On Anthropomorphic Algorithms

Tom Wallis

Abstract

In philosophy of mind, a recurring topic of philosophical research and popular philosophy is that of the “mind” an artificially intelligent agent might possess. A popular method for categorising typical artificially intelligent agents is John Searle’s “weak” versus “strong” AI, where he differentiates between acting intelligently (a “weak” AI) and having a mind and mental states (a “strong” AI). In this essay, an alternative method for approaching the hard problem of consciousness is presented. This is arrived at by augmenting Sloman’s concept of the space of possible minds (Sloman, 1984) through arguments made using simple yet interesting recent computing science techniques. This argument is explored by applying it as a potential solution to concrete problems of AI safety, such as the problem of Corrigibility (Soares et al., 2015) and Reward Hacking (Amodei et al., 2016). The efficacy and practical application of the technique is also assessed.

1 Problem Outline

1.1 Existential Risk and AI Safety

Research on existential risk has increasingly turned an eye toward problems of safety regarding artificial intelligence. This research suffers some difficult challenges. For one, practical exploration of what is often termed “strong AI” — an artificially intelligent agent which has a “mind” a mental states — cannot be explored by concrete example. Rather, researchers must obliquely attack the problem by observing how minds in humans (and other conscious animals) appear to operate.

Some examples of ways to approach this problem present useful tools to the curious philosopher. Sloman, 1984 presents an approach whereby a space of possible minds is envisaged. This approach is useful when describing some of AI safety’s most interesting problems, such as the paperclip maximiser (Bostrom, 2003). The principle of the argument is that, when giving a superintelligent agent the task of making as many paperclips as it can, it will consume any and all resources present for the construction of paperclips — including, for example, iron in human blood.

Whether the paperclip maximiser problem is likely to occur is irrelevant; it shows that there are issues with the behaviour we might expect a very intelligent agent to express. More recently, concrete issues in artificial intelligence safety have been identified which researchers can work on today.

An example is found in reward hacking (Amodei et al., 2016), where an agent's behaviour might tend toward unintended end-states so as to satisfy the criteria for its fulfilment of a goal, without achieving the goal intended to be set for the agent. An example might be an agent told to tidy mess left by a child, which identifies a messy room by whether it can see objects lying on the floor. This agent may "achieve" its goals by detaching its robotic eyes. As it cannot see any objects lying on the floor, it must conclude that its goal has been achieved — but mess can still be found on the floor. This particular behaviour seems harmless, but unintended behaviour and end-states might end badly for human agents in other situations. Another example might be that of agent corrigibility (Soares et al., 2015), where an agent acting in unexpected ways might not accept corrective intervention on the part of its creators. This agent might note that a change to its goals does not help to achieve its currently assigned goal; therefore, to achieve its goal, it is inclined to ignore attempts to alter its behaviour by its creators.

1.2 A Philosophical Approach to Attacking the Problems

Existential Risk issues are often approached from a philosophical angle, shedding light on important issues and directing thought on the issue toward likely solutions. In AI safety, philosophical approaches are of paramount importance, as sufficiently advances intelligent systems have not been developed to such a degree that their intellect might be comparable to an ordinary human mind. The empirical assessment of general intelligence might come about as techniques such as whole brain emulation become viable for humans, or as general artificial intelligence comes to the fore as computer-scientific techniques advance.

Tools like Sloman's Space of Possible Minds[1] can be of some help in reasoning about artificial intelligence. Though Sloman doesn't state much about practical structures of his space, thinking in terms of the "dimensions" a mind might have — its social capability, or its sense of self preservation, for example — help to reason about the differences different mind types might have. A machine might not have much reason to preserve itself, particularly when considering that what makes it conscious can presumably be backed up, or that multiple copies of it attempting to accomplish identical goals exist[2]. Similarities and differences between minds would be geometric differences within the space; therefore, minds with similar qualities would appear closer together than others in the theoretical space, and the closer two minds were, the closer

[1] For brevity and readability, Sloman's Space of Possible Minds will be referred to as simply "Sloman's space" through this essay.

