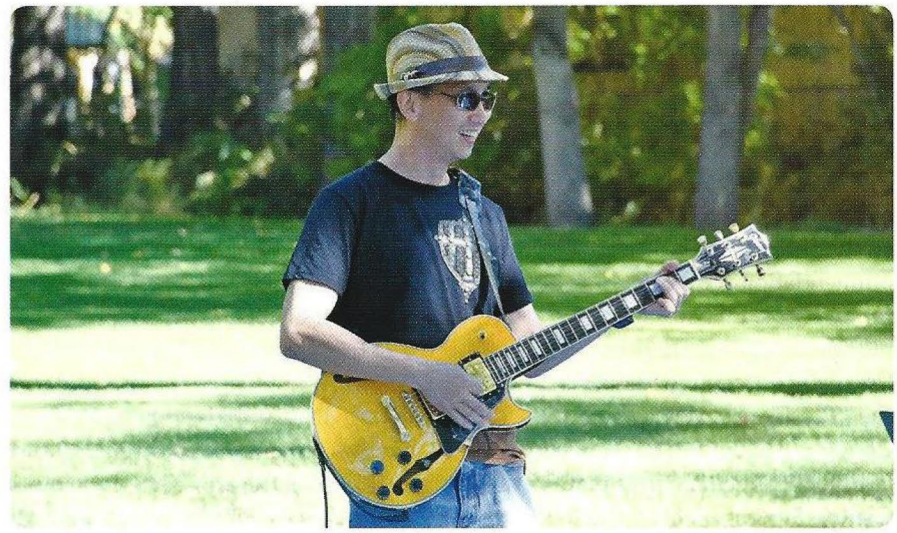
The Control Problem

Have you heard of the *control problem*? No, I don't mean *a* control problem or even the control problem you are working on now. I mean *the* control problem. Not sure what I'm talking about? Let me explain.

The rise of automation and artificial intelligence (AI) has been heralded in the popular media as a very significant sea change. Already, we have technology for self-driving cars, game playing, music composing, fraud prevention, and even software to write articles automatically (although, I'm sorry to disappoint you, not this one) [1]. There are even claims that "robots paved the way for Donald Trump" [2].

But why do we not see articles in the popular media on the "rise of control" or how "control is changing the world?" Are the problems and technologies of automation, robotics, or even AI not also those of our domain? Do control theorists and engineers have nothing to contribute to this huge wave sweeping the world? Even a prominent AI researcher concedes to the role of control in AI [3]:

> AI is thus a control problem, at least in a trivial sense, but also in a deeper sense. This view is to be contrasted with AI's traditional view of itself, in which the central paradigm is not that of control, but of *problem solving* in the sense of solving a puzzle, playing a board game, or solving a word problem. Areas where the problem solving paradigm does not naturally apply, such as robotics and vision, have been viewed as outside mainstream AI. I think that the control viewpoint is now much more profitable than the problem solving one, and that control should be the centerpiece of AI and machine learning research.



Ed Chong at a fundraising gig.

Although this quote is from a paper published in 1988, not much has changed since then regarding "control" being "the centerpiece," at least publicly. Perhaps we've been too preoccupied with proving the optimality of our solution or the asymptotic stability of this or that controller. Perhaps our field has an image problem.

But wait, not so fast. Apparently, the public media has now associated a central problem in AI with what is called the control problem [4]:

> Nick Bostrom, an influential thinker on the subject of AI, calls this the "control problem." In essence, any sufficiently intelligent artificial mind could be capable of having devastating effects on the world, so approaches to controlling such a creation should be carefully analyzed beforehand.

To quote Boström himself from his *New York Times* best seller [5]:

> The human brain has some capabilities that the brains of other animals lack. It is to these distinctive capabilities that our species owes its dominant position. If machine brains surpassed human brains in general intelligence, then this new superintelligence could become extremely powerful—possibly beyond our control. As the fate of the gorillas now depends more on humans than on the species itself, so would the fate of humankind depend on the actions of the machine superintelligence.

It turns out that the control problem, central to the current AI narrative, is the problem of controlling machines of the future that will be more intelligent and powerful than human beings, posing an existential risk to humankind. This fear is not some obscure viewpoint held by a small group of fringe quacks. Indeed,

Digital Object Identifier 10.1109/MCS.2016.2643238
Date of publication: 17 March 2017

it is a fear held by the likes of Stephen Hawking, Elon Musk, and Bill Gates [6].

