## Problem, Background and Rationale

Emerging applications of artificial intelligence (AI) are highlighting the limitations of established machine learning (ML) approaches in situations involving humans. Smart cars, for example, 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. Currently, the automated recognition of present behaviour builds on the success of deep learning [1]. The latter is based on artificial neural networks with an extended number of layers and architectures suitably designed to support learning at various levels of abstraction, for instance by forcing computations to take place locally and using shared connection weights. Deep networks can efficiently identify motion patterns in streaming videos, such as those associated with human actions [2,3]. Motion patterns, however, can be deceiving as humans can suddenly change their minds based on their own mental dynamics and things they see around them. In our example, children previously walking on the pavement towards school may spot an ice cream van on the other side of the road, and suddenly decide to cross the road to get an ice cream. No predictive system purely functioning based on past observed motion could be able to cope with this sort of unpredictability. In fact, human beings can predict future behaviour even when no motion is present, just by quickly assessing the 'type' of person involved (e.g. an elderly person standing in a hallway is likely to take the elevator rather than the stairs) and the scene. Theory of Mind (ToM) [4] capabilities, i.e., the ability to 'guess' another agent's mental states, are then crucial to the development of a next generation artificial intelligence designed to understand humans and share a common environment. Indeed, in absence of such a change of paradigm, none of the exciting applications of AI now considered to be within reach will be possible. In a mutually beneficial process, computational models developed within AI could in turn provide new insights into the way ToM works in humans. While a number of anatomical structures have been identified as correlated with Theory of Mind functionalities in the human brain, there remains a lack of proper, detailed modelling of the way they work, as well as an understanding of how reasonably well understood mechanisms on a cellular level translate into higher-level cognitive functions.

## State of the Art

***Within cognitive neuroscience*** 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. The theory-theory (TT) [5] paradigm supports the existence of a set of rules humans possess regarding human mind functioning. According to supporters of this point of view [6,7], 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. Problematic about TT is its assumption that an individual generates, and stores, a very large number of theories about other people and their behaviour. Such a system would be wasteful, whereas the principle of cognitive economics supports the tendency to obtain the most information with the least amount of cognitive resources [8]. The simulation-theory [9] paradigm, instead, defends a simulation process which consists of taking someone else's perspective to understand their reasoning. An individual would, under this hypothesis, activate 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?". Importantly, 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 [10].

***Within artificial intelligence*** work towards equipping machines with some degree of ToM abilities mainly takes place in the multi-agent systems domain. An intelligent agent is an entity that autonomously tries to reach goals. To this extent Belief-Desire-Intention (BDI) models [11], which attempt to model these three basic mental states, are so far the dominant force in computational Theory of Mind. As they lack the ability to learn from experience and heavily rely on prior knowledge, however, they have been much criticised. Efforts have also been made from other theoretical perspectives. For instance, as they can model goals, beliefs and dynamics, partially observable Markov decision processes have been proposed as the basis for a Bayesian model of ToM [12]. Recent computational approaches also span multi-agent reinforcement learning [13], evolutionary robotics [14], and game theory [15]. Overall, however, computational ToM models based on machine learning remain in their infancy. A solution, we argue here, may be provided by networks whose topology dynamically adapts to the data [16,17], a concept recently extended to recurrent neural networks [18]. Models which assemble neural networks starting from a collection of composable modules have recently been proposed for query answering from images [19]. Network layouts are there sampled and assessed, while network parameters are jointly learned via reinforcement learning [20].

## Hypotheses and Objectives

Our **fundamental hypothesis (1)** is that the integration of neuroscience and machine learning could enable significant advances in both fields, by allowing: (i) within neuroscience, the formulation and empirical validation of computational Theory of Mind models in humans leveraging current frontier efforts in ML and AI, forging towards a more detailed understanding of these functionalities in the human brain; (ii) within artificial intelligence, the development of machine Theory of Mind models informed by the latest neuroscientific studies and evidence. The latter will be able to go beyond both human prediction methods based on simple pattern recognition, and rule-based approaches which neglect the learning component, to allow prediction in complex, human-centred scenarios. Our **second hypothesis (2)** is that a simulation theory of mind standpoint, which describes human ability to understand other people as the result of an internal 'simulation' of their mental processes, is the most suitable to bring together frontier machine learning and neuroscience in a radically new approach to the problem. Finally (**hypothesis (3)**), we claim that efficient simulations of human behaviour can be implemented using reconfigurable (deep) neural networks, able to change topology to adapt to different agent classes and scenes. Overall:

