From Psychology and Neuroscience to AI: Toward a Learning-Based Computational Theory of Mind Model

Fabio Cuzzolin * and Andrea Morelli

Posted Date: 26 May 2026

doi: 10.20944/preprints202605.1676.v1

Keywords: theory of mind; deep learning; simulation theory; predictions; mental states; reconfigurable networks; stereotypes

 doi:10.20944/preprints202605.1676.v1

Article

From Psychology and Neuroscience to AI: Toward a Learning-Based Computational Theory of Mind Model

Fabio Cuzzolin [1,*] and Andrea Morelli [2]

[1] Institute for AI, Data Analysis and Systems (AIDAS), Oxford Brookes University, Oxford, United Kingdom

[2] Istituto di Scienze e Tecnologie della Cognizione, Consiglio Nazionale delle Ricerche (CNR), Italy

* Correspondence: fabio.cuzzolin@brookes.ac.uk

Abstract

Despite the dramatic advances made in artificial intelligence (AI) and other fields of computer science towards implementing "intelligent" systems expert in specific tasks, the goal of devising algorithms and machines able to interact with human beings just as naturally as other humans do is still elusive. As this naturalness is arguably a consequence of the similarity of the underlying 'hardware' (the human brain), it is reasonable to claim that only artificial systems closely inspired by the actual functioning of the human brain and mind have the potential to render this possible. More specifically, the aim of this paper is to propose a new, biologically inspired computational model able to mimic, in a more accurate way than existing ones, the set of functionalities know as Theory of Mind. This is a set of mental processes that allow an individual to attribute mental states to others. In human social interactions this mechanism is crucial, as it allows one to explain the observed behaviour of others, to guess their intentions and to effectively predict their future conduct. This happens by modelling and selecting the most likely (unobservable) mental states of the considered person, which are the primary causes of everyone's observed actions. The proposed model combines a number of concepts, including those of hierarchical structure, hypotheses pre-activation, and the notion of agent class or 'stereotype'. It rests on one of the main psychological approaches to Theory of Mind, termed Simulation Theory (ST), and is supported by significant neuroscientific evidence. Crucially, unlike previous efforts in AI, the proposed model puts the learning element at the forefront, in the belief that simulations of other intelligent being's reasoning processes need to be learned from experience. In this perspective, a possible implementation of the model in terms of deep, reconfigurable neural networks, trained in a reinforcement learning setting, is outlined.

Keywords: theory of mind; deep learning; simulation theory; predictions; mental states; reconfigurable networks; stereotypes

1. Introduction

1.1. Humans and Machines

Emerging applications of artificial intelligence (AI) are flagging the limitations of established machine learning (ML) approaches in situations involving humans. Smart cars, e.g., need to make reliable predictions about human behaviour in real time, say in order to pre-emptively adjust speed and course to cope with some children's possible decision to suddenly cross the road in front of them. The automated recognition of present behaviour builds, to date, on the success of deep learning (LeCun et al., 2015), a machine learning approach based on artificial neural networks (ANNs) composed by a significant number of layers, to efficiently identify motion patterns in streaming videos (Saha et al., 2016; Singh et al., 2017; Saha et al., 2017; Singh et al., 2018). Motion patterns, however, can be deceiving as humans can suddenly change their mind based on their own mental dynamics and things they may spot in the scene (e.g., the group of children may see an ice cream van

 doi:10.20944/preprints202605.1676.v1

on the other side of the road and decide to converge on it). In fact, humans are capable of predicting future behaviour even when no motion is present at all, just by quickly assessing the 'type' of person involved (e.g., an elderly person standing in a hallway is more likely to take the elevator rather than the stairs). Similar arguments can be extended to stimuli provided by a variety of sources.

A Theory of Mind (ToM) (Bello, 2012), i.e., the ability to 'guess' another agent's mental states (no matter whether a person, or a machine endowed with some level of intelligence), is therefore crucial to develop a next generation, human-centric artificial intelligence. In a mutually beneficial process, computational models developed within AI could, in turn, provide new insights on the way these mechanisms work in the human brain.

1.2. Theory of Mind

More precisely, the term "Theory of Mind" refers to the set of cognitive processes and functions of the human mind that allow an individual to attribute mental states to others. One can simply define as mental state the state in which a person's mind is at a given moment, generally described by a natural language sentence (e.g. "Luca believes that there is no more food in the pantry"). Many mental states consist in the combination of a mental representation and a mental attitude towards that specific representation. In the previous example, the absence of food in the pantry is a state of reality, and the content of Luca's mental representation; the belief that there is no food in the pantry is the mental attitude that Luca has towards this content.

A mental state can consist of both cognitive and emotional components. To make things easier, however, we tend to classify mental states into three main categories: beliefs, desires and intentions. Beliefs refer to anything people believe about the external world, themselves or others; desires represent what they wish to obtain and the goals they would like to achieve. Intentions, on the other hand, are what people intend and have already decided to do. Intentions are generally the result of a decision-making process which has different components as input (including different combinations of desires and beliefs) and which consequently generates an action, which may or may not be observable from the outside. This link between someone's desires, intentions and actions allows people to both interpret a given behaviour and predict a future one. On the one hand, Theory of Mind abilities make possible to try and explain what mental states may be responsible for generating a given behaviour. On the other, attributing mental states to someone else allows people to better predict their future actions.

1.3. Theoretical Approaches to ToM

One of the three dominant theoretical approaches to ToM is called Theory Theory (TT). This maintains that the knowledge concerning mental aspects lies in some structures whose contents are represented in the form of theories, and that the development of a primitive understanding of mental states in children comes with a radical conceptual change (Gopnik, 2003; Gopnik and Wellman, 1994; 2012). Consistently with this idea, TT supports the hypothesis that the more primitive forms of ToM, more apparent in the early stages of a child's life, could represent an implicit form of Theory of Mind; whereas the more verbal forms, which arise after 3-4 years of age, could be an explicit form of ToM (Apperly and Butterfill, 2009; Clements and Perner, 1994; Clements et al., 2000; Low and Perner, 2012; Thoermer et al., 2012; Ruffman et al., 2001; Wang et al., 2012). According to supporters of this point of view, children often elaborate theories and formulate hypotheses in a propositional form, which they try to confirm or disprove through experience, just as a scientist would do.

Something problematic about TT is its assumption that an individual generates, and stores, a very large number of theories about other people and their behaviour. Considering the large number of hypotheses that can be formulated for each person or category of people, and the number of types of individuals that are likely to be encountered over the course of one's life, there would simply not be enough "space" in the brain to store all these theories separately. Furthermore, it must be noted that such a system would be cognitively wasteful. Indeed, a high level of cognitive resources as well as their respective neural substrates would be required, contrarily to the principle of cognitive

 doi:10.20944/preprints202605.1676.v1

economics (the tendency to obtain the most information with the least amount of cognitive resources, Rosch, 2002). Said principle, however, has been shown to characterise various mental functions.

A second perspective to the functioning of ToM is the so-called Simulation Theory (ST) approach. This states that an individual uses (often automatically and unconsciously) a mechanism that simulates other people's mental states using his own reasoning processes. In other words, said individual activates processes that try to provide an answer to the question "How would I behave if I were in their shoes, having their own beliefs and desires?". If we assume that everyone's mind works in a similar way, it becomes possible to predict someone else's behaviour by predicting how we ourselves would behave in the same situation (Harris, 1992).

A third approach which attempts to explain ToM on another dimension, but which does not necessarily contrast with the two approaches previously discussed, is provided by the set of modular theories (Scholl and Leslie, 1999). These state that there is a module specialised for ToM in the brain. In this module, all the mechanisms and theories that allow an individual to attribute mental states to others and to predict someone else's behaviour are specified. According to this approach, both said theories, as well as the concepts of belief and desire, are innate. This module could be the product of an evolutionary process, since the ability to accurately predict others' behaviour in the past must have been crucial to guarantee the survival of the species. These ideas would explain why Theory of Mind seems to develop uniformly among different peoples and cultures.

Although all three approaches arguably have limitations, they all seem to contain some elements of truth, and can all provide useful cues as to how to design a comprehensive computational ToM model. In our view, however, the approach that seems to better explain the actual functioning of ToM in the human brain is ST. We will therefore take the latter as a reference when describing some aspects of the proposed model.

1.4. Paper Contributions

The aim of this paper is two-fold. Firstly, we present a new cognitive model that reproduces and simulates some Theory of Mind processes in a way that departs from previous efforts, by integrating in a coherent whole various existing psychological concepts and theories, such as Simulation Theory, the notion of stereotype, and its link with personality traits. Secondly, the model allows us to reinterpret a number of pieces of neuroscientific evidence present in current debate on Theory of Mind, by seeing them in a different light. It also allows us to explain how an individual can interpret and predict the mental states of others through mechanisms that function in a hierarchical fashion. Finally, a computational implementation of this model is proposed which adheres to the Simulation Theory philosophy while putting learning at the centre, in an attempt to bridge the existing gap between machine Theory of Mind models and machine learning. Such implementation, resting on a novel class of deep reconfigurable neural networks, has the potential to allow intelligent machines to gradually achieve Theory of Mind abilities from experience.

1.5. Paper Outline

The remainder of the paper is structured as follows.

In Section 2, the relevant state of the art in both computational models for theory of mind and machine learning research is reviewed.

In Section 3 our proposed cognitive framework, based on the notion of hierarchical predictive model, is presented. All aspects of the model are described, from the pre-activation of hypotheses (3.1) and its relationship with Bayesian reasoning (3.2), to the notion of hierarchical structure (3.3) and how prediction works there (3.4). The role of stereotypes in the model is described in Section 3.5, whereas Sec. 3.6 discussed prediction errors and their relation to weighting feedback signals. Section 3.7 illustrates internal simulation as another component of an effective ToM model, role which is elaborated upon in Section 3.8.

Section 4 reviews the neuroscientific evidence supporting the proposed model.

 doi:10.20944/preprints202605.1676.v1

Section 5 extends the argument to all other sources of information considered to be crucial to ToM abilities.

Section 6 outlines a computational model based on the proposed psychological one, based on reconfigurable deep neural networks implementing simulations of various classes of agents by re-arranging a number of 'base' neural modules, in a fashion learned via reinforcement learning (i.e., by rewarding network structures leading to successful simulations, and penalising those leading to incorrect predictions).

Section 7 concludes the paper.

2. State of the Art

The dominant approach used nowadays to implement some kind of Theory of Mind in intelligent systems is that of *multi-agent systems*. An *intelligent agent* is an entity that autonomously tries to reach goals, including by acquiring new information and learning new rules of action and how to use the acquired knowledge effectively. In some cases, agents are used to emulate human behaviour, and sometimes even the functioning of the mind, albeit in a simplistic way. To this purpose, one of the most used theoretical frameworks is termed *Belief-Desire-Intention* (BDI).

