On Anthropomorphic Algorithms

Tom Wallis

Abstract

The space of possible minds is a notion which has quickly permeated AI Safety literature since its introduction in Sloman, 1984. While the concept is useful, it sees rare empirical usage due to the fact tat the structure and nature of the space is difficult to define. This essay presents a technique by which one can model aspects of the expected behaviour of a system through an abstraction over the space. Further, it shows that this technique can apply to both human and artificial intelligences, and create qualitative identities across an arbitrary subspace of the space of possible minds. It explores some of the applications of limitations of this technique, and through the current limitations shows how future work can mitigate these issues to create a stronger practical technique for the exploration of AI Safety.

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" and 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.

Sloman, 1984 presents an approach to the problem 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). Paperclip maximisers are intelligences which pursue goals — such as the creation of ever more paperclips — while showing indifference to any consequences, such as the eventual mulching of humans into stationary.

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 move 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. 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. Finding solutions to these problems is a challenging undertaking, yet is also imperative for a future with safe AI.

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 advanced intelligent systems have not been developed to such a degree that their intellect might be comparable to an ordinary human mind. Therefore, empirical study can be impractical, and a-priori analysis through philosophy is a researcher's most valuable tool.

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 sense of self preservation, for example — help to reason about the traits differentiating types of minds[2]. Similarities and differences between minds can be considered 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 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 an-

---

[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 explored in "The Basic AI Drives".

[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.

thropomorphic 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, as well as future work solving these limitations. The essay concludes in section 5 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 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 two numerically identical minds would, by definition, occupy the same point in Sloman's space.

However, Sloman's space describes the fundamental properties of a mind; emergent phenomena[5] 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. Kripke was the first to discuss that a behavioural artifact with the same rigid designator can arise from different minds (Kripke, 1972) — a simple corollary being that behaviour cannot distinguish different minds, as their properties can converge on the same behaviour.

---

[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 essay.

[5]Emergent phenomena are generally complex phenomena where the same behaviour can arise from different initial conditions, or entirely different systems. They are often studied in sociotechnical systems research, which analyses systems of interrelated social and technical components working in synchrony.

2.1 Anthropomorphic Algorithms for Specifying Behaviour

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). Some sociotechnical work attempts to bridge the divide between the sociological and psychological approaches through algorithmic formalism (Marsh, 1994).

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). Many formalisms listed are an algorithmic formalisms, meaning that they 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 simply 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, because they 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.

2.2 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, as this would imply a behaviour cannot be shared by humans and computers. However, that the thesis is compatible with non-reductive physicalism means that it can co-exist with existing literature on theory of mind, and that the theories of mind it relies on are at least partially defined already.

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, athletes will on occasion make use of performance enhancing drugs — their reward function is fulfilled by winning gold medals, but the goal itself isn't technically fulfilled if it makes use of a loophole in some regulation of a sport.

There are 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 in their early stages (Wallis, 2016).

3.1.2 Corrigibility

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[6]. 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. An equilibrium between the behaviours of responsibility and stubbornness would therefore be ideal, and stable combinations of different anthropomorphic traits can be expected to feature in solving similar instances of this dilemma.

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. Agents can therefore be made corrigible in more, if not all, scenarios.

---

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

3.2 Anthropology

Refining and developing formalisms of behaviour of certain groups of people is a task often taken on by sociologists and psychologists (Gambetta, 1988; Luhmann, 2000). However, as these formalisms become more advanced, they permit two applications for modern anthropology.

First, anthropologists may use the formalisms to predict the behaviour of new cultures, as well as how those cultures might shift, as more socially involved artificial minds interact with human minds more frequently. Studying this is important, as the impact of minds with a subset of human traits might be significant: a greater emphasis might be put on social constructs of respect, for example, should a collection of people begin using respect as an interaction mechanism with the technology around them.

It is not unknown for technological interaction to put more — or less — emphasis on the aspects of human culture it touches. For example, modern communication techniques such as instant messaging alter the dynamic of communication: within the space of only a century or two, the expected time to wait for a response to a message has shortened from the days it took to respond to a letter to the minutes it takes to read and reply to an instant message. In turn, people's expectations about the promptness of communication alter accordingly; therefore technological innovation can have a marked and significant impact on culture. Introducing artificial minds with a subset of human traits lends a ripe field of study for anthropologists to research.