***Our Goal is to propose a simulation theory of mind approach in machines based on novel neuroscientific models, implemented by a new generation of reconfigurable deep networks.***

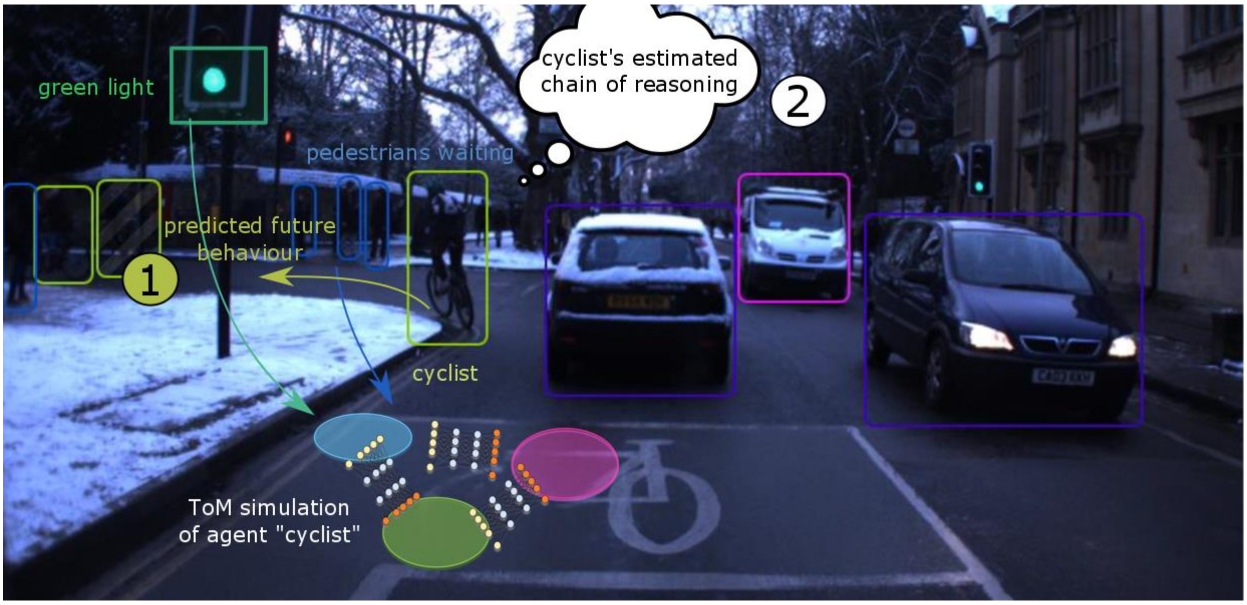

green light
pedestrians waiting
cyclist's estimated chain of reasoning
2
predicted future behaviour
1
cyclist
ToM simulation of agent "cyclist"

**Figure 1. Concept**. The detection of an intelligent agent (e.g., a cyclist, in the yellow bounding box) triggers a ToM simulation of its reasoning processes, in terms of reconfigurable neural networks (drawn below the cyclist). The simulation is also driven by relevant scene elements (e.g., the green light, the presence of pedestrians before him, in the blue boxes). The simulation outputs both (1) the likely predicted behaviour (the cyclist will turn left, but slowly to avoid collisions) and (2) a possible explanation for their actions (e.g., there is snow on the ground so it might be slippery), depicted in the comic bubble.

The concept is depicted in Fig.1. The proposed ToM simulations will allow, e.g., a smart car to predict future human (and other cars') behaviour, as well as provide an explanation for the observed behaviour.

## Significance and Timeliness

This project aims at laying the stepping stone for a new paradigm in artificial intelligence, one in which models and algorithms are by design capable of understanding human reasoning.