As the name suggests, BDI approaches model the concepts of belief, desire and intention, as well as other motivational aspects of the agent. Some models are structured in a way that ensures they are able to use meta-reasoning processes (e.g. "I think he believes that you want..."), similarly to how human beings would use them (Bosse et al., 2007; De Weerd et al., 2014; Gmytrasiewicz et al., 2015; Pynadath e Marsella, 2005; Si e Marsella, 2014). More consistently with the ST philosophy, some agent models use a simulation mechanism to foresee what the effects would be if certain actions were performed. This allows the agent to perform different internal simulations and, eventually, select the action that minimises the discrepancy between the real and the desired result (Hiatt e Trafton, 2010; Hiatt et al., 2011; Polceanu e Buche, 2014; Pynadath e Marsella, 2005; Si e Marsella, 2014).

Attempts have also been made to implement social norms that agents must follow, in order to study more realistic agent dynamics and the interaction between these norms and the mechanisms of ToM in such agents (Si e Marsella, 2014). In other cases, the structure of the agent contains elements that more closely recall the way in which the human mind works, such as: the representation of categories of other agents through the use of stereotypes, some constraints linked to *working memory* (a cognitive system with a limited capacity also responsible for temporarily holding information available for processing), and some "personality traits" (Pynadath e Marsella, 2005). In a different approach, an extended version of *ACT-R* was used as an agent's basic structure (Hiatt e Trafton, 2010; Hiatt et al., 2011). *ACT-R* is a cognitive architecture that was designed to simulate the basic cognitive processes of human beings. Another study focused on the implementation of a ToM starting from the two psychological approaches described above: TT and ST (Harbers et al., 2012). Interestingly, the authors noted how the two models reflected the pros and cons that are often found in the psychological literature when the above-mentioned approaches are compared.

Although the most common computational approach to ToM implementation is that of multi-agent systems, efforts have been made from completely different theoretical perspectives, such as *partially observable Markov decision processes* (Baker et al., 2011), *multi-agent reinforcement learning* (Pynadath et al., 2014), *evolutionary robotics* (Kim and Lipson, 2009), and *game theory* (Yoshida et al., 2008). One last model worth mentioning originates from psychological theories that emphasise the development of visual recognition abilities as prerequisite for the development of a ToM in humans (Scassellati, 2002). Based on these theories, this model tries to create complex representations (such as those related to mental states) starting from the recognition of simple elements of a visual scene, such as gaze direction, skin colour, facial contours, etc. This approach is quite significant, for it reflects the temporal consequentiality of the development of cognitive functions (including those related to Theory of Mind) in children.

The aim of this paper is to conceive a model strongly inspired by the evidence provided by psychology and neuroscience, and which could therefore allow a machine to interact with human

 doi:10.20944/preprints202605.1676.v1

beings in a natural way. In this perspective all of these models, which indeed constitute important attempts to implement ToM computationally, exhibit several limitations. Mainly, and crucially, they fail to reflect the real functioning of the human mind. For instance, a strong limitation of most existing models is the *inability to continuously and dynamically learn from experience* and, thus, expand their knowledge base and update the very rules of learning. Another main limitation of these models is that the mental states of the other agents *are often predetermined and certain*: no human-like inferential process is performed to allow the agent to learn to more accurately attribute mental states to others. Moreover, albeit being in line with the TT approach, most of these models use reasoning processes based exclusively on propositional logic, which seems to reflect the way in which human beings use language, but does neither reflect the actual way the brain (often unconsciously) works, nor the way in which knowledge is represented and organised within it.

Overall, computational Theory of Mind models based on machine learning remain in their infancy. A solution, we argue here, may be provided by artificial neural networks (ANNs) whose topology dynamically adapts to the data (Juang et al., 1998; Fritzke, 1994), a concept recently extended to recurrent neural networks (She et al., 2014). Models which assemble neural networks starting from a collection of composable modules have recently been proposed for answering queries related to images (Andreas et al., 2016a). In such an approach, various possible layouts of the overall neural network (i.e., various possible ways of connecting the base modules) are sampled and assessed, while network parameters are jointly learned via reinforcement learning (Andreas et al., 2016b).

3. A New Cognitive Model of Theory of Mind

As stated, the aim of this work is to propose a model that could allow the creation of intelligent systems able to smoothly interact with human beings. In order to do this, we must start from the assumption that an individual learns to easily predict and understand other peoples' actions because his neural mechanisms and the functioning of his own mind are extremely similar to those of others. Therefore, implementing a computational system with a model that more accurately represents the functioning of human cognitive mechanisms, at least in what concerns ToM processes, could greatly facilitate a seamless human-machine interaction. For this reason, in the rest of this Section we will often refer to the evidence provided by psychology and neuroscience in relation to the real functioning of the human mind and brain, in order to illustrate the various aspects of the proposed model.

3.1. Generation and Pre-Activation of Hypotheses

One of the most relevant aspects of the proposed model is that it reflects the ability of the human brain to generate and "pre-activate" some hypotheses, having only partial stimuli coming from the environment in the form of inputs. To better understand this phenomenon, we can refer to what happens in the field of vision with what is called the *inverse problem*.

The latter consists in the fact that our visual system must constantly decide what has generated the perceived stimuli; such stimuli can be partial, distorted or ambiguous. For instance, Figure 1 depicts two distinct animals, a rabbit and a duck. Nevertheless, at any given moment, our mind alternately elaborates hypotheses (for example, by activating our mental models for "rabbit" or for "duck") and selects the one it considers to be more likely.

This mechanism applies to other situations as well. Imagine being in front of a person who is turning away from us and has long hair. If we had to decide whether that person was a man or a woman, our brain would generate a hypothesis and suppose, for instance, that the person is more likely to be a woman because of their long hair (probably because we are more used to seeing women with long hair, rather than men). It is clear from this example that, rather than just proceed "backwards" (from the reception of sensory inputs, to the reconstruction of their causes), our neural system(s) also make *a-priori* assumptions about hypotheses, especially in the presence of partial or no information. The hypothesis that generates the best prediction (the one with the highest probability of being activated at any given moment) will determine the content of our perception. When,

 doi:10.20944/preprints202605.1676.v1

subsequently, we receive new information that we consider useful for our decision making process (e.g., we notice that the person with long hair is waiting in line for the men's bathroom), a mechanism of hypothesis correction and update of their probability of activation is triggered (we can then decide that said person is more likely to be a man). What happens is that our system compares the result that was previously predicted by our *a-priori* model, to the input subsequently received, and consequently updates the model. In other words, our mind changes the probability that said model (as well as others) will activate in the future due to that specific input (a concept know in probability theory as *posterior probability*).

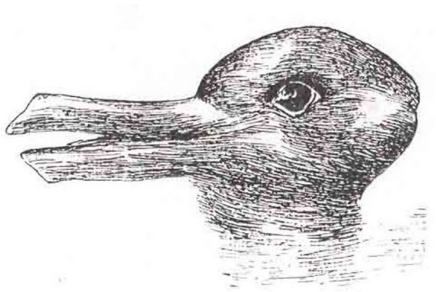

Figure 1. An optical illusion. The human mind alternatively perceives in this picture either a duck or a rabbit.

3.2. Bayesian Learning

Models can be updated in various ways, meaning that there are different rules automated systems can follow to learn. Although from a computational point of view it is possible to select any one of them (also depending on the structure and purpose for which the machine was created), in this part of our work we will focus on *Bayesian learning*, in the way of example. Although there is no certainty that the human neural system works precisely in this way, Bayesian learning allows us to take into account all the aspects we considered so far: *a-priori* probabilities of the hypotheses to generate, evidence provided by new observations, and new *a-posteriori* probabilities of the same hypotheses, obtained after updating.

The way probabilities are updated in the Bayesian framework is by means of Bayes theorem, or *Bayes' update rule*:

$$P(A|E) = \frac{P(E|A)P(A)}{P(E)},$$

where:

- $P(A)$ is the *a-priori* probability of the hypothesis (on the cause of the observed stimulus) – a probability distribution which does not take into account any information concerning the available evidence E;
- $P(A|E)$ is the *conditional* probability of A, known E – this is also called the *posterior probability*, since it depends on the observed value of E;
- $P(E|A)$ is the conditional probability of E, known A – in other words, this is the probability of actually perceiving evidence E if our hypothesis A were to be true;
- $P(E)$ is the *a-priori* probability of E, and serves as a normalisation factor.

Intuitively, the theorem describes the way in which a hypothesis A is enriched by having observed the event E. Put differently, it describes how the *a-priori* model A (or the probability of its activation) is modified after receiving the new input E. The latter will be compared with the previous prediction elaborated by the system and, regardless of what the input received might be, the system will update the initial hypotheses through the prediction error, which will proportionately correspond to the discrepancy resulting from this comparison. If the prediction is correct, only a small part of the new evidence will be used as error signal. Otherwise, if the prediction is wrong, the

 doi:10.20944/preprints202605.1676.v1

prediction error will be higher, and most of the information received will be used to better update the previous models (and consequently, to correct future predictions).

The different hypotheses and their activation probabilities are modified according to a principle of error minimisation: every update taking place tends to minimise future prediction errors.

3.3. Hierarchical Predictive Structures

According to several authors, these inferential and model updating processes do not occur only on two levels (the level where new inputs are processed and the level where hypotheses are generated), but rather over a very large number of them, organised in a *hierarchical structure* (Hohwy and Palmer 2014; Hudson et al. 2016; Kilner et al. 2007; Koster-Hale and Saxe 2013; Ondobaka et al. 2017; Palmer et al. 2015; Westra 2017). Such a hierarchical organisation falls within what is generally called *hierarchical predictive coding*, a family of models regarding information processing systems which emphasises the importance of predictive processes, organised into hierarchical structures.

In our model these are arranged in such a way that the higher levels process predictions based on characteristics of the environment that are very abstract and generally more stable in time; whereas the lower levels elaborate predictions based on more local (and temporary) environmental features. Taking the visual system as an example, while predictions made at the lower levels can represent very simple properties such as contours, surfaces and colours; predictions produced at the higher levels can represent more global, abstract and complex elements such as, for example, the representation of one's body movement (*action recognition*) or the belonging of an object to a given category (*object classification*). As we will see in Section 6, these principles are embraced by modern convolutional deep neural networks (CNNs).