Secondly, when observing new types of culture and shifts in common cultures, anthropologists will begin to note how some fundamental human traits — rather than emergent phenomena of the ordinary human mind — shift over time. This will help to:

1. Understand more of the nature of Sloman's space, as it pertains to the fundamental traits of a human mind

2. Understand more of the structure of the human subspace of minds, and what different types of human minds might exist

The utility of this work to the philosopher, and to AI Safety research, is that integrating empirical observations about the world allows for more informed and practical thought experiments or practical exercises in AI .

The theory then applies to anthropological study in the deeper understanding of the human mind from a cultural perspective, and a pragmatic framework by which anthropologists might note precisely what sorts of minds their field studies. It might also permit anthropological study to broaden, from humans and their culture to minds with human-like traits — observing how minds with human properties might engage with culture, and what traits affect culture in different ways. This practice has the added benefit of the anthropological study contributing to the original theory, strengthening empirical evidence which can be used in the study of AI and its safety, as well as lending practical applications to non-reductive physical theories of mind through their working in tandem with the theory presented.

3.3 Human-Computer Interaction

A better understanding of the formalism of minds does more than bolster our understanding of artificial minds: it also helps to perceive how a human operates and therefore, how interfaces between artificial minds and human minds might be better achieved.

Currently, much human-computer interaction research centres around interaction mechanisms and design for input and output devices: styluses, spherical displays, phone applications and so on. Application of this formalism-centric approach, however, would permit a human-computer interaction paradigm which centres around how "feelings" change: humans and computers might work together to build trust, or to avoid danger through a sense of fear, as these are all human behaviours which can theoretically be formalised and applied to an artificial mind.

An example might be trust regards privacy: a user's phone might present fewer security barriers to accessing sensitive information when being accessed in a hospital, for example, if the data accessed was related to health. Then, sensitive medical data can be unlocked by the trust in a user or location which can be learned over time; if a hospital shuts down and medical professionals are no longer accessing data from the location, trust in that location would dwindle just as human trust in old notions might dwindle over time also.

In the field of human-computer interaction, it's unusual for rigorous mathematical study to take place; empirical "laws" of interaction, such as Fitt's Law (Fitts, 1954), are rare. However, this theory presents an approach which, with research into the space's structure, present another mathematical and rigorous study of interaction.

4   Limitations and Solutions

As it stands, the theory proposed has several useful applications. This is not to say, however, that it presents a complete and foolproof measure of how minds might interact in Sloman's space, nor is the theory as complete as one might hope it to be. These issues, regardless of their significance presently, do not seem to be ultimately fatal to the theory. We shall explore some of these issues now, and discuss future work which will address some of these issues.

4.1   Incompatibility with certain Theories of Mind

One criticism of the proposed theory is that it is incompatible with some theories of mind, and relies on non-reductive physicalism, which — while a popular theory — has come under criticism with several counter arguments in the last few centuries. An argument can be made that any physicalist theory of mind is sufficient for the theory to work: all it relies on is that artificial minds belong in the space of minds with human minds. However, to permit equivalent behaviour across different Sloman subspaces, other physicalist theories

such as identity theory require workarounds, because the bridge laws they often rely on limit the minds that experience certain phenomena based partly on *hardware*.

Detailing potential workarounds for alternative theories of mind is beyond the intention of this essay; future work might apply similar abstractions to other theories of mind, exploring this technique's utility assuming, say, a reductive physicalist theory of mind.

4.2   Omitted Formalisms

As may have been noticed already, behaviour formalisms alone can't solve problems in AI safety. One factor which conspires to complicate existential risk research is the difficulty of dealing with edge cases. To demonstrate: should a general artificial intelligence be created, it is unlikely that the algorithms that create it would be contained indefinitely. This is because these technologies to create the intelligence will be rediscovered many times, and commitment to the protection of this data from the public requires potentially international collaboration.