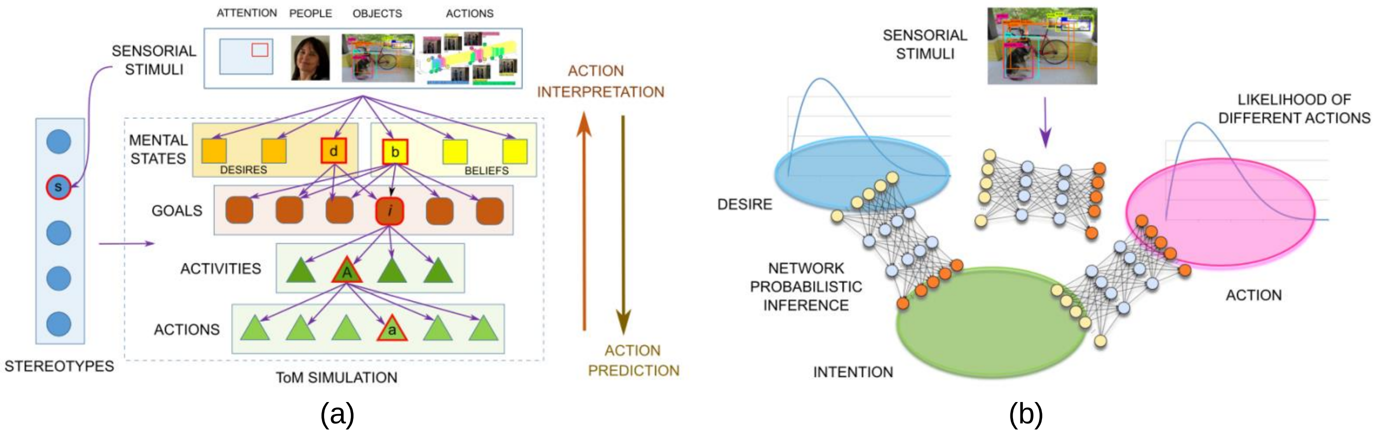
Success will lay the foundations for the creation, among others, of autonomous vehicles able to negotiate complex road situations in mixed scenarios [21] involving humans as well as human-driven and autonomous vehicles. In the field of healthcare, next-generation robotic assistant surgeons [22] could be developed 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 NHS, especially when dealing with autism and similar conditions. The new ToM models we propose may improve the efficacy of psychological treatments, such as cognitive behavioural therapy or mindfulness. In customer service and the financial sector, a new generation of 'bots' able to interact more effectively and empathetically with humans would be possible. In the longer term robotic companions for disabled people can be imagined, based on the capabilities we aim to develop. From a more foundational point of view, this work would also impact on the current debate on moral AI [23], helping machines make ethical, human-like decisions in critical situations. Trust in artificial intelligence will be reinforced, as people realise machines capable of truly reasoning like them become more widespread.

## Track record and vision

Cuzzolin and Sahakian are both world-leading researchers in their respective fields: Sahakian in cognitive neuroscience and psychology, including both non-emotional ('cold') and emotional cognition, Cuzzolin in artificial intelligence and machine learning, with a focus on recognition and prediction of human behaviour. Cuzzolin's team, in particular, enjoys a leadership position in the detection of human action and activities [PI1-8], and is now pioneering the field of future action prediction [24-25]. Their original deep learning approach has topped all competitors on accuracy, while demonstrating better than real-time capabilities [PI5], but has also shown the limitations of human prediction approaches based on pure pattern recognition. Sahakian's expertise [CI1-4] is in the area of objective measurement of various forms of cognition and the neuroimaging of their underlying brain circuits. She is co-inventor of both the CANTAB (www.cambridgecognition.com) and the EMOTICOM computerised neuropsychological and Theory of Mind tests (https://www.frontiersin.org/articles/10.3389/fnbeh.2016.00025/full), which are used globally. Neither investigator would be able to carry out this project in isolation. The work's obvious cross-disciplinary nature only can produce substantial progress towards both a better modelling and understanding of theory of mind in humans, and a next generation artificial intelligence capable of understanding humans by design.

## Methodology

Composable deep nets have the potential to provide solid foundations for a simulation-theory approach to Theory of Mind in both machines and humans. The reasoning of various classes of complex agents can be flexibly simulated by dynamically rearranging the structure of the connections linking a base set of neural modules. The way the simulation adapts to agent class and scene is learned from experience via reinforcement learning. The work will be broken down into WorkPackages (WPs), described below.



(a) (b)

**Figure 2**. **(a)** ***Predictive hierarchical structures*** *allow us to both predict future behaviour in complex agents, and interpret the observed actions in terms of mental states.* **(b)**. *Computationally, such models can be implemented as* ***composable neural networks*** *in which the available stimuli 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 inferences between mental states.*