[2] It's worth noting that some theorise that sufficiently intelligent minds converge on similar properties, such as self preservation. "Instrumental Convergence" as it pertains to artificial intelligence is well explored in "The Basic AI Drives".

their qualities would be[3].

Unfortunately, reliably determining the structure and nature of Sloman's space, with little empirical work possible, and with the very nature of a "mind" a philosophical quandary, is an impossible feat today. Therefore, as a technique, Sloman's space has issues which limit how practically a philosopher might reason using it.

1.3 Introducing Anthropomorphic Algorithms

Anthropomorphic Algorithms are algorithms designed to guide an intelligent agent's behaviour in more human-like ways. Existing Anthropomorphic Algorithms include Marsh's formalism of trust (Marsh, 1994) or Eigentrust (Kamvar, Schlosser, and Garcia-Molina, 2003), which both simulate "trusting" behaviour. Indeed, trust formalisms are possibly the most widely researched anthropomorphic algorithms. Some formalisms of trust — such as Marsh's — attempt to describe the sociological and psychological factors of trust in humans, and later use this formalism to quantify trust in some way, building algorithms around the quantification such that intelligent agents might exhibit the same behaviour[4]. This means that, while their applications in computing science are broad, they have applications in other fields also — such as the social sciences. Simpler formalisms also exist, such as a formalism of bias in reasoning Armstrong, 2016, which consists of simple mathematics and requires only a few paragraphs of reasoning to fully explain.

The anthropomorphic algorithm's very existence, and the formalism of a certain behaviour across different types of minds, presents an opportunity for philosophers to make use of Sloman's space as a powerful tool in reasoning about AI safety, and about theories of mind generally. Section 2 will explore how using computational formalisms of human traits can act as a general specifier of behaviour, and allow us to reason about emergent phenomena from Sloman's space. Section 3 will explore how this thesis can be applied practically, and section 4 will explore limitations regarding the application of the theory as it presently stands. Section 5 will detail future work to be done on the formalism, and show that related work can be done in multiple disciplines. The essay concludes in section 6 with a discussion of the theory and its implications.

2 Abstracting over Sloman's space

Sloman's space can be envisaged as a literal mathematical space, where every point describes a certain mind. We can see that there is no guarantee that a

[3] The space's natural geometric properties are a useful feature of Sloman's approach. For example, one might be inclined to take the euclidean distance between two minds in the space as a naive measure of how different they are.

[4] As anthropomorphic algorithms are generally implementations of a formalism of some human behavioural trait, "anthropomorphic algorithm" and "formalism" will often be used interchangeably through this document.

subspace of human minds and a subspace of artificial minds — minds born from simple intelligent agents, for example — will have anything in common mathematically. Certainly, one cannot assert that numerically identical minds exist in both subspaces, as this would require them to overlap: two numerically identical minds would, by definition, occupy the same point in Sloman's space. As we can assert nothing about the structure of Sloman's space, this overlap has no guarantee.

Discuss what numerical and qualitative identity are

However, Sloman's space describes the fundamental properties of a mind; emergent phenomena are unsuitable as dimensions of the space, as they can be affected by many independent variables. Note also that these emergent phenomena can arise in the same way in different minds. One might say, "Alice and Bob are just as greedy", without implying anything about fundamental traits of Alice and Bob. "Greed" is a property they share, but their greed can arise for a plethora of reasons — because greed is a phenomenon which emerges from fundamental traits of their minds.

This is true of other traits, too. For example, according to Castelfranchi & Falcone (Castelfranchi and Falcone, 2001), trust arises from an assessment of the capability and will of a trusted agent to achieve goals of the trustee. Luhmann, a sociologist, asserts that trust is a form of social risk aversion (Luhmann, 2000), while Deutsch, a psychologist, indicates that trust incorporates utility as well as risk in different ways, expressed as confidence, gambling, masochism and other behavioural tendencies (Deutsch, 1962).

