 5) when there are no more non-influential nodes to remove with respect to I_{ToM}. In this way MBs differ compared to RNNs (see Figure 6). When averaging the I_{ToM} for all MBs we find a flat slope at the beginning and suddenly dropping when about 5 nodes are left, indicating that 10 nodes do not carry I_{ToM} whereas the remaining 5 do (see Figure 6 Panel 1). In RNNs, I_{ToM} decays right away and much more continuous similar to an inverted logistic function (see Figure 6 Panel 2). This suggests that the information is shared among more nodes and isolated with much less definition — a phenomenon that is expected and has been observed before (Hintze et al., 2018; Bohm et al., 2022b). This likely arises because these MBs have discrete components and few of them that in combination sparsely store representations (Marstaller et al., 2013). In contrast, standard neural networks with backpropagation tend to recruit all weights during training (Hintze, 2021), leading to more distributed computations and more distributed representations.

Consequently, due to this informational smearing caused by deep learning, it should be more difficult to isolate which nodes in an RNN carry the I_{ToM}. To test this hypothesis, we performed a perturbation analysis similar to the one for MBs. Each identified set of nodes was set to a random value when agents entered the final room. That way, any information stored in those nodes was destroyed or at least corrupted. When the greedy algorithm properly identified and noised the nodes carrying the I_{ToM}, the ability for agents to report the color the other agent experienced should be more significantly altered than their ability to report the color of their own starting room. The assumption here is that the information about the original room colors for self and other are stored using separate nodes. We observe this for MBs (see Figure 7 Panel 1).

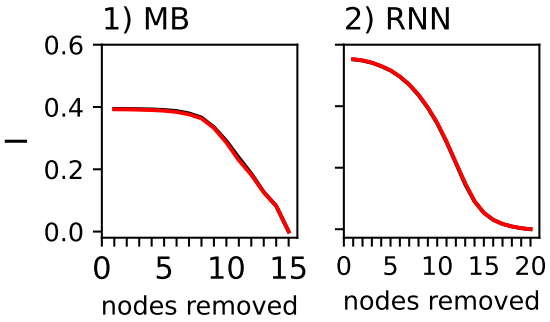

Figure 6: Average I_{ToM} for all perfectly-performing Markov Brains (top panel 1), and RNNs (bottom panel 2) following the shrinking sets identified by the greedy algorithm. In black the measurement for the North agent, and in red for the South agent. We assume differences between the red and blue line come from stochastic noise caused by the probabilistic gates used in the Markov Brain, as this phenomenon does not appear in the RNNs.

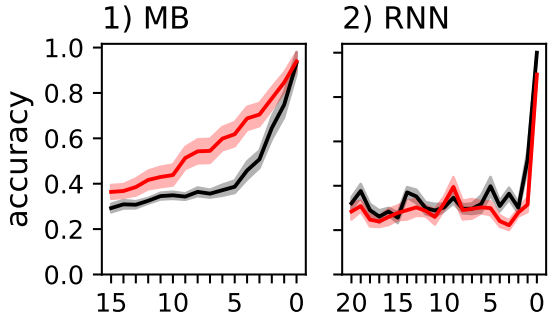

Figure 7: Results of the perturbation experiment. The greedy algorithm identifies sets of shrinking size which always contain the highest amount of I_{ToM}. When agents enter the last room, nodes in these sets are perturbed by setting them to a random number ($[-1,1]$). Performance on reporting their own original room color is shown in red, while the ability to report the color the other agent saw is shown in black. Top panel 1) for Markov Brains, bottom panel 2) for RNNs. Shadows show the standard error computed over all perfect or near-perfect performers.

However, in RNNs information seems much more dis-

tributed over the hidden nodes and I_{ToM} appears to overlap with the representation about their own original room color (see Figure 7 Panel 2). An alternative explanation might be that information is not localized, but moves from one hidden node to the next. For instance, imagine two nodes each node carrying a different representation, but at each update they swap the information they carry between them. As long as the rest of the neural substrate knows where the information is at what time, then the system will function correctly (Bohm et al., 2022a). It is possible that a similar phenomenon occurs in the MBs tested here. If this so, then it happens to a much lesser extent than in this toy example.In confirmation, the perturbation analysis shows that the greedy algorithm applied to the I_{ToM} identifies nodes with representations about the mental states of the other agent.