**WorkPackage 1 – Novel cognitive Theory of Mind models** [months 1-24]. In a combination of conceptual and experimental work, we will develop a new cognitive ToM model by building on the latest neuroscientific evidence, starting from the notion of predictive hierarchical structure [26] (Figure 2(a)). The latter is typical of some brain areas, such as the *primary visual cortex* [27]. Information flows from one level to another via both a top-down and a bottom-up process, allowing an individual to both infer the mental states generating a specific behaviour in others, and to predict future behaviour by ascribing to them suitable mental states. Brain areas exhibiting a similar functioning (amygdala, *superior temporal sulcus* [28], *temporo-parietal junction* [29]) are also involved in ToM processes. This takes place through three main mechanisms, whose joint activation is expected to improve prediction accuracy. Firstly, observed behaviours are associated with possible intentions. *Mirror neurons* in the brain seem to support this mental function, for they are activated both when an individual actually performs an action and when they see someone else doing the same [30]. Secondly, internal simulation [31] takes place which consists in projecting ourselves into a situation by recreating the scene in our mind. In a third process, humans put other people into 'stereotypical' categories characterised by rough personality traits, in order to speculate what they might do. Indeed, the ventral and dorsal regions of the *medial prefrontal cortex* specialise in responding to elements related to the 'self' and to the 'other' [32].

The proposed initial model is illustrated in Figure 2 (a). The available stimuli (top) drive the recognition of the class of agent ('stereotype', left) observed, in turn shaping the structure of the hierarchical ToM simulation (middle) designed to explain their behaviour. The latter assumes the form of a series of logical implications (purple arrows) between mental states (e.g. from desires to goals/intentions) and observed behaviours (from simple actions to more complex activities). The model will be experimentally validated as described in Task 3.1. Test results, in conjunction with any new psychological and neuroscientific evidence available, will drive the development of a series of ever more refined and realistic models.

**WorkPackage 2 – Computational models for a Machine Theory of Mind** [m. 1-30]. The final Theory of Mind model devised in WP1, as well as the successive approximations towards that, will be implemented by developing a novel class of 'composable' deep neural networks, as the building block of a functioning machine Theory of Mind incorporating learning as a crucial element. This concept will be iteratively refined in the light of WP1's work, in a continual feedback cycle between WP1 and WP2.

*Task 2.1 – Reconfigurable network architecture* [m. 1-24]. The setting builds on the most recent related work on neural module networks [18-20], and is general enough to accommodate various distinct psychological models. Deep ToM simulations for separate agent classes or 'stereotypes' (e.g. children, workers, etc) are created by flexibly assembling basic modules, rather than learned from scratch for each agent type, in accordance with the principle of cognitive economics. Our starting hypothesis is that such modules represent relations between mental states (purple arrows in Fig. 2(a)), implemented as deep networks (shown in (b)) expressing probabilistic logic inferences [33] (e.g. if the person I see is angry, they are likely to want to hurt somebody). Various alternatives exist as to implementing probabilistic inferences in the form of a neural network [34,35], and will be explored. Crucially, a suitable objective function to optimise needs to be designed. In line with the most recent multi-task approaches [36] making use of encoder-decoder architectures [37], we will focus on optimising a combination of reconstruction errors for both the (observable) external behaviour of the agent, and its (latent) mental states in a mutally reinforcing process.

*Task 2.2 – Reinforcement learning* [m. 7-30]. Unlike the query answering domain [19,20], in which the topology of the overall network is driven by the gramatical structure of the query itself, in this project the optimal topology of the Theory of Mind simulation network is learned by rewarding configurations accurately predicting the observed behaviour, and by penalising those leading to inaccurate predictions via deep reinforcement learning [38]. To kickstart the learning process, a priori expert knowledge from the field of neuroscience provided by the Cambridge team will be mined to produce initial suitable simulation architectures. 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 Fig. 2(a)). The observed behaviour may also be explained in terms of the agent's reasoning process (upward arrow). Various reinforcement learning approaches will be investigated [39].

**WorkPackage 3 – Validation** [months 4-30].

*Task 3.1 – Neuroscientific validation* [m. 4-30]. To determine to what extent people have Theory of Mind, tests to assess emotion from their eyes/faces are typically used [40,CI1]. In opposition, Co-I Sahakian [CI2] has co-devised novel Theory of Mind tests that looks instead at ambiguous social situations. For this project we will modify these and develop novel ones based around the ice cream van/child scenario (cfr. Background) [CI3], in order to compare the performance of WP1's successive models to the ability of humans to anticipate other people's actions. Variations will be explored, e.g. by swapping children with teenagers seeing friends on the other side of the road. Results will be fed back to WP1 and WP2.