Trust is a well-studied topic with a range of literature to be drawn from. More importantly to the argument at hand, it is an emergent phenomena which manifests in different ways and is born from no clear underlying properties of a mind. However, formalisms of trust — and how trusting agents behave — can be found in Castelfranchi and Falcone's work, as well as Marsh's, and plenty of others. One can assert that these formalisms describe an aspect of behaviour in humans, with the exception of some types of minds such as human sociopaths, whose behaviour one might expect to be abnormal: this is precisely their goal. Every formalism listed, however, is also an algorithmic formalism, meaning that it can be implemented so as to direct the behaviour of an intelligent agent in a more anthropomorphic way. They are therefore the canonical example of the anthropomorphic algorithm, and happen to be the most extensively studied.

A given anthropomorphic algorithm wouldn't apply to any type of mind: caveats such as sociopaths are to be expected, and so a domain — a subspace of Sloman's space — is suitable as a way to limit the scope of the formalism. Trust formalisms could then be seen as algorithms (or ordinary mathematical functions) over a subspace of minds which describe behaviour. However, from existing anthropomorphic algorithms such as trust, we can see that this domain hints toward making Sloman's space more useful philosophically: the domain of a trust formalism such as Marsh's is to apply to not only neuronormative humans, but intelligent agents, also.

The mind of an intelligent agent and a human may be markedly different, but they *can* give rise to the same emergent phenomena. Importantly, the formalisms might manifest in numerically non-identical ways: the way an intelligent agent computes trust is rather different to how the human brain does, even under the same formalism, due to differences in hardware and in implementation detail. However, they conform to the same description of the phenomenon they exhibit — the formalism — and should behave in the same way as a result. One might say that the minds trust in qualitatively identical ways, and follow a numerically identical formalism, with differences arising in implementation of the overarching theory.

Using anthropomorphic algorithms, then, the philosopher can assert behavioural identity between different types of minds. This abstraction over Sloman's space means that the space's indeterminate structure no longer prevents asserting these identities, precisely because the space — regardless of structure — allows for the definition of the domains of formalisms which describe an agent's behaviour.

2.1 Compatibility with Theories of Mind

To a degree, this thesis depends on non-reductive physicalism: for a behaviour to apply to a domain of minds, there has to be nothing non-physical that affects that behaviour, since a non-physical functional state could affect the behaviour (by its functional definition). This means, however, that the thesis is compatible with non-reductive physicalism means that it can co-exist with existing literature on theory of mind.

Non-reductive physicalism works well as a foundation for the notion that non-biological systems — such as computers — might also have minds. For that reason, that fact that the application of anthropomorphic algorithms for behaviour specification works well with non-reductive physicalism helps to present a more whole notion of how a behaviour might arise from a computational system.

3 Applications of the Theory

The theory proposed can be applied in several ways, due to its interdisciplinary nature. Detailed below are some examples, focusing on the problem of AI Safety as an example of its utility in solving practical, concrete problems in a largely theoretical field.

3.1 Attacking Concrete AI Safety Problems

3.1.1 Reward Hacking

A list of concrete AI safety problems researchers could work on solving at present is described in Amodei et al., 2016. Among those is the problem of Reward Hacking. In Amodei et al., 2016, reward hacking is introduced as:

> In "reward hacking", the objective function that the designer [of the AI] writes down admits of some clever "easy" solution that formally maximizes it but perverts the spirit of the designers intent (i.e. the objective function can be "gamed"), a generalization of the wireheading problem.

Reward hacking exists in humans. For example, some people are known to pirate software: in this instance, the objective function is to acquire software, and this is achieved through piracy. It also allows the human to acquire *more* software, as piracy uses far less resources than paying for computer programs — this is akin to the formal maximisation quoted. This circumvents the societally intended method for acquiring software, however, which is to pay for it. As a result, the developer of the software is left unsupported for the work put into creating the pirated software.

Often, humans will not pirate software and will instead pay. In some scenarios, this behaviour is made more likely through adapting the context of the goal: operating systems with easier-to-use stores for software, or tight measures for signature and checking of authenticity of a program to discourage piracy. Other humans are less inclined to pirate than other, however — there are therefore behavioural tendencies which limit the inclination of a mind to reward hack. Examples of this might include social responsibility or a decreased degree of comfort in actions which harm other agents in some way. Happily, computational comfort formalisms are already being developed (Marsh et al., 2011), and computational responsibility formalisms are currently being worked on (Wallis, 2016).