Lastly, we tried to test every evolved agent with every other evolved agent instead of pairing them together with their clones. For this, each of the 40 evolved controllers (MB or RNN) were used to control the North agent, and paired with all other 39 possible controllers as the South agent. No combination achieved perfect or even near-perfect performance, with the majority failing at over half of all color combinations. When introspecting the communication beeping signal we found that each agent had a specific sequence, timing, and co-occurring behavioral pattern. While it is possible that the same pattern emerges given enough samples, here it did not occur within the 40 controllers tested.

Discussion and Conclusion

Theory of Mind is an important mental property that allows intelligent agents to interact with other agents. In its simplest form it allows agents to understand that other agents have internal states, which is also known as "knowledge about the animacy of things." This notion is extended to knowing the exact mental state, which becomes important in negotiations, interpreting intentions, or efficiently communicating with minimal interaction. However, previous work either uses an external (human) judge to interpret otherwise arguably obvious situations, or studies psychological mechanisms on which Theory of Mind might be built. We found no approach that tries to objectively quantify the amount of information an agent has about the mental states of another agent. For this reason we have taken the first step in this direction. We also now present a method to identify those hidden states containing the I_{ToM}. Our approach is a juxtaposition to psychological or cognitive approaches, which do not necessitate quantification or require a computationally working model (Margolis et al., 2012).

We created a virtual environment (see Figure 1) in which agents must make observations, communicate those observations, and then report about both their own observation and about the observation the other agent made. The environment is constructed in such a way that agents can't cheat, but it requires that agents store their own observations and those of the other agent as mental states. We defined Theory of Mind to be based on exactly this relation, where an agent has mental states about another agent's mental states. While in humans these mental states can be complex constructs like love and hate, here the mental states are simple memories about three possible colors agents experienced.

The reader may feel uneasy about the claim that this simplified measure includes mental states. As a reminder, Theory of Mind is defined as the ability to impute mental states on itself or about others (Premack and Woodruff, 1978). This definition makes no implications about the nature of those mental states. Consequently, if there is disagreement that this simplified measure can measure any kind of mental states, then perhaps the initial definition may be inadequate. Thus far, we do not have that impression and would argue that the nature of the states is irrelevant, and that Theory of Mind can exist about arbitrary states.

Our metric simplifies the measurement of mental states to those that are relevant to the environment, and excludes spurious correlations — or imaginings. This seems to be a harsh constraint, because the ability to imagine something and communicate that to another agent seems an integral part of human interaction, and by saying someone is very imaginative we impute exactly these mental states about imagined content. Our metric seemingly excludes those.

However, this exclusion would only happen if those mental states could never be observed in the environment. The moment an agent utters those imagined states in any form then they do change the world, and could thus be captured in the environmental random variable (E_t). Consequently, as long as the environmental random variable used in our metric is conditioned properly to include how mental states are communicated or how they affect the environment, then our metric will be able to capture them while excluding coincidental information between mental states.

Lastly, we questioned the necessity of a general Theory of Mind in order to have an agent-specific Theory of Mind. Here, agents evolved only agent-specific Theory of mind in that they will not function in combination with each other, but only with their clones. An agent evolved in one experiment fails to complete the task when paired with an agent evolved in a different replicate. Consequently, these agents evolved an agent-specific Theory of Mind without having a general Theory of Mind.

Acknowledgements

The computations were performed on resources provided by SNIC through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under Project SNIC 2020-15-48 and by the Institute for Cyber-Enabled Research at Michigan State University. This work was in part sponsored by the BEACON Center for the Study of Evolution in Action NSF Cooperative Agreement No. DBI-0939454.

References

Ahmed, F. S. and Stephen Miller, L. (2011). Executive function mechanisms of theory of mind. *Journal of autism and developmental disorders*, 41:667–678.

Albantakis, L., Hintze, A., Koch, C., Adami, C., and Tononi, G. (2014). Evolution of integrated causal structures in animats exposed to environments of increasing complexity. *PLoS computational biology*, 10(12):e1003966.

Albrecht, S. V., Stone, P., and Wellman, M. P. (2020). Special issue on autonomous agents modelling other agents: Guest editorial.

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

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., and Plumb, I. (2001). The "reading the mind in the eyes" test revised version: a study with normal adults, and adults with asperger syndrome or high-functioning autism. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, 42(2):241–251.

Barto, A. G., Sutton, R. S., and Anderson, C. W. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE transactions on systems, man, and cybernetics*, (5):834–846.