Is this fear even worth worrying about? Surely the emergence of superintelligent machines won't happen anytime soon, definitely not within our own lifetimes. Or will it? Prominent futurists speak of something called the *technological singularity*, "the hypothesis that the invention of artificial superintelligence will abruptly trigger runaway technological growth, resulting in unfathomable changes to human civilization" [7]. The basic idea is that when machines reach an intelligence level that surpasses that of human beings, they will perpetuate a cycle of building ever-more-intelligent machines, a phenomenon that Kurzweil calls the "law of accelerating returns" [8]. The remarkable fact is that futurists predict the technological singularity to be some time in the surprisingly near future [7]:

> Ray Kurzweil predicts the singularity to occur around 2045 whereas Vinge predicts some time before 2030. At the 2012 Singularity Summit, Stuart Armstrong did a study of artificial general intelligence (AGI) predictions by experts and found a wide range of predicted dates, with a median value of 2040.

It behooves us as control theorists and engineers to take a closer look at the AI control problem. Apparently, in control terms, the AI control problem arises from the risk posed by the lack of controllability of machines. More specifically, the risk here is the instability (of sorts) of controllers. In essence, the control problem is one of controlling controllers. Surely this is a legitimate problem in our field of control. In fact, it's not even all that different, at least in principle, from the kind of control problems that we find in control textbooks.

Indeed, the control problem, as understood here, is quite familiar to us. When we have a plant we wish to control, and we design a controller for it, there is always the risk that the closed-loop system will not behave as desired. More specifically, it might be unstable, resulting in all kinds of undesirable consequences, ranging from saturated amplifiers to catastrophic loss of life. So what can we do? We can analyze the stability of the controller before even implementing. Alternatively, we can design a "higher-level" (supervisory) controller that stabilizes the closed-loop system. Two immediate questions arise. First, what if the closed-loop system is too complicated to be analyzed? Second, and worse, what if closed-loop system is, in some sense, fundamentally unstabilizable? These questions lie at the heart of the AI control problem.

There is a third question that we must ponder. What if the closed-loop system behaves exactly as designed, but the design approach itself causes undesirable but nonobvious consequences? This is a fear reflected by prominent AI commentators [9]:

> However, the real risk posed by AI—at least in the near term—is much more insidious. It's far more likely that robots would inadvertently harm or frustrate humans while carrying out our orders than they would become conscious and rise up against us.

Boström has a whimsical illustration of this [5]:

> An AI, designed to manage production in a factory, is given the final goal of maximizing the manufacture of paperclips, and proceeds by converting first the Earth and then increasingly large chunks of the observable universe into paperclips.

This predicament, sometimes called "perverse instantiation," should also be quite familiar to control theorists and engineers. Surely we have come across optimal controller designs with unintended consequences.

An article [10] discusses some lessons learned from the area of adaptive control and describes the distrust of adaptive controllers on the part of some control practitioners, not unlike the kind of distrust expressed by current AI commentators [10].

> ... we explain why such distrustfulness is warranted, by reviewing a number of adaptive control approaches which have proved deficient for some reason that has not been immediately apparent. The explanation of the deficiencies, which normally were reflected in unexpected instabilities, is our main concern. Such explanations, coupled with remedies for avoiding the deficiencies, are necessary to engender confidence in the technology.

Notice the similarities between the above and the AI control problem. The deficiencies were the result of the control approaches, they were not immediately apparent, and they were reflected in unexpected instabilities. Fortunately, [10] also discusses remedies to these problems. The real question is whether remedies can be found for the AI control problem. While this remains to be seen, it seems at least plausible that control theorists and engineers, researchers in our own community, have important contributions to be made to *the control problem*.

REFERENCES

[1] A. Davis, "How artificial intelligence has crept into our everyday lives," *The Institute*, June 8, 2016.

[2] R. Newman, "How robots paved the way for Donald Trump," *Yahoo Finance*, July 14, 2016.

[3] R. S. Sutton, "Artificial intelligence as a control problem: Comments on the relationship between machine learning and intelligent control," in *Proc. IEEE Int. Symp. Intelligent Control*, 1988, pp. 500–507.

[4] J. Dye, "Google outlines some key ways to keep an AI from taking over the world," *Android Authority*, June 22, 2016.

[5] N. Bostrom, *Superintelligence: Paths, Dangers, Strategies*. London, U.K.: Oxford Univ. Press, 2014.

[6] M. Sainato, "Stephen Hawking, Elon Musk, and Bill Gates warn about artificial intelligence," *Observer*, Aug. 19, 2015.

[7] [Online]. Available: https://en.wikipedia.org/wiki/Technological_singularity

[8] R. Kurzweil, *The Age of Spiritual Machines*. Viking Press, 1999.

[9] O. Solon, "The rise of robots: Forget evil AI—The real risk is far more insidious," *The Guardian*, Aug. 30, 2016.

[10] B. D. O. Anderson, "Failures of adaptive control theory and their resolution (dedicated to Professor Thomas Kailath on his 70th birthday)," *Commun. Inform. Syst.*, vol. 5, no. 1, pp. 1–20, 2005.

Edwin K.P. Chong