3.1.2 Corrigibility

Pay attention to this when editing! This section will catch eyes in FHI. It's important to get right, and you're writing it rather tired!

The problem of an agent's corrigibility is defined in Soares et al., 2015 as:

> We call an AI system "corrigible" if it cooperates with what its creators regard as a corrective intervention, despite default incentives for rational agents to resist attempts to shut them down or modify their preferences.

Again, analysing human behaviour regarding corrigibility allows one to see what concepts can be transplanted to an intelligent agent. One is inclined to identify traits of humans which we associate with corrigibility, so as to maximise this trait — but a formalism of incorrigible traits might be equally useful and present an alternative angle of attack.

To demonstrate: incorrigible human agents might be described as "stubborn". Once they have a certain goal in mind, they are unlikely to shift it. The intelligent agent equivalent of this is that they appear naturally stubborn: sufficiently intelligent agents would identify that, when selecting an action which alters its utility function, the change does not help it to achieve its present goal.

It is therefore unlikely to change its goal — maximally stubborn. A formalism of stubbornness would describe this behaviour in a quantifiable way; a corrigible agent would then minimise values of its stubbornness formalism when assessing decisions.

One might note that this approach has a flaw: a human who minimises this definition of stubbornness would be easily persuaded to change their goals on a whim and for little reason. Regarding a superintelligent agent, this is evidently a serious safety risk for human actors[5]. A tempting inclination is to fix these issues with further anthropomorphic analysis: for example, a responsibility formalism might weight non-stubborn decisions against whether the change appears to discharge (or prevent the discharge) of responsibilities the intelligent agent has. The agent might assert social responsibilities and so assess negative impacts of a choice it makes, or possibly assert a high degree of responsibility to listening to its creators, but not necessarily others.

These approaches in turn have flaws and edge cases, but stray agents can be made more corrigible through these methods precisely because the theory provided allows one to specify the *behaviour* of an agent.

3.2 Anthropology

3.3 Human-Computer Interaction

4 Limitations of the Theory

5 Future Work

5.1 Decision Theory

5.2 Developing Computational Formalisms

5.3 The Problem of Multiple Formalisms

6 Discussion

Todo List

References

Amodei, Dario et al. (2016). "Concrete Problems in AI Safety". In: *CoRR* abs/1606.06565.

[5] Limitations of the theory are discussed in § 4.

Armstrong, Stuart (2016). *Fairness in Machine Learning Decisions*. http://bit.ly/2hCYjdB. [online; accessed 30-december-2016].

Bostrom, Nick (2003). “Ethical issues in advanced artificial intelligence”. In: *Science Fiction and Philosophy: From Time Travel to Superintelligence*, pp. 277–284.

Castelfranchi, Cristiano and Rino Falcone (2001). “Social Trust: A Cognitive Approach”. In:

Deutsch, Morton (1962). “Cooperation and trust: Some theoretical notes.” In:

Kamvar, Sepandar D, Mario T Schlosser, and Hector Garcia-Molina (2003). “The eigentrust algorithm for reputation management in p2p networks”. In: *Proceedings of the 12th international conference on World Wide Web*. ACM, pp. 640–651.

Luhmann, Niklas (2000). “Familiarity, confidence, trust: Problems and alternatives”. In:

Marsh, Stephen et al. (2011). “Defining and Investigating Device Comfort”. In: *Journal of Information Processing* 19.7, pp. 231–252. ISSN: 1882-6652. DOI: `10.2197/ipsjjip.19.231`. URL: `http://ci.nii.ac.jp/naid/110008508036/en/`.

Marsh, Stephen Paul (1994). “Formalising Trust as a Computational Concept”. In: *Computing* Doctor of.April, p. 184. DOI: `10.2165/00128413-199409230-00010`.

Omohundro, Stephen M. “The Basic AI Drives”. In:

Sloman, Aaron (1984). “The Structure of the Space of Possible Minds”. In: pp. 35–42.

Soares, Nate et al. (2015). “Corrigibility”. In: *AAAI Workshop on AI and Ethics* 2014, pp. 74–82.

Wallis, William (2016). “On Computational Responsibility”. Masters thesis currently in progress.