On Anthropomorphic Algorithms

Tom Wallis

Contents

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 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[1]. 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 their qualities would be[2].

[1] 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".

[2] 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.

2 Assertions

3 Proposed Argument

4 Applications

5 Future Work

6 Discussion

References

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

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

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.