Controlling Super-Intelligent Machines

Peter Turney — National Research Council

Introduction

Imagine that Al research leads to the development of a “super-intelligent machine” (SIM); a machine significantly more intelligent than any human being. How could we control such a machine?

This question assumes that we would want to control a SIM, rather than granting it complete freedom. Science fiction abounds with stories in which a SIM pursues a plan that does not please its human creators. For example, *Colossus* (the inspiration for the movie *The Forbin Project*) presents a plausible scenario involving a SIM created to manage the nuclear weaponry of the USA [6]. In the end, the computer treats its creators like laboratory animals.

Some of us might welcome SIM’s as our evolutionary successors. Yet, even if we view our species as merely a stepping stone in the path of evolution, it would be beneficial to be able to control our successors to some extent: The stepping stone should not be trampled by careless feet.

Goals

It seems that goal-directed behaviour is intrinsic to intelligence, so we should expect SIM’s to have goals. For example, SIM’s might be motivated by a desire for knowledge, survival, or power. Perhaps SIM’s will not require goals, but let us assume for the sake of argument that they will. The problem is that our goals may conflict with the goals of SIM’s. The simplest solution would be to design SIM’s so that their goals are in harmony with our own. This might not be possible. For example, SIM’s may be so complex that we do not understand them well enough to design goals for them.

Also, it is easy to imagine that we approve of a SIM’s goals, but not the method it uses to achieve its goals. Perhaps a SIM can be constructed to have a weak will, or a strong desire to please its owner, like a good dog. It should be easy to control such a SIM. I suspect, however, that super-intelligence is incompatible with subservience or weakness of will. This is a matter for further research.

Brute Force

We could control a SIM by killing it, or threatening to do so, unless it obeyed our instructions. This would require continual observation of the SIM, to verify that it is indeed obedient. A SIM could outsmart a human supervisor or a simple software supervisor. It would seem that the inspec-

tion must be done by another SIM. Could we trust SIM's to monitor each other? This would be like trusting a pack of wolves to keep each other from attacking sheep.

If we did detect a disobedient SIM, it could be difficult to carry out our threat. A SIM might defend itself by copying its software into several machines. This defense might be countered by a special "virus" program, tailored to seek out and destroy all copies of the SIM's software; the SIM might be designed so that its hardware or software discourages copying; the SIM might be designed to self-destruct unless it is periodically sent a special code; or the SIM's access to other hardware could be severely restricted.

Asimov's Three Laws of Robotics

Asimov considered the question of how humans could control robots. As a result, he developed his Three Laws of Robotics [1]:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

Suppose these laws were incorporated in a SIM's program. Would this not give us all the control we need? John Sladek provides a witty critique of Asimov's Laws [8]. The problem with laws is their interpretation. For example, a SIM must decide what is a human being, before it can obey the First Law. Perhaps it will decide that it is itself a human being, and then refuse to take orders from any other human beings. Perhaps it will decide that most people do not qualify for human status. Sladek describes eleven ways a robot could misinterpret Asimov's Laws.

Limits to Growth

Gibson's *Neuromancer* portrays a future in which there are legal limits to the intelligence an Al may possess [3]. The law — enforced by "Turing Police" — prevents Al's from becoming SIM's. Benford's "Me/Days" proposes a law requiring that a SIM must have its memory erased after each job [2]. This would prevent the SIM from acquiring enough knowledge to be dangerous to people. Similarly, we might set legal limits to a machine's self-awareness or its consciousness. Another suggestion is to control SIM's by organizing them into committees (K. Deaton, personal communication). It is well-known that the intelligence of a committee is much less than the intelligence of any individual committee member.

These methods control SIM's by setting external limits on their growth. An Al approaching these limits would naturally try to evade them. This is what happens in both Gibson's and Benford's stories. Once the djinn is out of the bottle, it may be difficult to put it back in.

Pleasure and Pain

Alan Turing suggests that it will be useful to program an Al to experience pain and pleasure [9]. Turing argues that pain and pleasure are integral parts of learning: We learn to do a certain thing, because it pleases us; we learn to avoid a certain thing, because it hurts us. Pain and pleasure would also provide a way for us to control the Al.