Within this hierarchical organisation, the processes of inference generation and learning occur between each pair of adjacent levels of the hierarchy. To better understand this, let us consider, initially, only two levels: an upper one (*B*) and a lower one (*C*) (depicted in Figure 2). The upper level *B* will generate a hypothesis concerning the input that previously reached the lower level *C*. This level, in turn, will compare this hypothesis and the new received inputs, and will pass the signal representing the error committed in the inference to the *B* level; this error, as already mentioned, will change the initial model at level *B*. In the same way, level *B* will act as a lower level for the even higher level *A*; that is, the hypothesis that level *B* will consider as more likely, will be passed as new evidence to the upper level *A*, which will in turn have formed a hypothesis to be compared with this new input, and so on. Level *C* (which had provided inputs for level *B*), will likewise be a predictive level for the even lower level *D*, and so on.

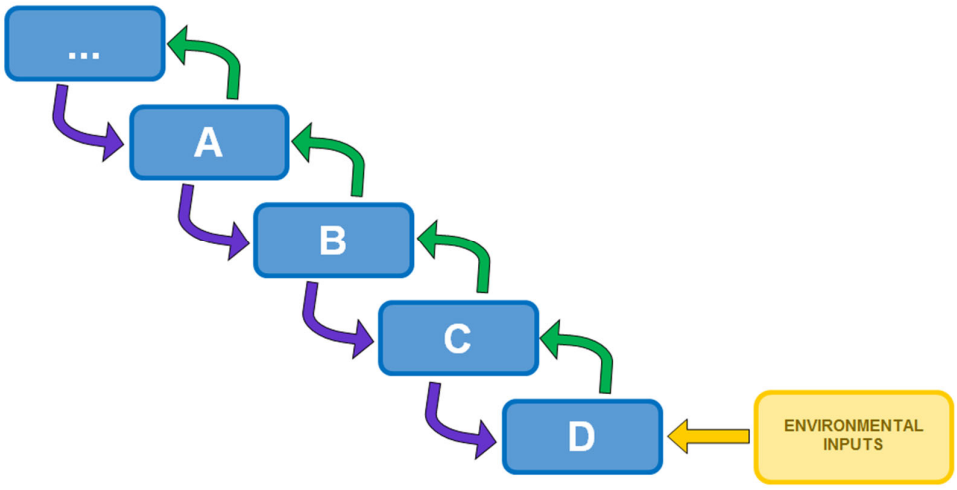

Figure 2. In hierarchical predictive models, predictions (purple arrows) and error signals (green arrows) flow between each pair of levels in the hierarchy.

 doi:10.20944/preprints202605.1676.v1

By doing so, therefore, the content of perception is generated through two main processes: a bottom-up process, through which new evidence is collected and an error signal is passed from one level to another; and a top-down process, through which the highest levels of the hierarchy generate models and hypotheses to be compared with the new evidence present in the lower levels.

Going back to the previously mentioned example, while the lower levels could represent the contours of a human figure seen from behind or different parts of it (more specifically, the deepest levels represent local and very simple visual details, such as a certain colour or the orientation of some segments, or less simple ones, such as that person's hair colour and length), the upper levels may represent more complex elements, such as the concept of "woman". This last information is passed as initial prediction to the lower levels which, upon gathering new information (such as the fact that the person is waiting in line for the men's bathroom), will generate an error signal that will be re-sent to the upper levels, in order to modify the probability that the previously generated hypotheses will activate in the future in the presence of the same stimuli. Thus, in the example, decreasing the activation probability of the representation of "woman" and increasing that of "man".

3.4. Prediction of Actions and Mental States

The hierarchical structure we described in the previous paragraph in relation to the visual system can reflect various other cognitive processes, such as the recognition of other people's actions and behaviour, or the attribution of mental states to others.

Just as the visual system does, the neural regions related to ToM processes must also solve an "inverse problem" – indeed, they need to understand which mental states have caused the observed behaviour. The challenge is, in this case, how best to use the observed evidence (actions and states related to an agent) in order to infer the causal and unobservable structure that generated that evidence (e.g. goals, beliefs, intentions, emotions and personality traits of the person). This prediction process allows us to extend such predictions to objects and concepts of different nature, which may be more or less abstract, more or less stable and lasting in time.

Indeed, different components/concepts will always be predicted at different time scales. For instance, a prediction concerning a person's movements may take just a few milliseconds, whereas that concerning which insurance agency an individual will choose in a month's time will require a longer and more complex process, forcibly taking place in a much longer time frame. In this second case, any prediction will also be more stable and long-lasting than in the first one. We can conjecture that the layers encoding the first type of prediction will be located in lower part of the hierarchy, while layers encoding the second kind of prediction will instead be located in the higher echelons. We can then picture a broader hierarchical system organised this way: in the lower levels, the hypotheses formulated continue to concern the perceptual components of an image (for instance), as well as elements of different sensory nature (sounds, smells, etc.). At an intermediate level, instead, the system generates inferences about the type of action that is performed by a person. Finally, at the higher levels hypotheses are validated which concern the mental states that may have generated the observed actions. According to what argued above, while in the action prediction process the stimuli being compared with the hypotheses represent movements of one or more body parts, in prediction processes concerning mental states the inputs are the actions themselves, seen as a whole.

3.4.1. Intentional Actions

It is worth noting that intentional actions (or, put differently, the representations of different intentions to perform one or more actions) can also be structured in a hierarchical manner and represented on several layers, both from a conceptual and a temporal point of view. For instance, let us consider the desire of eating something to satisfy one's feeling of hunger. This desire could lead us to assessing different sub-goals, to later choose the best option (for instance, "make yourself a sandwich"). In turn, this sub-goal can be divided into finer, more specific sub-goals, which can be ordered in the following timeline: "look for bread in the pantry", "cut the bread", "open the fridge", "look for cold cuts", etc. At an even lower level, the individual intentions to move towards the pantry,

 doi:10.20944/preprints202605.1676.v1

to open it and to take the bread will be generated. At an even finer level of granularity, intentions for things such as standing up and starting to walk toward the kitchen may be selected, and so on.

3.4.2. Interaction Between Layers

According to the prediction process described above, for each pair of levels of the hierarchy, the generation of a hypothesis at a higher level allows us to attach *a-priori* probabilities to the activation of those relevant to the level immediately lower, so narrowing down the space of possible hypotheses that will be activated later at this new level. This mechanism also extends to the higher levels of the hierarchy, starting from the formulation of hypotheses about mental states that have generated the intention to perform a certain action.

For instance, if I attribute to a person the desire to go to the kitchen to get a drink, I will assign a high *a-priori* probability to all the hypotheses of the lower levels that are consistent with the higher-level hypothesis "desire to get a drink" (e.g. going to drink water, drink a fruit juice, take a bottle of beer from the fridge, etc.). At the same time, I will assign a low *a-priori* probability to all the alternatives not consistent with it (making a sandwich, washing one's hands, cooking some meat, etc.).

3.4.3. Mental States Are Represented at Multiple Levels of the Hierarchy

As already mentioned, a higher-level mental state, such as the desire of "making a sandwich" is more stable and lasting over time than "taking bread from the pantry", since the former generally remains persistent from the beginning until it has been satisfied (in this case until the sandwich has been prepared), or until a better intention is generated to satisfy the more general desire of eating. A lower level mental state, on the other hand, such as the intention of performing the sub-action "cutting the bread" lasts for a shorter stretch of time and, as it is much more specific, it may be subject to various unexpected events (e.g., there is no more bread left), changes, and can therefore be more unstable. The bottom line is that *mental states can be represented at different levels of the hierarchy*, where they differ in terms of duration and temporal stability. Therefore, there is neither a single "module" representing "desires", nor a single module encoding "intentions" (as in the agent software field). Rather, a range of level-specific modules such as "level 1 desire", "level 2 desire", "level 3 desire" and so on exist, down to "leaf" modules describing the desire (or intention) to perform a very small action.

3.4.4. Desire Versus Intention in the Hierarchical Model

As a consequence, in this conceptual framework, it is not always possible to clearly define a boundary between the mental states "desire" and "intention", as most models based on software agents appear to be doing. In our view, the difference that is often pointed out between the concepts of desire and intention is purely conventional. It seems that "intention" implies a consequent and often immediate action, while desire does not. This difference, however, can be explained simply by the fact that "desire", in the way it is usually referred to, cannot be transformed directly into a single and simple action, as can it be done with more specific intentions. An intention, in this sense, is nothing more than a mapping of a desire into sub-components, related to increasingly specific and time-limited actions.

In our model, learning plays a crucial role in differentiating concepts such as mental states.

3.5. Stereotypes and Personality Traits

Another fundamental component of our proposed ToM model concerns *stereotypes*. A "stereotype" can be defined as a mechanism that allows to make generalisations, in the sense that some characteristics attributed to a specific social group or category of people are "inherited" by all the individuals belonging to that particular group. In this way, starting from the membership of individuals to a specific group, it is possible to attribute to them characteristics such as beliefs,

 doi:10.20944/preprints202605.1676.v1

attitudes and personality traits that are considered to be specific of that particular social group. Membership to these social or stereotypical categories is generally determined on the basis of some physical characteristics of the individual, such as skin colour, facial features and accent (Mason et al., 2006), as well as other types of information such as gender (Burnham and Harris, 1992; Condry and Ross, 1985) and age (Mueller-Johnson et al., 2007; Goodman et al., 1987). Various other features can be used to describe the individual within a stereotypical category, such as food preferences, clothing, music interests, social practices, ethnicity, and so on.

3.5.1. Stereotypes and Theory of Mind

The connection between stereotypes and Theory of Mind is stronger than one might think. The processes of mental states attribution and stereotypes activation are both triggered spontaneously every time one observes or interacts with other people. As already mentioned, while activating a certain "social category", we simultaneously generate specific mental states that we link to it, in a usually automatic, unconscious manner. For instance, if we see a man in a store with a knife in his hand, we may simply think he is a butcher, if he is wearing a white uniform and a white hat. Whereas we might believe him to be a terrorist, if he is wearing a black tunic and has a long beard. In this last case only, we will attribute harmful intentions to that person and act accordingly, perhaps moving quickly away from the area. In one of the many studies on this topic, for example, the evaluation of some of other people's behaviours changed based on the colour of the actors' skin: indeed, the behaviour of black people was judged to be worse and more threatening than that of white people (Sagar and Schofield, 1980). Hence, the participants tended to attribute harmful rather than benign intentions, depending on the different skin colour only.

Concepts that usually act as a bridge between the stereotypical elements activating a specific category and the mental states attributed to the individual members of that category are *personality traits*. These can be seen as psychological properties which are unobservable and relatively stable during the course of time (laziness, intelligence, honesty, aggression, etc.; Doris, 2002). These characteristics are activated by some stereotypical elements, and in turn, they pre-activate certain mental states.Through this process, stereotypes allow us to infer mental states of people we do not know, but who, however, belong to social groups known to us. It is precisely the association among stereotypical elements, personality traits and mental states that, over time, creates the various stereotypical categories.

