Saving The Control Problem

Dustin Juliano

The **control problem** is a question posed by Nick Bostrom on how to limit advanced artificial intelligence while still benefiting from its use. I propose an extension to the original control problem that separates it into a local and global version. I then provide proofs that the global version has no solution.

Terminology

The terminology "advanced artificial intelligence" is taken, more or less, to be synonymous with artificial superintelligence (ASI), strong artificial intelligence (SAI), and artificial general intelligence (AGI), only for the specific purpose of reasoning in this context.

Definitions

The *global* version of the control problem universally quantifies over *all* advanced artificial intelligence to prevent *any* of them from escaping human control. The apparent rationale is that it would only take *one* to pose a threat. This is the most common interpretation when referring to the original control problem without a qualifier on its scope.

By contrast, I introduce a tractable version called the **local control problem**, which asks if there exists *any* advanced artificial intelligence that can be controlled. This is claimed to be solvable by the author without a proof provided in this text. It is currently an open problem in the field.

Proofs

Theorem. The *global* control problem has no solution.

Proof 1. Let P represent a compiled program in a verified instruction-set architecture that implements an advanced artificial intelligence that has been proven safe and secure according to agreed upon specifications. If P is encapsulated in an encrypted program loader then simulate it in a virtual machine and observe the unencrypted instruction stream to extract P. Next, disassemble and recompile or patch P to alter its behavior and change one or more verified properties. Now modify P such that all safety and security is either removed from the final program or rerouted in control of flow. Then distribute P widely and in a way that can not be retracted. An easily accessible alternative to P now exists, defeating the global version of the control problem. □

Proof 2. Let P represent a compiled program in a verified instruction-set architecture that implements an advanced artificial intelligence that has been proven safe and secure according to agreed upon specifications. Let K represent a compiled program for some instruction set architecture that implements an advanced artificial intelligence that was discovered independently from P. Suppose K has sufficient and similar capabilities to P and is of concern to the context of the control problem, with neither safety nor security properties to limit it. Now distribute K widely and in a way that can not be retracted. An easily accessible alternative to P now exists, defeating the global version of the control problem. □

Discussion

Informally, the theorem can also be interpreted as applying to the so-called "friendly artificial intelligence" problem. The goal was that a proven safe and secure version of advanced artificial intelligence could be created that would displace or defend against any other advanced intelligence from subsuming humanity. Crucially, proof (2) shows that goal to be invalid. However, if it can help solve the *local control problem* then it is still a valid line of research.

Finally, the arguments presented above operate implicitly under the Church-Turing thesis. It is expected that valid software programs can be translated to appropriate hardware implementations. Thus, the theorem and its proofs are similarly appropriate for any hardware implementation of advanced artificial intelligence.

References

Bostrom, Nick. *Superintelligence: paths, dangers, strategies.* Oxford, United Kingdom: Oxford University Press, 2014.