Summary

Abstract

Emerging applications of artificial intelligence are raising an awareness of the limitations of established approaches in situations involving humans. We support the view that a fruitful cross-fertilisation of neuroscience and machine learning can enable significant advances in both fields. Theory of Mind capabilities, i.e., the ability to 'read' other sentient beings' mental states, are crucial to develop a next generation, human-centric artificial intelligence capable of understanding the behaviour of complex agents. In a mutually beneficial process, computational models developed within artificial intelligence may also provide new insights on the way these mechanisms work in the human brain.

Context

Artificial intelligence is being increasingly deployed to our benefit. Nevertheless, its emerging applications are flagging serious limitations of current approaches in situations involving humans. Smart cars, for instance, cannot be deployed among human-driven vehicles, pedestrians or cyclists without being able to make reliable predictions about human behaviour in real time (for instance, in order to pre-emptively adjust speed and course to cope with a group of children's possible decision to suddenly cross the road). The automated recognition of present behaviour builds, to date, on the success of deep learning. The latter is based on artificial neural networks characterised by innovative architectures and an extended number of layers. Deep networks have indeed proved able to efficiently identify motion patterns associated with human actions in streaming videos. Motion patterns, however, can be deceiving as humans can suddenly change their mind based on their own mental dynamics, or things they see in the scene. 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 a cone. No system making predictions purely based on past observed motion would be able to cope with this apparent unpredictability.

In fact, human beings 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 likely to take the elevator rather than the stairs) and the scene (is the person carrying a knife?), and by putting themselves in the other's shoes. Such 'Theory of Mind' capabilities, i.e., the ability to 'guess' another intelligent agent's mental states, are crucial to develop a next generation artificial intelligence inherently designed to understand humans and share a common environment.

Objectives

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. Computational models developed within AI may, in turn, provide new insights on the way Theory of Mind works in the human brain. Whereas various anatomical structures within the human brain have been identified as correlated with Theory of Mind in people, proper, detailed models of the way we are able to guess others' intentions and mental states are still lacking.

Our objectives are: (i) within neuroscience, the formulation and empirical testing of computational Theory of Mind models in humans leveraging the latest efforts in AI, as a step towards 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 evidence, capable to go well beyond current human behaviour anticipation methods based on simple pattern recognition.

Significance and originality

***Significance*.** Success will lay the foundations for the creation, among others, of autonomous vehicles able to negotiate complex road situations in scenarios involving human drivers and pedestrians. In healthcare, next-generation robotic assistant surgeons can 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 NHS, especially when dealing with autism and similar conditions. The new Theory of Mind models we propose may improve the efficacy of psychological treatments, such as cognitive behavioural therapy or mindfulness. From a more foundational point of view, this work will also impact on the current debate on 'moral AI', by helping machines make ethical, human-like decisions in critical situations. Trust in AI will be reinforced, as people realise machines capable of truly reasoning like them become more widespread.

***Originality*.** In visual AI methods for predicting human behaviour are currently based on learning from examples of human actions and activities, without any attempt to model human reasoning, i.e., to provide explanations for the observed behaviour. Existing approaches to computational Theory of Mind, on the other hand, lack the ability to learn from experience. Overall, computational Theory of Mind models based on machine learning remain in their infancy. The notion proposed here (see Methodology) of simulating human reasoning using reconfigurable neural networks is an absolute first.

Methodology

Machine Theory of Mind capabilities, we argue, may be provided by deep networks whose structure can flexibly adapt to the situation. In our proposal, networks implementing simulations of the reasoning of people or other intelligent machines are created by assembling a number of basic neural modules, each describing how two mental states are related (e.g., if a person is carrying a knife, then they are likely to intend to assault someone).

The reasoning of various classes of intelligent agents can then be flexibly simulated by rearranging the structure of the connections linking these basic neural modules, depending on what is observed in the scene (the type of agent, the objects around them, and so on). The way the simulation actually adapts to agent class and scene is learned from experience, just as in human beings. This is done using 'reinforcement learning', which basically rewards 'successful' simulations (those that match with what observed) and penalises incorrect ones.

The work is broken down into a number of WorkPackages (WPs):

WP1 (Novel cognitive Theory of Mind models), led by Sahakian, will develop a new cognitive ToM model by building on the latest neuroscientific evidence, starting from the notion of predictive hierarchical structure, typical of some brain areas such as the primary visual cortex.

WP2 (Computational models for a Machine Theory of Mind), led by Cuzzolin, will implement the ToM model devised in WP1, as well as the successive approximations towards that, in terms of a novel class of 'composable' deep neural networks (explained above).

WP3 (Validation) will jointly validate both the cognitive Theory of Mind models developed in WP1, in terms of their ability to explain the capacity of humans to anticipate other people's actions, and the machine Theory of Mind models constructed in WP2, by measuring their prediction performance on novel benchmark datasets appropriately designed.

Why the Leverhulme Trust?

The proposed research is radically transformational, as it aims at bringing together neuroscience and machine learning to stimulate breakthroughs in both the way AIs understand humans, and in our comprehension of Theory of Mind mechanisms in ourselves.

The Leverhulme Trust has consistently proven be most open to support blue-sky thinking initiatives such as ours, whose impact may not necessarily unfold in the short term but is likely to be enormous in a longer timeframe. Its success will likely affect all applications of artificial intelligence involving humans, including smart vehicles, empathic healthcare and ethical AI, as well as psychological treatments and behavioural therapies.

The proposed is obviously highly interdisciplinary, and will set up a mutually fruitful interaction in which cognitive and neuroscientific evidence inform novel models for a machine Theory of Mind, with the latter employed, in turn, to provide new insights into how these mechanisms work in the human brain at both cognitive and biological level.

The proposed work constitutes an integral part of the vision of both scientists, bringing together the Principal Investigator's ambition to design machines capable of mimicking human-level intelligence and flexibility, and the co-investigator's ground-breaking, widely recognised work on cognitive neuroscience. Both the PI and the CI are leading experts in their respective fields, who have won awards in recent years.

The degree of ambition of this project also implies a certain level of risk, thus making the proposed research (at this stage) less suitable for other sources of funding, such as EPSRC and Horizon2020.