*Task 3.2 – Validation in the AI domain* [m. 4-30]. Novel, bespoke benchmarks will need to be designed and constructed with the specific goal of benchmarking machine theory of mind and intention prediction abilities. We will build in particular on the new READ (Road Events and Activities Dataset) produced by Cuzzolin's team [41] by supplementing the Oxford RobotCar dataset (https://robotcar-dataset.robots.ox.ac.uk/) with complex road event annotations, including multiple labels and detection boxes for action, agent, activity and location classes. Mental state annotations will be added to READ, in close collaboration with the Cambridge team, to allow for the testing of the accuracy of the simulation. Performance will be measured by assessing the accuracy of the predictions produced by ToM simulations versus the actual actions taking place (using standard performance measures such as video- and frame-mean average precision, mAP), as well as directly comparing machine-generated mental state attributions with human guesses in combined, ground-breaking new phychological/technological experiments.

## Project management

Day-to-day work will be conducted by two Postdoctoral Research Assistants (PDRA1 and PDRA2). PDRA1 (based at Oxford Brookes University, under PI Cuzzolin's supervision) will mainly focus on WorkPackage 2, the design and subsequent refinement of machine Theory of Mind models based on reconfigurable deep networks. PDRA2 (at Cambridge University under Co-I Sahakian's supervision) will take responsability for WP1, to drive the development of computational ToM models which actually mirror human behaviour and agree with neuroscientific evidence. Validation in WP3 will be joint responsability of the two teams. Tests in the neuroscience and AI domains will be intertwined to ensure cross-fertilisation. Regular monthly meetings of the Project Team (including the PI, the CI and the PDRAs), either physical or virtual, will allow us to coordinate the work at the two sites. In addition, regular physical visits between Oxford and Cambridge will take place to facilitate the exchange of information among Investigators and RAs. An advisory board has been set up to help steering the project towards success. Oxford University's Philip Torr and British Columbia's Leonid Segal will advise on the AI work (WP2, Task 3.2). Professor Verity Brown will assist the team on the psychological and neuroscientific aspects (WP1, Task 3.1). Co-I Sahakian's time on the project will be contributed by Cambridge University at no cost to the Leverhulme Trust.

| WP/Months | 1-3 | 4-6 | 7-9 | 10-12 | 13-15 | 16-18 | 19-21 | 22-24 | 25-27 | 28-30 |
|---|---|---|---|---|---|---|---|---|---|---|
| **WP1** | 🟨 | 🟨 | 🟨 | 🟨 | 🟨 | 🟨 | 🟨 | 🟨 | | |
| **WP2: Task 2.1** | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | | |
| **Task 2.2** | | | 🟦 | 🟦 | 🟦 | 🟦 | 🟦 | 🟦 | 🟦 | 🟦 |
| **WP3: Task 3.1** | | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 |
| **Task 3.2** | | 🟪 | 🟪 | 🟪 | 🟪 | 🟪 | 🟪 | 🟪 | 🟪 | 🟪 |

## Risk and Mitigation

Given its ambition to radically depart from previous pratice in human behaviour understanding (in terms of vanilla deep learning, or absence of strong links between computer science and neuroscience) this project carries some risk. The proposed agent simulations based on reconfigurable neural networks have never been tried, and would constitute a significant step forward towards general AI. To mitigate this, we will explore various design strategies and test each component separately, as described in the Workplan. Feedback loops between the three WPs will help us identify issues early on and adjust course if necessary.

## Outcomes, Publication and Dissemination

The material outcomes of the project will be: (1) novel, more refined computational ToM models in humans; (2) a new framework for reinforcement learning-based machine Theory of Mind, implemented as a PyTorch codebase, to be made publicly available; (3) new annotated benchmarks and psychological tests on ToM abilities, both in humans and in machines. Results will be published at top-tier conferences and journals in AI, ML, computer vision, psychology and neuroscience, starting from end of year 1. In particular, we will target top conferences such as IJCAI (the premier AI venue), AAAI and UAI, as well as top AI journals such as AIJ and JAIR. We also aim to publish in the best machine learning conferences, such as NeurIPS (attended by 6,500+ researchers in the field), as well as machine learning journals, such as IEEE Transactions on Pattern Analysis and Machine Intelligence (impact factor 9.455). Within neuroscience, publication at Society for Neuroscience (SfN, more than 30,000 attendees), and FENS will be our main objective. Psychological Medicine (i.f. 6.159) will be a suitable journal venue. Top vision conferences such as ICCV and CVPR (5,500 submissions in 2019, 5% chance of oral presentation) will also be considered. We will aim for 3/4 conference publications and 1/2 journals per year. Given the potentially very significant outcome, we will consider publication on Nature as well. A project web site will disseminate the results to fellow researchers, make the datasets gathered and tests conducted in the course of the research publicly available, and contact new partners to kick-start follow-up projects at a more mature stage.