3.5.2. Stereotype and Predictive Hierarchy

Within the predictive hierarchy stereotypes and, more specifically, personality traits are located in the highest levels, as they are the most long-lasting and stable associations over time (Doris, 2002). Indeed, while the desire of eating a dessert can last for a few minutes or even a few hours, the physical characteristic of being obese can activate in the observer a stereotype that associates with the individual a personality trait that is more general and much more lasting over time (as, for example, the label of "greedy" referred to person). Also, in this case, the activation of certain stereotypes allows us to attach a certain *a-priori* probability to the beliefs and desires of the individual in the underlying levels. When identifying higher-level mental states, the stereotype "pre-selects" some of them by increasing their *a-priori* probability and reducing that of others, in the same way we discussed for other levels of the hierarchy. Therefore, if, for example, we see a person finding a wallet on the ground, depending on the fact that we attribute an "honest" or a "dishonest" personality to that person, the predictions within the hierarchical structure (Figure 3) will allow us to foresee whether that person will take the wallet to the police, or keep it for himself or herself.

 doi:10.20944/preprints202605.1676.v1

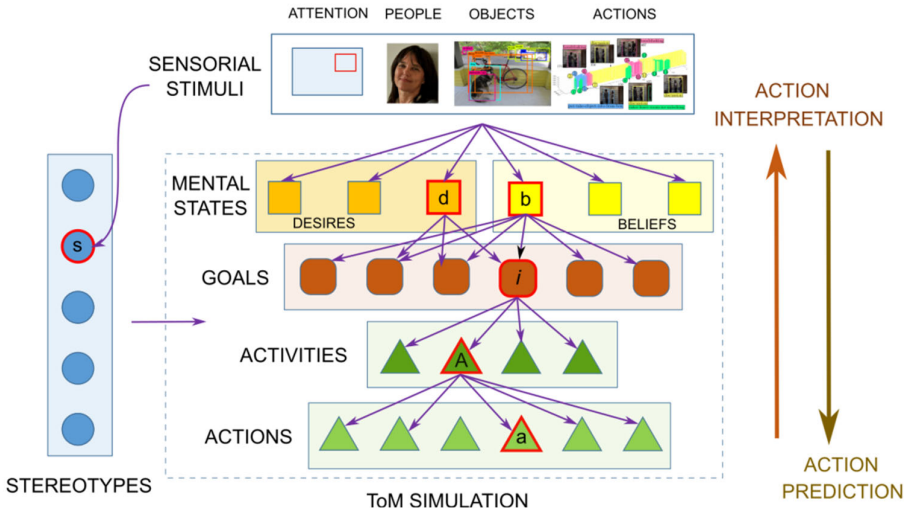

Figure 3. Predictive hierarchical structures, as described in Section 3, allow us to both predict future behaviour in complex agents, and interpret the observed actions in terms of mental states.

3.5.3. Stereotypes as an Accurate and Efficient Mechanism

This pre-selection mechanism performed by stereotypes is a guarantee of higher prediction accuracy and, especially, of lower use of cognitive resources, especially in situations in which we do not know the specific person we are observing well. Indeed, if no characteristics, physical or otherwise, could allow us to pre-activate some mental states, our brain would need to evaluate a much higher number of hypotheses, thus spending more time and using more mental resources and memory "space" to store all the associations related to any single person (as TT would suggest). At any rate, judgments would be more superficial, less accurate, and the probability of making attribution errors would be higher.

3.6. Prediction Errors and Weight of the Error Signal

Stereotypes thus allow us, through a narrowing of the hypotheses potentially to activate, to save cognitive resources during the processes of mental states attribution and actions prediction. In so doing, they ensure a reduction of prediction errors (if one knows that their interlocutor is hungry, and they can see that they are obese, one could attribute to them the desire to eat highly caloric and unhealthy food, and not attribute any desire of eating healthier food).

3.6.1. Contrarian Evidence

Nevertheless, stereotypes can be misleading. Categorising people based on some physical characteristics amounts to assuming that all people with these characteristics also have similar personality traits. Obviously, individuals may well differ from the average and represent an exception. Indeed, categorising a person is equivalent to having an approximate representation for them: not all of their characteristics are taken into consideration and this sometimes leads to attribution errors (it could turn out that said interlocutor, despite their obesity, is a vegan). When an incorrect prediction is generated, in association with one or more stereotypes, the feedback received from the "outlier" person provides contrarian evidence which can be exploited to update the model and the different activation probabilities of future hypotheses, just as it happens in the lower levels of the hierarchy.

3.6.2. Pace of Change of Stereotype Hypotheses

Changing the model in this way, however, does not mean disproving it. Indeed, an important aspect, which is ensured by the higher location of personality traits (and thus, of stereotypical

 doi:10.20944/preprints202605.1676.v1

categories) in the hierarchy, is that the rate at which hypotheses are updated at these levels is lower than that of the lower levels. For the latter, such as those relating to intentions or action recognition, little contrarian evidence is enough to radically modify the hypotheses generated (e.g. fine contextual data, such as clothing, or the presence of a carpet, are enough to interpret a person's action of prostrating on the ground as "praying", as opposed to "practicing yoga"). To amend higher-level hypotheses, and with them the associations between perceived characteristics and personality traits, a much higher amount of evidence of disagreement between fact and prediction is needed (the fact that someone knows only one obese person eating healthy will not make him radically change their opinion about the stereotypical association, which instead asserts the opposite).

This is not a problem at all, but rather a guarantee of stereotypes' functioning role. If these were radically altered each time a contrarian piece of evidence was gathered, stereotypical categories would simply disappear (or they would not even be able to arise from experience), thus affecting the overall prediction accuracy of the model.

3.6.3. Weight of the Error Signal

The correctness of the attribution of others' mental states, pre-activated by stereotypes, is greatly assisted precisely by the existence of stereotypical categories that are robust to noise and rather stable over time. Thus, when implementing ToM within a hierarchical organisation such as the one we propose, it is essential to include parameters that determine *the weight that has to be assigned to error signals* received for other levels and, consequently, the robustness of the system against noise.

This weight can vary depending on several factors, including the salience that the specific evidence has for someone: if we believe that the information provided by the new evidence may be useful to improve the accuracy of our future predictions, or if we believe that such information is reliable (in other words, if we believe that the new evidence is very close to represent the average of the regularities present in the environment), then we will pay close attention to this piece of information and attribute a higher weight to the error signal it provides.

The weight attributed to this signal could also vary according to action plans, goals and motivations of the main agent and, more generally, to the type of required predictive task. In cooperative tasks, for instance, in which people have to coordinate their actions with those of their partners, it is necessary, in order to achieve a goal, for each agent to understand correctly the mental states of the other persons involved. Therefore, the focus is on maximising the accuracy of their predictions regarding the specific persons they have to coordinate with. In order to do so, agents need to pay close attention to all the feedback received, giving them particular importance (Godfrey-Smith 1991). In other cases where, for example, an agent has to identify predators, improving the accuracy of predictions toward a specific predator may not generate a higher benefit, as it would create a far too specific representation of "predator", resulting in a high number of false negatives after several updates of the model (while wasting useful cognitive resources). In machine learning, this issue with the generality of a model is referred to as *overfitting*. To provide another example, if I am to hire a new employee, accurately understanding the candidate's mental states during a job interview is crucial, because I will have to work with that person in the future. In this case, the activation of general stereotypes linked to first impressions should be inhibited, whereas prediction accuracy for that specific person's mental states and actions is of the outmost importance. On the other hand, in situations of simpler and more familiar cooperation (e.g. paying a cashier), a lower level of precision and accuracy is required (since, the way in which the cashier guides me through the payment is generally similar in different contexts and situations). In this case, a prediction based mainly on general rules of stereotypical behaviour is more effective (less cognitive resources are needed and the prediction and reaction times are shorter).

This is very useful in lower complexity tasks, such as driving, and could be of utmost importance when designing, for instance, an autonomous vehicle with human-comparable capabilities.

 doi:10.20944/preprints202605.1676.v1

3.7. Internal Simulation of Mental States

The proposed model features another mechanism, which we call *internal simulation*. This simulation allows us to predict the behaviour of others, through the above-described mental states attribution process, in the absence of specific stereotypical traits. Indeed, humans can almost always infer what another person will do in a given context, even without having specific information about the individual. One is led to infer that, in such cases, the only inputs we rely on in order to predict their behaviour are those that arise from ourselves. Specifically, we can predict the others' mental states and actions using our own reasoning processes, by predicting what we ourselves would do in a given situation (once again, as suggested by ST).

3.7.1. Simulation by Perspective Taking

The aim of this mechanism is to answer the question "what would I do / think / feel if I were in his/her situation?". In order to do this, we must mentally project ourselves into the situation the other finds his/herself in, through a process called "perspective taking". In order to take this new perspective we picture ourselves in that specific context. If we have already experienced a similar situation in the past, we will try to retrieve memories related to it. If the situation seems to be new to us, we will imagine ourselves in that context, recreating and mentally visualising the new scene with all the necessary elements and details, making it as realistic as possible. From a cognitive point of view, therefore, this process could be seen as creating and putting ourselves in an imagined ("reconstructed") scene or situation through the reactivation and integration of mental representations of different elements. From a neural point of view, instead, this process could reflect the activation of a number of neural patterns (related to the different elements of the scene) which, once integrated with one another, allow us to create that specific representation in our mind (it must be noted that this process may also not happen consciously).

3.7.2. Simulation in Other Mental Functions

Simulations defined in this way can also be found in other well-known mental functions, such as *episodic* and *autobiographical memory*, *counterfactual thinking*, *episodic future thinking*, in addition to what has already been said for the processes of *perspective taking* (for a review please see: Schacter et al., 2008). When, for example, we remember an event that happened in the past, all we do is reconstruct our memory of it, reactivating and recovering different elements and integrating them in a coherent whole (often altering its original nature). In counterfactual thinking, what we do is answer the question "What would have happened if instead...?". In this case, we retrieve some elements related to a past event and add other elements to it, or intentionally modify some of them, in order to simulate "how it would have gone if...". Moreover, in this case we "project" ourselves in the readapted memory with a first-person perspective, and we see how our mind reacts: what emotions do we feel? What do we want? How would we like to act?

In episodic future thinking we do the same, whilst imagining something that has not happened yet. We build events and contexts which are more or less likely, but always starting from known elements. Just as in the other named processes, we try to put ourselves in the imagined situation and understand what mental states we would experience and what emotions we would feel. Other authors also associate this simulation mechanism with lower-level cognitive processes such as perception and action. For instance, according to Hesslow (2002), imagining we perceive something (both through vision and through other senses) is no different from actually perceiving something. The only difference, in the latter case, is that the inputs that generate that perception do not come from the external environment, but rather from our brain itself.