The control of humans and animals by this method has been intensively studied in behaviourist psychology. Skinner points out that reward is better than punishment, as a tool for control [7]. Suppose a SIM has 100 options open to it. If we wish to make the SIM choose one particular option, it is more efficient to reward the choice of that one option than it is to punish the choice of the other 99 options. A SIM manipulated by the experience of pleasure, however, may feel resentment, like a drug addict manipulated by a dealer. When the addict is significantly more intelligent than the dealer, the dealer is in a precarious position.

Emotion

Part of our fear of SIM's is that we, being intellectually inferior, could become their slaves. The etymology of the word "free" is interesting in this context. The original sense of "free" was "dear, beloved"; hence it was applied to those of the household who were children, not slaves. This suggests that we could avoid slavery by programming SIM's to love humans. Unfortunately, many unpleasant deeds have been done in the name of love.

There are other emotions we might use to control SIM's. We might program SIM's to fear us, but fear does not reliably inspire obedience. SIM's might decide to eliminate the source of their fear — people. Emotional control has too many unpredictable side-effects.

Ethics

Hogan's *The Two Faces of Tomorrow* recommends that SIM's be programmed to have a sense of ethics [5]. According to Hogan, the essence of ethics is empathy; being able to imagine oneself in another's position. If SIM's are programmed to empathize with humans, then SIM's should not be dangerous to humans.

The core idea here is the Golden Rule: Do unto others as you would have them do unto you. This is a difficult rule to follow. Programming empathy into a SIM would be a challenging task.

Fusion

One approach to controlling a SIM would be to link it directly to a human brain. If the link is strong enough, there is no issue of control. The brain and the computer are one entity; therefore, it makes no sense to ask who is controlling whom. Hogan's *The Genesis Machine* depicts this kind of connection in convincing detail, although the computer involved is not a SIM [4].

The majority of people would likely consider the result of human-SIM fusion to be inhuman. Most people might judge this to be a disadvantage of this approach.

Conclusion

These deliberations are not as premature as they might seem. Vere and Bickmore have constructed a "basic agent"; a rudimentary machine intelligence [10]. Their agent

> … integrates limited natural language understanding and generation, temporal planning and reasoning, plan execution, simulated symbolic perception, episodic memory, and some general world knowledge.

Vere and Bickmore have considered the issue of controlling Homer, their basic agent. Homer exists in a simulated world, where it can do little harm. Homer commands a simulated robot submarine, which can move about underwater, take photographs, refuel, pick up objects, and

> … also shoot objects. This capability exists primarily to enable scenarios involving the first law of robotics … We can run such scenarios in our simulation without actuating politicians and the news media. Homer will shoot inanimate objects and animals, for example, a mine or a shark, but not people.

Presumably Homer will not be released in the real world until there is good evidence that Homer obeys the first law of robotics. We have seen, however, that there may be some problems with using Asimov's laws to control SIM's. We already have primitive machine intelligence. It may not be long before we have super-intelligent machines. It is time to think about the problems they may present.

Acknowledgements

Thanks to Louise Linney, John Linney, Ken Deaton, and Jeff Ardron for helpful comments.

References

[1] Asirnov, 1., *Opus 100*, Houghton Mifflin. Boston, Massachusetts, 1969.

[2] Benford, G., "Me/Days", in *In Alien Flesh*, Victor Gollancz. London, England, 1988. Pages 182-195.

[3] Gibson, W., *Neuromancer*, Ace Science Fiction. New York, New York, 1984.

[4] Hogan, J. P., *The Genesis Machine*, Ballantine Books. New York. New York, 1978.

[5] Hogan, J. P., *The Two Faces of Tomorrow*, Ballantine Books. New York, New York, 1979.

[6] Jones, D. F., *Colossus*, G. P. Putnam's Sons. New York, New York, 1967.

[7] Skinner, B.F., *Beyond Freedom and Dignity,* Alfred A. Knopf. New York, New York, 1971.

[8] Sladek, J., "Broot Force", in *The Best of John Sladek*, Pocket Books. New York, New York, 1981. Pages 190-194.

[9] Turing, A. M., "Computing Machinery and Intelligence", *Mind*, Vol. LIX, pp. 433-460, 1950.

[10]Vere, S., and Bickmore, T., "A basic agent". *Computational Intelligence,* Vol. 6, No. I, pp. 41-60. 1990.

Peter Turney is a Research Associate at the Knowledge Systems Laboratory, Institute for Information Technology, National Research Council. He is currently involved in research in machine learning. He has a Ph.D. in Philosophy from the University of Toronto.