Any safety measure which is not essential for the creation of a general intelligence should therefore be neglected as a complete solution: someone is likely to leave the algorithms out of their general intelligence, and only one such instance presents a major safety risk.

Though it is tempting for this to inspire a rather bleak outlook, many processes may be discovered which may keep humanity safe from an artificial general intelligence. One such process is behaviour formalism. To fix the issue, therefore, future work should be undertaken so simplify formalisms and make them easy to include in an AI's construction, perhaps exploring engineering techniques to simplify the integration of formalisms and traditional AI systems design.

4.3   Limited numbers of formalisms

One limitation of the theory as it currently stands is that, whole it is valuable in its practical merits, computational formalisms of human behaviour are limited in scope currently. Formalisms of trust are plentiful, and work is done on comfort and responsibility, but the spectrum of human behaviour and experience is hard to formalise. Lots of this work has not yet been undertaken.

Over time, this limitation should lessen; regardless of the utility of the philosophical theory, formalisms as a tool for computer scientists, sociologists, and psychologists are useful and can be expected to grow in scope as a result. Therefore, this limitation is expected to be short-lived. Nevertheless, actual solutions to concrete AI problems may not be feasibly produced until a broader range of formalisms have been developed, and formalising an array of behaviours should be a priority of future work.

4.4 Formalism Accuracy

One issue with formalisms as a field of study is whether they "accurately" represent the behaviour they claim to. For example, Eigentrust (Kamvar, Schlosser, and Garcia-Molina, 2003) uses the concept of trust a metaphor to draw conclusions from reputation, which it focuses on modelling. The argument often provided in these situations is that the behaviour "formalised" is a useful metaphor for an intended outcome or perceived behaviour on the part of the algorithm. Therefore, whether it accurately represents trust is a moot point, as it has no effect on Eigentrust's efficacy in its *intended* purpose.

I propose two solutions to this issue:

1. The term "anthropomorphic algorithms" should apply specifically to behaviour formalisms which have their roots in well-defined sociology, psychology, and anthropology. In this way, researchers and AI developers should be able to distinguish between socially accurate behaviour formalisms and those derived from analogy for engineering purposes.

2. Development of rigorous formalism analysis in the fields that the formalisms originate from: this would enable a richer and more diverse range of formalisms, which is hard to produce without interdisciplinary collaboration with these social sciences and humanities subjects.

5 Discussion

Behaviour can be modelled as an abstraction over Sloman's space, realised by formalisms of human behaviour which can also apply to non-human minds. This approach requires some development to be practically useful, but it can aid in providing a mitigating technique which makes an artificial intelligence's behaviour easier for a human developer to predict and reason about.

The technique has a number of benefits, including:

- Helping to attack problems in AI safety, such as corrigibility and reward hacking

- Providing further opportunities for interdisciplinary study

- Opportunities to develop practical solutions to previously theoretical problems

- Progressing theories of mind into testable, implementable theories

The practical approach requires further development in both theory and application, including:

- Development of more formalisms, from sociology, psychology, and anthropology

- Analysis of possible structures of Sloman's space

- Application of the formalisms to more AI safety problems

- Work on combining the behaviours described by formalisms of different behaviours

- Analysis of the theory as it pertains to theories of mind other than non-reductive physicalism

The theory itself, and the application of it as an appropriate model of behaviour, requires further argument and analysis. However, the theory integrates easily with existing and popular theories of mind, and popular concepts in AI safety literature (particularly Sloman's space).

What remains to be done, therefore, is not so much proof-of-concept work for the theory as it is the theory's applications. As these applications are presently realisable, a great body of work in a series of fields should be undertaken so as to test the hypothesis, and provide fertile ground for the development of further pragmatic approaches to AI safety and the integration of philosophy and computing science as a whole.

References

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

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:

Fitts, Paul M (1954). "The information capacity of the human motor system in controlling the amplitude of movement." In: *Journal of experimental psychology* 47.6, p. 381.

Gambetta, Diego (1988). *Trust: Making and Breaking Cooperative Relations*. Blackwell.

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.

Kripke, Saul A (1972). "Naming and necessity". In: *Semantics of natural language*. Springer, pp. 253–355.

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.