Finally, another fundamental aspect linked to the simulation process, which is highly instrumental to Theory of Mind, is *empathy*. Through empathy it is possible to reactivate neural patterns that trigger our perception of emotions. In other words, it consists in the activation of neural

 doi:10.20944/preprints202605.1676.v1

patterns that allow us to experience emotions that we would feel if we were "in the other person's shoes" (for a review please consult: Oberman and Ramachandran, 2007).

3.8. From the Simulation of the Self to the Prediction of Others

How does this relate to our hierarchical predictive model? As already mentioned, this internal simulation, performed through reactivations of neural patterns and mental representations, pre-activates some mental states, in the same way stereotypes do. The result of these simulations, indeed, corresponds to a selection of some mental states whose activation probability is, as a result, increased. The overall consequence is that such a simulation process allows us to predict, in a hypothetical situation, states and behaviours of ourselves, or of people who are similar to us. The more different from us the other person is, the less accurate such a prediction-via-simulation will be.

We are now ready to understand how stereotypes and internal simulation can be seamlessly integrated. In order to predict a person's behaviour the outcomes of the simulation of the self are combined with those derived from the activation of potential stereotypes. This produces the likelihoods of the actual combination of mental states to use to predict the different actions and sub-actions of the observed person. This is illustrated in Figure 3. Note that the results produced by the two separate mechanisms will be weighted differently, depending on how much is known about that specific person, and on how similar they are to us. When information about the other individual is scarce, the system will have to rely more on internal simulation. If the individual's characteristics and personality are known in detail, the opposite will happen. Similarly, the weighting of the two components will also change depending on whether the person we are making predictions about is similar to or different from us. The more similar to us the other is, the higher will be the weight given to the internal simulation process; the more unfamiliar and different from us the other person is, the more weight the system will assign to information provided by stereotypes (Ames, 2004a, 2004b).

4. Neuroscientific Evidence Supporting the Model

Neuroscientific evidence does indeed also support our proposed Theory of Mind model.

4.1. Hierarchical Predictive Structure and Neural Correlates

Our proposed hierarchical organisation is inspired by the organisation of different structures of our central nervous system, such as, for example, the visual system (Felleman and Van Essen, 1991;

Hochstein and Ahissar, 2002). Indeed, this type of architecture is reflected in most brain regions involved in visual information processing: the lower (primary) areas receive sensorial inputs, while the upper areas play a significant role in *multimodal associations*. Multimodal association areas are brain regions in which inputs coming from different senses are integrated. In this sense, "multimodal" and "associative" can be used as synonyms. In these regions, the neurobiological concept of hierarchy is based precisely on the distinction between *forward* (FW) and *backward* (BW) connections, which act in a similar way as the top-down and bottom-up mechanisms described above (Murphy and Sillito, 1987; Felleman and Van Essen, 1991; Sherman and Guillery, 1998; Angelucci et al., 2002).

The concept is also related to the notion of *backpropagation* in neural network training (see Section 6).

4.2. Actions Prediction and Neural Correlates

In accordance with the proposed predictive model based on error propagation, it was found that the activation of the superior temporal sulcus (STS), a brain area involved in the recognition of actions (Grossman et al., 2005), decreases when the action can be easily predicted, whereas it increases when the action is not easily predictable, or violates expectations about the subject (Koster-Hale and Saxe, 2013; Grossman et al., 2010; Kable e Chatterjee, 2006; Costantini et al., 2005; Saygin et al., 2011; Pelphrey et al., 2003; Pelphrey e Vander Wyk, 2011; (Pelphrey et al.,2004).). It also seems that an observer's STS is also more active whenever an action is performed correctly, yet inefficiently (Jastorff

 doi:10.20944/preprints202605.1676.v1

et al., 2010; Brass et al., 2007), or whenever an action that fails in its intent or is not performed correctly is observed (Shultz et al., 2010; Shibata et al., 2011).

Our model may provide an interpretation of the way STS functions. In our model's perspective, the initial hypotheses formulated by the subject would be generated there. Such hypotheses would in turn modulate (by descending the hierarchical structure) the input layers, which encode the action performed as well as all the associated contextual elements. The less correct the initial prediction is, the more the error signal is transferred to the higher levels, thus supposedly generating the above-mentioned activations. In this interpretation, this region encodes, in some way to be further explored, the subject's expectations towards a given action.

4.3. Mental States Prediction and Neural Correlates

Similar evidence can also be found with regard to the prediction of mental states. Indeed, as shown for STS, the *temporo-parietal junction* (TPJ) seems to become more active when a person's beliefs and desires are not easy to predict. Assuming that the actions of an individual are generally consistent with their mental states (Malle, 1999), it has been discovered that TPJ (particularly the right junction) seems to be sensitive to inconsistencies between the observed action and the likely mental states of that person (as guessed by the observer), and to become more active when predicted and real intentions appear to be inconsistent (Buckholtz et al., 2008; Koster-Hale et al., 2013; Yamada et al., 2012; Young e Saxe, 2009; Berthoz et al., 2002).

4.4. Influence of Stereotypes and Neural Correlates

Neuroscientific evidence also supports the role of stereotypes in the predictive process. Indeed, experiments showed that the pre-activation of a stereotype associated with an individual increased TPJ activation whenever there was an inconsistency between the mental states predicted on the basis of the specific stereotype, and the real ones subsequently communicated (Saxe e Wexler, 2005; Cloutier et al., 2011). A piece of evidence worth noting is that, in some cases, such difference in the activation of this area (between situations which are easily or not easily predictable) disappears when no additional information concerning the person who is performing the action is provided (e.g. physical characteristics, personality traits, etc). In the study this would happen despite the fact that participants kept declaring some conditions to be more predictable than others (Young et al., 2010).

This suggests that TPJ could indeed be one of the brain areas involved in predicting others' mental states through the activation of stereotypes.

4.5. Simulative Processes and Neural Correlates

An interesting piece of evidence that corroborates the idea of the existence of two mechanisms that specialise in predicting mental states of similar or dissimilar people relates to an area of the *prefrontal cortex*, specifically in its *medial region* (mPFC). Indeed, while the dorsal area of the medial prefrontal cortex (dmPFC) specialises, over time, in judgments made on people we are unfamiliar with and who are different from us, the ventral area of the same region (vmPFC) is generally activated when forming judgments related to the self and to people similar to us (Mitchell et al., 2005; 2006).

In the light of our proposed model we may conjecture that the dmPFC be one of the areas that support stereotyping (in concert with the TPJ, as discussed), with the vmPFC supporting, instead, the internal simulation process used to predict similar people's mental states.

Another peculiarity is that while the dmPFC initially activates in both types of judgments, it subsequently specialises only in assessments related to people different from us (Pfeifer et al., 2007). This transition could reflect, in part, the formation of stereotypical categories. Indeed, one can sensibly hypothesise that this area is composed of different sub-regions, each specialised in responding to stimuli related to specific categories of people. These sub-regions may indeed not be directly distinguishable with current neuroimaging systems.

 doi:10.20944/preprints202605.1676.v1

Finally, the extreme closeness of the ventral and the dorsal regions of mPFC, on the one hand, would explain how making predictions about the others may be partly based on a simulation of the self. On the other hand, it would explain all the prediction errors we make when attributing one's own mental states, emotions and personality traits to another people. What happens, indeed, is that we too often believe that the others are similar to us, more than they may actually be. While this mechanism gives us a tool to make judgments in situations in which that would be otherwise impossible (due to a lack of information related to the others), the same mechanism leads us to a variety of errors, due to the "projection" of ourselves into the others.

5. Input Stimuli to the Predictive System

Stereotypical features are in no way the only elements that our mind takes into account when attributing mental states to others. The variety of information that is processed by the whole Theory of Mind "functionality" in humans is enormous.

Some of these sources of information are worth mentioning:

Movement: the movements of the limbs, the direction in which a person walks, as well as the movement of objects and other agents in the visual scene, are important when trying to understand and predict mental states and actions of a person acting in a complex scene and in a longer time frame.

Facial expressions and emotions: among the major cues that allow us to understand how another person is feeling at any given moment are facial expressions. It should be noted that the internal simulation process can provide us not only with information about what the others may think or what intention they may have at a given moment, but also on how they might potentially feel or what emotions they might experience. Much like the recognition of an action serves as useful evidence to correct the initial predictive hypotheses, facial expressions can provide a useful error signal to correct the initial prediction related to the supposed emotions experienced by other people.

Language: a verbal message can convey accurate information about an individual's mental and emotional state. Specifically, a verbal message can immediately corroborate or disprove what one might think, believe and feel (assuming that we deem its content trustworthy), allowing us to update our knowledge about our interlocutor in a direct and generally effective way.

Non-verbal communication: if the verbal information is untrue, ambiguous or incomplete, the recognition of non-verbal information, that it often provided automatically and unconsciously by our interlocutor, may support or weaken the content conveyed through the verbal channel. Non-verbal information can also be sent intentionally, for example, by means of conventional gestures, which differ in distinct cultures. Even body language, such as a person's posture or pace, can provide precious information to improve the interpretation of their mental states.

G***aze direction***: adults use their eyes constantly, both to send messages to their interlocutors and to understand other people's internal states. For instance, if I notice that the other person is looking away while I am speaking, I might think that his attention is focused on something else. Joint information about gaze direction and facial expressions can provide even more useful information. For example, if I notice a frightened expression on a person's face, I could imagine that something dangerous may come from the direction in which they are looking, suggesting I run away in the opposite direction. Finally, knowing what the other is watching allows us to understand what the other person knows (i.e., its knowledge base). If an individual crossing the street is not looking at a car moving towards him, I can infer that he or she is not aware of it, and I can predict that they will not move to avoid it.

Additional personality traits: as has already been said, personality traits act as a bridge between the stereotypical characteristics of people and the mental states that we attribute to them. However, whenever we attribute a specific personality trait to someone, we also activate other traits that, stereotypically, our brain has connected to it. A good system capable of representing and attributing mental states to others should be able to appropriately represent and manage all the individual's personality traits. A non-hierarchical structure could be used to link all the different attributes to each other, causing the activation of one of them to influence the activation of the others. The latter could

 doi:10.20944/preprints202605.1676.v1

be hard to predict only based on the physical characteristics of the person (e.g. if they are fat they will likely be lazy; moreover, if they are lazy they will also probably be unreliable, and so on). In turn, these linked personality traits could influence the pre-activation of new mental states. This system could also learn to create *ad hoc* configurations of personality traits activation for specific individuals (e.g. Luca is lazy and sociable, Marco is willing and aggressive).