Bohm, C., Kirkpatrick, D., Cao, V., and Adami, C. (2022a). Information fragmentation, encryption and information flow in complex biological networks. *Entropy*, 24(5):735.

Bohm, C., Kirkpatrick, D., and Hintze, A. (2022b). Understanding memories of the past in the context of different complex neural network architectures. *Neural Computation*, 34(3):754–780.

Descartes, R. (1901). Discourse on method, and metaphysical meditations.[translated by gb rawlings].

Edlund, J. A., Chaumont, N., Hintze, A., Koch, C., Tononi, G., and Adami, C. (2011). Integrated information increases with fitness in the evolution of animats. *PLoS computational biology*, 7(10):e1002236.

Elkins, K. and Chun, J. (2020). Can gpt-3 pass a writer's turing test? *Journal of Cultural Analytics*, 5(2).

Gallagher, H. L. and Frith, C. D. (2003). Functional imaging of 'theory of mind'. *Trends in cognitive sciences*, 7(2):77–83.

Goldberg, D. E. (2013). *Genetic algorithms*. pearson education India.

Goldman, D. S. (2023). A stateful multi-context aware design using openai's gpt (towards digital sentience).

Happé, F. G. (1994). An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of autism and Developmental disorders*, 24(2):129–154.

Hintze, A. (2021). The role weights play in catastrophic forgetting. In *2021 8th International Conference on Soft Computing & Machine Intelligence (ISCMI)*, pages 160–166. IEEE.

Hintze, A. (2022). Understanding the four types of ai, from reactive robots to self-aware beings.

Hintze, A. and Adami, C. (2022). Neuroevolution gives rise to more focused information transfer compared to backpropagation in recurrent neural networks. *Neural Computing and Applications*, pages 1–11.

Hintze, A. and Adami, C. (2023). Detecting information relays in deep neural networks. *arXiv preprint arXiv:2301.00911*.

Hintze, A., Edlund, J. A., Olson, R. S., Knoester, D. B., Schossau, J., Albantakis, L., Tehrani-Saleh, A., Kvam, P., Sheneman, L., Goldsby, H., et al. (2017). Markov brains: A technical introduction. *arXiv preprint arXiv:1709.05601*.

Hintze, A., Kirkpatrick, D., and Adami, C. (2018). The structure of evolved representations across different substrates for artificial intelligence. *arXiv preprint arXiv:1804.01660*.

Horst, S. (2011). The computational theory of mind. *Zalta, Edward N., The Stanford Encyclopedia of Philosophy*.

Kuniyoshi, Y., Yorozu, Y., Ohmura, Y., Terada, K., Otani, T., Nagakubo, A., and Yamamoto, T. (2004). From humanoid embodiment to theory of mind. In *Embodied Artificial Intelligence: International Seminar, Dagstuhl Castle, Germany, July 7-11, 2003. Revised Papers*, pages 202–218. Springer.

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

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Margolis, E., Samuels, R., and Stich, S. P. (2012). *Theory of Mind Alvin I. Goldman, in The Oxford handbook of philosophy of cognitive science*. Oxford University Press.

Marstaller, L., Hintze, A., and Adami, C. (2013). The evolution of representation in simple cognitive networks. *Neural computation*, 25(8):2079–2107.

Miller, J. F. and Harding, S. L. (2008). Cartesian genetic programming. In *Proceedings of the 10th annual conference companion on Genetic and evolutionary computation*, pages 2701–2726.

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

Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. A., and Botvinick, M. (2018). Machine theory of mind. In *International conference on machine learning*, pages 4218–4227. PMLR.

Roy, D. (2005). Semiotic schemas: A framework for grounding language in action and perception. *Artificial Intelligence*, 167(1):170–205. Connecting Language to the World.

Scassellati, B. M. (2001). *Foundations for a Theory of Mind for a Humanoid Robot*. PhD thesis, Massachusetts Institute of Technology.

Sclar, M., Neubig, G., and Bisk, Y. (2022). Symmetric machine theory of mind. In *International Conference on Machine Learning*, pages 19450–19466. PMLR.

Singer, T. and Tusche, A. (2014). Understanding others: Brain mechanisms of theory of mind and empathy. In *Neuroeconomics*, pages 513–532. Elsevier.

Stone, V. E., Baron-Cohen, S., and Knight, R. T. (1998). Frontal lobe contributions to theory of mind. *Journal of cognitive neuroscience*, 10(5):640–656.

Tononi, G. and Koch, C. (2015). Consciousness: here, there and everywhere? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1668):20140167.