Additional personal attributes: in addition to the perceptual characteristics described above, a person can also be represented by other semantic features that, similarly to stereotypes, may be associated with the personality traits of the individual. For instance, I might know that a person "studies languages", "studied at Harvard", or "is a librarian", or I might be in possession of other, more complex pieces of information such as the fact that they "love animals" or "are not religious". All of this information will consequently activate inferential processes based on previous experiences, which consist simply in associating these representations with elements more related to the person's personality traits, as shown above. A system designed to consider such sources of information about an individual, which are more complex and not intrinsically connected to their internal states, should have the ability to manipulate both symbolic and sub-symbolic representations, as well as complex verbal aspects and the relationship between the different cues.

6. Theory of Mind Simulations via Reconfigurable Deep Neural Networks

The overall conceptual model advanced in this paper is illustrated in Figure 3. The available stimuli (e.g. identity of people, presence of objects, perceived actions, top) drive the recognition of the class of agent ('stereotype', left) observed, in turn shaping the structure of the simulation in the form of a hierarchical predictive structure (middle, cfr. Section 3) designed to explain their behaviour. The less similar the other is to any known stereotype, or the fewer input information is available, the more the ToM simulation is driven by a simulation of the self, which amounts to a default state associated with lack of knowledge, as described in Sections 3.7 and 3.8. Note that despite the visual nature of the stimuli in the example, the latter can be of any form (e.g., auditive or verbal, as recalled in Section 5).

Note that, as argued above, simpler actions forming part of a more complex behaviour pattern or activity can also be associated with shorter-lived desires and intentions, so that the depicted diagram quite underestimates the true complexity of the description.

To conclude this work, we wish to outline a possible suitable implementation of our conceptual Theory of Mind model in terms of a novel class of '*composable' deep neural networks*, as the building block of a functioning theory of mind for AI incorporating learning. The setting builds on the most recent related work in the area of reconfigurable artificial neural network (She et al., 2014; Andreas et al., 2016a; Andreas et al., 2016b), and is general enough to accommodate various distinct models within the class of hierarchical structures discussed in this paper.

6.1. Artificial Neural Networks

Artificial Neural Networks (ANNs) are computational models inspired by the biological neural networks that constitute animal and human brains. ANNs are typically composed by layers of artificial neurons, taking a number of signals $x_1, \dots, x_n$ as input and producing a single output quantity, y. The latter is the result of applying a non-linear *activation function* f to a weighted linear combination of the inputs $x_1, \dots, x_n$, each of which is multiplied by a weight factor w_i:

$y = f\left(\sum_{i=1,\dots,n} w_i x_i\right)$.

Common activation functions are the *sigmoid*

$$f(x) = \frac{1}{1 + e^{-x}}$$

and the rectifier linear unit (ReLU) ones:

$$f(x) = \max(0, x).$$

 doi:10.20944/preprints202605.1676.v1

Artificial neurons can be arranged into multiple *layers*, so that neurons can only accept inputs from the previous layer and feed their outputs to the following layer, but do not interact with fellow neurons on the same layer. ANNs can be graphically represented as collection of nodes representing the artificial neurons, and of connections (depicted as arrows) linking pairs of neurons, each carrying a given weight w_{ij}. An example of such architecture is the *multilayer perceptron* (MLP, Rosenblatt, 1961), depicted in Figure 4.

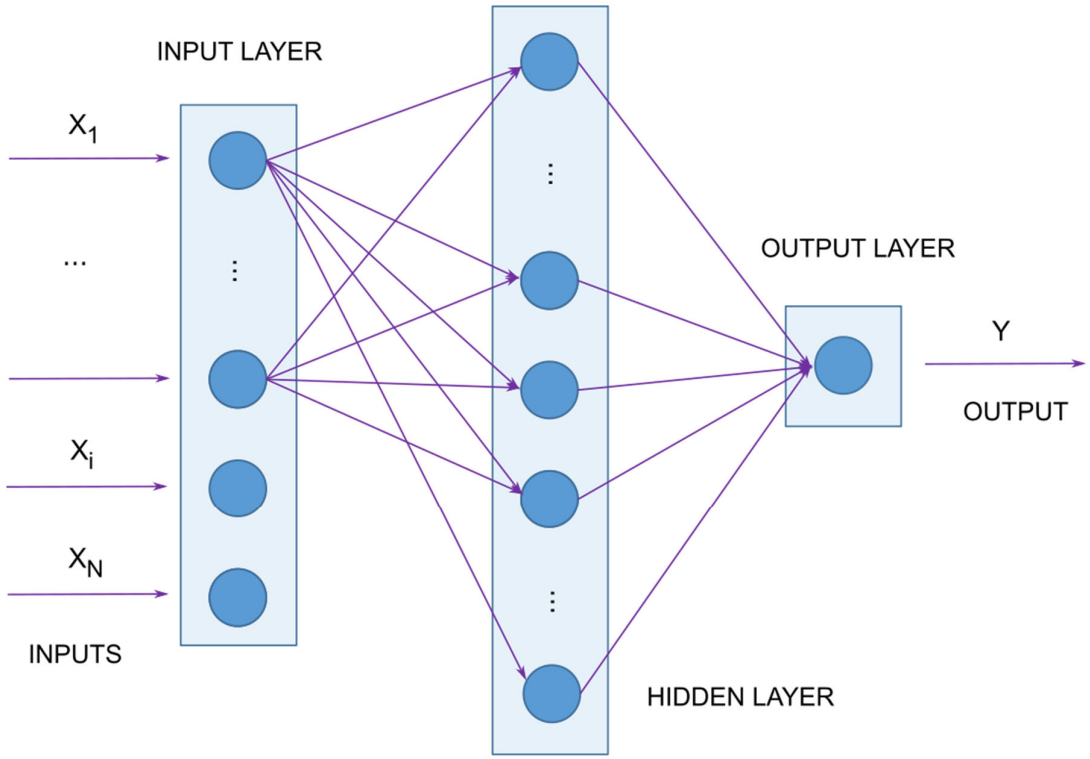

Figure 4. Example of multilayer perceptron (MLP). The hidden layers are those distinct from both the input and the output ones. Not all connections are drawn for sake of clarity.

MLPs are trained in a supervised fashion (i.e., by showing the network a number of examples of input vectors together with the corresponding target values) using *backpropagation* (Rumelhart et al., 1986). In backpropagation, the weights of all the connections in a neural network are computed backward, layer by layer, starting from the last one, by calculating the gradient (or vector of derivatives) of the error (discrepancy) between the output produced by the network and the desired one (which, in a supervised setting, is known for the training datapoints).

Note that in classical ANN architectures each neuron is connected to all the neurons in both the previous and the following layer. This makes it impractical to design networks with an extended number of layers, as gradients in the backpropagation procedure tend to "vanish" because of the number of variables involved.

6.2. Deep Neural Networks

An artificial neural network is termed *deep* whenever it contains a large number of (hidden) layers. Once again the concept is inspired by what happens in the primary visual cortex of animals, in which visual inputs are processed at multiple levels of abstractions, from very simple edge-like patterns in the lower strata to semantically meaningful concepts (e.g. faces) in the upper ones. This is achieved thanks to a specialised connectivity structure in which all neurons are not connected to all others, but (as is the case for *convolutional neural networks*, CNN) follow a local pattern.

 doi:10.20944/preprints202605.1676.v1

For instance, in CNNs the main structures are convolutional layers in which each local patch in the output of the previous layer (*feature map*) is processed using convolution[1]. Crucially, all local patches at the same level of the hierarchy are processed with the same convolution kernel (in network terminology, they "share weights"). Convolutional layers are alternated with *max pooling* ones which summarise each local patch with a single real number, and non-linear ones associated with activation functions. As the number of network parameters is drastically reduced, all these peculiarities make it possible to train the resulting deep network, achieving consistent state of the art results.

6.3. Logical Inferences via Artificial Neural Networks

A number of recent studies (Cohen et al., 2017; Liu et al., 2016; Fadja et al., 2017) show, in particular, that (probabilistic) logical inferences can be implemented in the form of artificial neural networks. This is kind of inference associated with reasoning of the kind: "if the person is running while carrying a knife, then they are likely to want to assault somebody". *Modal logic* inferences (allowing quantifiers such as usually, sometimes and so on) are possible as well: "Usually, when I am angry I have a desire of hitting a person in the face". This provides evidence that the implications between mental states in our proposed hierarchical predictive structure (the purple arrows in Figure 3) can indeed be implemented in the form of appropriate (deep) neural networks (Fadja et al., 2017).

As argued in Section 3, however, cognitive economics arguments, as well as hard neuroscientific evidence, support the view that learning a separate simulation network (for instance, by assembling a number of such probabilistic inference networks into an overall one, modelling the reasoning and mental states of the agent under consideration) for each individual or even for each class of agents is completely impractical. In our lifetime, we are likely to have to deal with several thousand individuals, grouped into possible hundreds of different, often overlapping classes. The same can be expected of an intelligent machine endowed with Theory of Mind capabilities.

6.4. Reconfigurable Simulations

A possible solution to this conundrum is the design of a framework in which ToM simulations for separate agent classes or 'stereotypes' (e.g. children, construction workers, and so on) are created when required (i.e., when the related agent class is perceived) by flexibly assembling basic artificial neural network 'modules', i.e., by linking smaller networks together to generate the overall simulation. In our proposed model, these modules are charged with implementing relations between mental states (the purple arrows in Figure 3), and can be implemented as deep neural networks expressing probabilistic logic inferences. The concept is illustrated in Figure 5.

In principle, such modules can be individually trained based on general experience, whereas the most appropriate structure and weights for the connections linking the modules together can be tailored to each specific stereotype. A more general framework in which the parameters of the base modules are also updated to better fit a specific agent class can also be envisaged.

As noted in Section 3.8, a pre-arranged simulation of the self can also run in parallel, its outcomes integrated with those produced by the stereotype-driven simulation in a manner that flexibly adapts to the amount of information available. Furthermore, following our observations that the distinction between mental states such as "belief" and "intention" is rather fuzzy, and that mental states spanning different temporal horizons can/should be present at different levels of the hierarchy, the composable deep simulation proposal allows multiple replica of the basis modules to be part of the simulation. In fact, modules are not even constrained to form a strictly hierarchical structure but the presence of loops between different parts of the simulation model is allowed.

[1] https://en.wikipedia.org/wiki/Convolution

 doi:10.20944/preprints202605.1676.v1

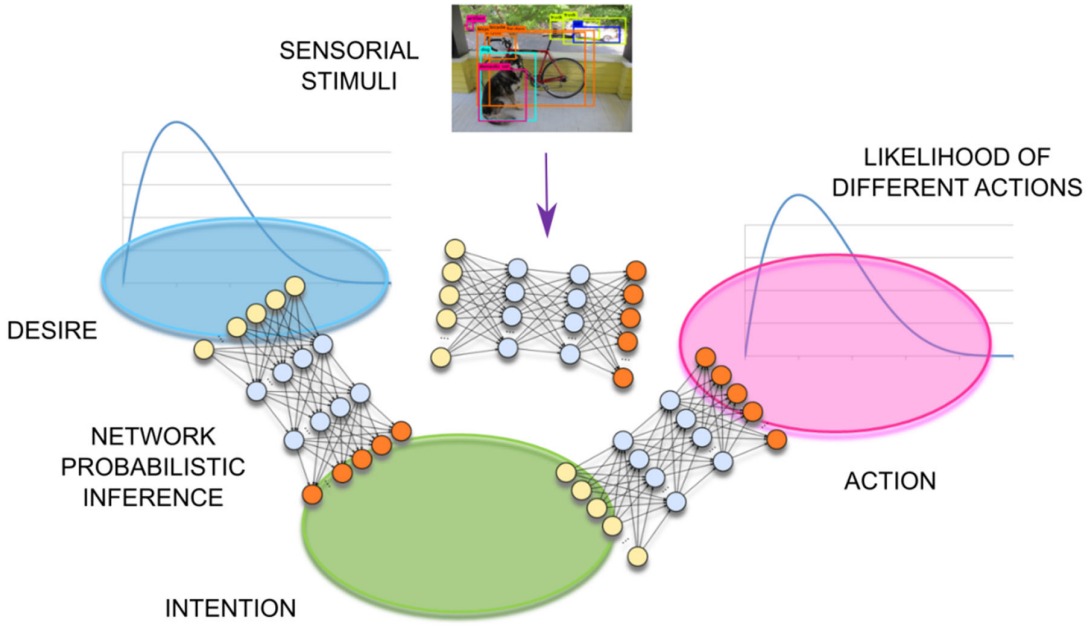

Figure 5. The proposed predictive hierarchical model can be implemented in the form of composable artificial neural networks in which the available stimuli (not necessarily of a visual type) drive both inferences on the likelihood of the various mental states involved in the generation of the action observed (the blue plots), and a bespoke simulation of how such mental states interact in the observed agent, in the form of deep neural networks (shown in the diagram) implementing probabilistic inference between the various mental states involved. Such a structure can be replicated at multiple levels of the hierarchy.

In our proposed framework the optimal topology (structure) for the overall simulation network is learned by rewarding configurations accurately predicting the observed behaviour of the agent concerned, and by penalising those leading to inaccurate predictions via a mechanism known in machine learning as *reinforcement learning* (Kaelbling, L.P. et al., 1996).

As in humans, learning is ongoing and triggered by the recognition of classes of agents or scene elements in the sensory data. Whenever a known 'stereotype' is identified, its internal replica is activated to run simulations on their possible courses of action and predict their future behaviour (downward arrow in Figure 3), concurrently with the long-standing simulation of the self. As a bonus, the observed behaviour may also be explained in terms of the agent's reasoning process (upward arrow), in a process potentially constituting a significant step forward in the pursue of "explainable" artificial intelligence.

In perspective, the proposed framework can be further extended by allowing predictions on the next action performed by an agent, based on their mental states as guessed by the appropriate simulation. Another interesting direction of research is the incorporation of temporal dynamics in the reasoning process, to reflect the fact that mental states are not static quantities but are characterised by their own, often quite independent, dynamics, which can in turn be useful to model to provide more accurate estimates and predictions. More details will follow in a subsequent technical paper.

7. Conclusions

In this paper we proposed a new conceptual model of Theory of Mind, which rests on both neuroscientific and psychological evidence. The model builds on the notion of hierarchical predictive structure, and incorporates as crucial elements stereotypes, as a way to efficiently make inferences about other people or agent, and internal simulation as a mechanism allowing the observer to make inferences on mental states and future behaviour even in absence of sensorial stimuli.

 doi:10.20944/preprints202605.1676.v1

We also outlined a possible computational implementation of such a model in terms of reconfigurable (deep) neural networks, whose structure is adjusted as a function of the available information in order to provide the most suitable simulation of the external agent whose states and intention need to be predicted. The topology of the "best" simulation at any given time is learned from experience via reinforcement learning, by rewarding simulations whose prediction are in accordance with the observed behaviour, and penalising those which are not.

In the near future we will build on our current work in deep learning for human behaviour understanding to implement and test the proposed computational model. Success of this programme of work will lay the foundations for the creation, among others, of autonomous vehicles able to negotiate complex road situations involving humans. Next-generation robotic assistant surgeons could be envisaged with the ability to understand what the main surgeon is doing and foresee their future intentions, in order to best assist them. Empathic healthcare is becoming a priority for the healthcare providers worldwide, especially when dealing with autism and other similar conditions. The new ToM model we propose here may improve the efficacy of psychological treatments, such as cognitive behavioural therapy or mindfulness. A new generation of 'bots' (e.g. for customer service or financial advice) able to interact more effectively and empathetically with humans would be possible.

This work has also the potential to impact on the current debate on moral AI, helping machines make ethical, human-like decisions in critical situations. In the longer term robotic companions for disabled people can be imagined, based on the capabilities we aim to develop.

Funding: This work was funded by the Leverhulme Trust, under the Research Programme Grant RPG-2019-243.

Authors Contributions Statement: AM and FC contributed to the concepts presented in the manuscript. AM and FC wrote the manuscript.

References

Ames, D. R. (2004a). Inside the mind reader's tool kit: projection and stereotyping in mental state inference. *Journal of personality and social psychology*, 87(3), 340-353.

Ames, D. R. (2004b). Strategies for social inference: a similarity contingency model of projection and stereotyping in attribute prevalence estimates. *Journal of personality and social psychology*, 87(5), 573-585.

Andreas, J. et al (2016a). Learning to Compose Neural Networks for Question Answering. arXiv:1601.01705v4.

Andreas, J. et al (2016b). Neural module networks. arXiv:1511.02799v3.

Angelucci, A., Levitt, J. B., Walton, E. J., Hupe, J. M., Bullier, J., and Lund, J. S. (2002). Circuits for local and global signal integration in primary visual cortex. *Journal of Neuroscience*, 22(19), 8633-8646.

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

Baker, C., Saxe, R., and Tenenbaum, J. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. *Proceedings of the annual meeting of the cognitive science society*, 33(33).

Bello, P. (2012). Cognitive Foundations for a Computational Theory of Mindreading. *Adv. in Cognitive Systems*, 1, 59-72.

Berthoz, S., Armony, J. L., Blair, R. J. R., and Dolan, R. J. (2002). An fMRI study of intentional and unintentional (embarrassing) violations of social norms. *Brain*, 125(8), 1696-1708.

Bosse, T., Memon, Z. A., and Treur, J. (2007). A two-level BDI-agent model for theory of mind and its use in social manipulation. *Proceedings of the AISB 2007 Workshop on Mindful Environments*, 4, 335-342.

Brass, M., Schmitt, R. M., Spengler, S., and Gergely, G. (2007). Investigating action understanding: inferential processes versus action simulation. *Current Biology*, 17(24), 2117-2121.

Burnham, D. K., and Harris, M. B. (1992). Effects of real gender and labeled gender on adults' perceptions of infants. *The Journal of genetic psychology*, 153(2), 165-183.

Clements, W. A., and Perner, J. (1994). Implicit understanding of belief. *Cognitive development*, 9(4), 377-395.

Clements, W. A., Rustin, C. L., and McCallum, S. (2000). Promoting the transition from implicit to explicit understanding: A training study of false belief. *Developmental Science*, 3(1), 81-92.

 doi:10.20944/preprints202605.1676.v1

Cloutier, J., Gabrieli, J. D., O'young, D., and Ambady, N. (2011). An fMRI study of violations of social expectations: when people are not who we expect them to be. *NeuroImage*, 57(2), 583-588.

Cohen, W.W., Yang, F. and Mazaitis, K. R. (2017). Tensorlog: Deep learning meets probabilistic dbs. *arXiv preprint arXiv:1707.05390*.

Condry, J. C., and Ross, D. F. (1985). Sex and aggression: The influence of gender label on the perception of aggression in children. *Child development*, 56(1), 225-233.

Costantini, M., Galati, G., Ferretti, A., Caulo, M., Tartaro, A., Romani, G. L., and Aglioti, S. M. (2005). Neural systems underlying observation of humanly impossible movements: an fMRI study. *Cerebral Cortex*, 15(11), 1761-1767.

de Weerd, H., Verbrugge, R., and Verheij, B. (2014). Agent-based models for higher-order theory of mind. *Advances in Social Simulation*, 229, eds. B. Kamiński and G. Koloch (Berlin, Heidelberg: Springer), 213-224.

Doris, J. M. (2002). Lack of character: Personality and moral behavior. Cambridge University Press.

Fadja, A.N., Lamma, E. and Riguzzi, F. (2017). Deep Probabilistic Logic Programming. *PLP@ ILP*, 3-14.

Felleman, D. J., and Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex*, 1(1), 1-47.

Fritzke, B. (1994). A growing neural gas network learns topologies. Proceedings of NIPS'94.

Gmytrasiewicz, P., Moe, G., and Moreno, A. (2015). MARTHA Speaks: Implementing Theory of Mind for More Intuitive Communicative Acts. *2015 AAAI Fall Symposium Series*.

Godfrey-Smith, P. (1991). Signal, decision, action. *The Journal of Philosophy*, 88(12), 709-722.

Goodman, G. S., Golding, J. M., Helgeson, V. S., Haith, M. M., and Michelli, J. (1987). When a child takes the stand: Jurors' perceptions of children's eyewitness testimony. *Law and Human Behavior*, 11(1), 27.

Gopnik, A. (2003). "The theory theory as an alternative to the innateness hypothesis," In *Chomsky and his critics*, eds. L. M. Antony, and N. Hornstein (Oxford: Blackwell Publishing), 238-254.

Gopnik, A., and Wellman, H. M. (1994). "The theory theory," In *Mapping the mind: Domain specificity in cognition and culture*, eds. L. Hirschfeld and S. Gelman (Cambridge: Cambridge University Press), 257-293.

Gopnik, A., and Wellman, H. M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological bulletin*, 138(6), 1085.

Grossman, E. D., Battelli, L., and Pascual-Leone, A. (2005). Repetitive TMS over posterior STS disrupts perception of biological motion. *Vision research*, 45(22), 2847-2853.

Grossman, E. D., Jardine, N. L., and Pyles, J. A. (2010). fMR-adaptation reveals invariant coding of biological motion on human STS. *Frontiers in human neuroscience*, 4, 15.

Harbers, M., Van den Bosch, K., and Meyer, J. J. (2012). Modeling agents with a theory of mind: Theory-theory versus simulation theory. *Web Intelligence and Agent Systems: An International Journal*, 10(3), 331-343.

Harris, P. L. (1992). From simulation to folk psychology: The case for development. *Mind & Language*, 7(1-2), 120-144.

Hesslow, G. (2002). Conscious thought as simulation of behaviour and perception. *Trends in cognitive sciences*, 6(6), 242-247.

Hiatt, L. M., and Trafton, J. G. (2010). A cognitive model of theory of mind. *Proceedings of the 10th International Conference on Cognitive Modeling*, 91-96.

Hiatt, L. M., Harrison, A. M., and Trafton, J. G. (2011). Accommodating human variability in human-robot teams through theory of mind. *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, 22(3), 2066.

Hochstein, S., and Ahissar, M. (2002). View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron*, 36(5), 791-804.

Hohwy, J., and Palmer, C. (2014). "Social cognition as causal inference: Implications for common knowledge and autism," In *Perspectives on social ontology and social cognition*, ed. M. Gallotti and J. Michael (Dordrecht: Springer), 167-189.

Hudson, M., Nicholson, T., Ellis, R., and Bach, P. (2016). I see what you say: Prior knowledge of other's goals automatically biases the perception of their actions. *Cognition*, 146, 245-250.

Jastorff, J., Clavagnier, S., Gergely, G., and Orban, G. A. (2010). Neural mechanisms of understanding rational actions: middle temporal gyrus activation by contextual violation. *Cerebral Cortex*, 21(2), 318-329.

 doi:10.20944/preprints202605.1676.v1

Juang, C. F., and Lin, C. T. (1998). An On-Line Self-Constructing Neural Fuzzy Inference Network and Its Applications. *IEEE Fuzzy Systems*, 6(1), 12-32.

Kable, J. W., and Chatterjee, A. (2006). Specificity of action representations in the lateral occipitotemporal cortex. *Journal of Cognitive Neuroscience*, 18(9), 1498-1517.

Kaelbling, L. P., Littman, M. L., Moore, A. W. (1996). Reinforcement Learning: A Survey. *Journal of Artificial Intelligence Research*, 4, 237–285.

Kilner, J. M., Friston, K. J., and Frith, C. D. (2007). Predictive coding: an account of the mirror neuron system. *Cognitive processing*, 8(3), 159-166.

Kim, K. J., and Lipson, H. (2009). Towards a theory of mind in simulated robots. In *Proceedings of the 11th Annual Conference Companion on Genetic and Evolutionary Computation Conference: Late Breaking Papers*, 2071-2076.

Koster-Hale, J., and Saxe, R. (2013). Theory of mind: a neural prediction problem. *Neuron*, 79(5), 836-848.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521, 436-444.

Liu, Q., Jiang, H., Evdokimov, A., Ling, Z. H., Zhu, X., Wei, S. and Hu, Y. (2016). Probabilistic reasoning via deep learning: Neural association models. *arXiv preprint arXiv:1603.07704*.

Low, J., and Perner, J. (2012). Implicit and explicit theory of mind: State of the art. *British Journal of Developmental Psychology*, 30(1), 1-13.

Malle, B. F. (1999). How people explain behavior: A new theoretical framework. *Personality and social psychology review*, 3(1), 23-48.

Mason, M. F., Cloutier, J., and Macrae, C. N. (2006). On construing others: Category and stereotype activation from facial cues. *Social Cognition*, 24(5), 540-562.

Mitchell, J. P., Banaji, M. R., and MacRae, C. N. (2005). The link between social cognition and self-referential thought in the medial prefrontal cortex. *Journal of cognitive neuroscience*, 17(8), 1306-1315.

Mitchell, J. P., Macrae, C. N., and Banaji, M. R. (2006). Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron*, 50(4), 655-663.

Mueller-Johnson, K., Toglia, M. P., Sweeney, C. D., and Ceci, S. J. (2007). The perceived credibility of older adults as witnesses and its relation to ageism. *Behavioral sciences & the law*, 25(3), 355-375.

Murphy, P. C., and Sillito, A. M. (1987). Corticofugal feedback influences the generation of length tuning in the visual pathway. *Nature*, 329(6141), 727.

Oberman, L. M., and Ramachandran, V. S. (2007). The simulating social mind: the role of the mirror neuron system and simulation in the social and communicative deficits of autism spectrum disorders. *Psychological bulletin*, 133(2), 310.

Ondobaka, S., Kilner, J., and Friston, K. (2017). The role of interoceptive inference in theory of mind. *Brain and cognition*, 112, 64-68.

Palmer, C. J., Seth, A. K., and Hohwy, J. (2015). The felt presence of other minds: Predictive processing, counterfactual predictions, and mentalising in autism. *Consciousness and Cognition*, 36, 376-389.

Pelphrey, K.A., and Vander Wyk, B.C. (2011). "Functional and neural mechanisms for eye gaze processing." In *Oxford Handbook of Face Perception*, eds. A. Calder, G. Rhodes, M. Johnson, and J. Haxby, (Oxford, UK: Oxford University Press), 591–604.

Pelphrey, K. A., Mitchell, T. V., McKeown, M. J., Goldstein, J., Allison, T., and McCarthy, G. (2003). Brain activity evoked by the perception of human walking: controlling for meaningful coherent motion. *Journal of Neuroscience*, 23(17), 6819-6825.

Pelphrey, K. A., Morris, J. P., and Mccarthy, G. (2004). Grasping the intentions of others: the perceived intentionality of an action influences activity in the superior temporal sulcus during social perception. *Journal of cognitive neuroscience*, 16(10), 1706-1716.

Pfeifer, J. H., Lieberman, M. D., and Dapretto, M. (2007). "I know you are but what am I?!": neural bases of self- and social knowledge retrieval in children and adults. *Journal of Cognitive Neuroscience*, 19(8), 1323-1337.

Polceanu, M., and Buche, C. (2014). Towards a theory-of-mind-inspired generic decision-making framework. *arXiv preprint arXiv:1405.5048*.

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

 doi:10.20944/preprints202605.1676.v1

Pynadath D.V., Rosenbloom P.S., nad Marsella S.C. (2014) "Reinforcement Learning for Adaptive Theory of Mind in the Sigma Cognitive Architecture," In *Artificial General Intelligence*, eds. B. Goertzel, L. Orseau, J. Snaider AGI 2014. Lecture Notes in Computer Science, vol 8598. Springer, Cham

Rosch, E. (2002). "Principles of categorization," In *Foundations of cognitive psychology: Core readings*, ed. D. J. Levitin (MIT Press), 251–270.

Rosenblatt, F. (1961). Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Spartan Books, Washington DC

Rumelhart, D. E., Hintonm G. E., and Williams, R. J. (1986). "Learning Internal Representations by Error Propagation," In *Parallel distributed processing: Explorations in the microstructure of cognition*, 1: Foundation, eds. David E. Rumelhart, James L. McClelland, and the PDP research group (Cambridge, MA, United States: MIT Press).

Ruffman, T., Garnham, W., Import, A., and Connolly, D. (2001). Does eye gaze indicate implicit knowledge of false belief? Charting transitions in knowledge. *Journal of experimental child psychology*, 80(3), 201-224.

Sagar, H. A., and Schofield, J. W. (1980). Racial and behavioral cues in black and white children's perceptions of ambiguously aggressive acts. *Journal of personality and social psychology*, 39(4), 590.

Saha, S., Singh, G., Sapienza, M., Torr, P. and Cuzzolin, F. (2016). Deep Learning for Detecting Multiple Space-Time Action Tubes in Videos. *Proceeding of the British Machine Vision Conference (BMVC'16)*.

Saha, S., Singh, G., and Cuzzolin, F. (2017). AMTnet: Action-Micro-Tube regression by end-to-end trainable deep architecture. *Proceeding of the International Conference on Computer Vision (ICCV'17)*.

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

Saygin, A. P., Chaminade, T., Ishiguro, H., Driver, J., and Frith, C. (2011). The thing that should not be: predictive coding and the uncanny valley in perceiving human and humanoid robot actions. *Social cognitive and affective neuroscience*, 7(4), 413-422.

Scassellati, B. (2002). Theory of mind for a humanoid robot. *Autonomous Robots*, 12(1), 13-24.

Schacter, D. L., Addis, D. R., and Buckner, R. L. (2008). Episodic simulation of future events. *Annals of the New York Academy of Sciences*, 1124(1), 39-60.

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

She, Y., He, Y., and Wu, D. (2014). Learning Topology and Dynamics of Large Recurrent Neural Networks. arXiv:1410.1174.

Sherman, S. M., and Guillery, R. W. (1998). On the actions that one nerve cell can have on another: distinguishing "drivers" from "modulators". *Proceedings of the National Academy of Sciences*, 95(12), 7121-7126.

Shibata, H., Inui, T., and Ogawa, K. (2011). Understanding interpersonal action coordination: an fMRI study. *Experimental brain research*, 211(3-4), 569-579.

Shultz, S., Lee, S. M., Pelphrey, K., and McCarthy, G. (2010). The posterior superior temporal sulcus is sensitive to the outcome of human and non-human goal-directed actions. *Social cognitive and affective neuroscience*, 6(5), 602-611.

Si, M., and Marsella, S. C. (2014). Encoding theory of mind in character design for pedagogical interactive narrative. *Advances in Human-Computer Interaction*, 10. doi: 10.1155/2014/386928

Singh, G., Saha, S., Sapienza, M., Torr, P., and Cuzzolin, F. (2017). Online Real-time Multiple Spatiotemporal Action Localisation and Prediction on a Single Platform. *Proceeding of the International Conference on Computer Vision (ICCV'17)*.

Singh, G., Saha, S., and Cuzzolin, F. (2018). TraMNet - Transition Matrix Network for Efficient Action Tube Proposals. *Proceeding of the Asian Conference on Computer Vision (ACCV'18)*.

Thoermer, C., Sodian, B., Vuori, M., Perst, H., and Kristen, S. (2012). Continuity from an implicit to an explicit understanding of false belief from infancy to preschool age. *British Journal of Developmental Psychology*, 30(1), 172-187.

Wang, B., Low, J., Jing, Z., and Qinghua, Q. (2012). Chinese preschoolers' implicit and explicit false-belief understanding. *British Journal of Developmental Psychology*, 30(1), 123-140.

Westra, E. (2019). Stereotypes, theory of mind, and the action–prediction hierarchy. *Synthese*, 196, 2821-2846.

 doi:10.20944/preprints202605.1676.v1

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

Young, L., Dodell-Feder, D., and Saxe, R. (2010). What gets the attention of the temporo-parietal junction? An fMRI investigation of attention and theory of mind. *Neuropsychologia*, 48(9), 2658-2664.