Journal of EVOLUTION & Technology

A peer-reviewed electronic journal published by the Institute for Ethics and Emerging Technologies
ISSN 1541-0099
26(1) – February 2016

Don't Worry about Superintelligence

Nicholas Agar
School of History Philosophy Political Science and International Relations
Victoria University of Wellington

nicholas.agar@vuw.ac.nz

Journal of Evolution and Technology - Vol. 26 Issue 1 – February 2016 - pgs 73-82

Abstract

This paper responds to Nick Bostrom's suggestion that the threat of a human-unfriendly superintelligence should lead us to delay or rethink progress in AI. I allow that progress in AI presents problems that we are currently unable to solve. However, we should distinguish between currently unsolved problems for which there are rational expectations of solutions and currently unsolved problems for which no such expectation is appropriate. The problem of a human-unfriendly superintelligence belongs to the first category. It is rational to proceed on that assumption that we will solve it. These observations do not reduce to zero the existential threat from superintelligence. But we should not permit fear of very improbable negative outcomes to delay the arrival of the expected benefits from AI.

Don't worry about superintelligence[1]

Nick Bostrom (2014) argues that we face an existential threat from superintelligence. Progress in artificial intelligence will predictably produce minds capable of improving their own cognitive powers. A process of iterated self-improvement should give rise to an intelligence explosion which will soon produce a superintelligence. This being would then have the capacity to send humanity extinct and could form the goal to do so. According to Bostrom, we desperately need a solution to the control problem – "the problem of how to control what the superintelligence would do" (2014, vii). Without a solution we cannot be confident that a superintelligence would act in ways that are compatible with human interests. Bostrom urges that we delay the arrival of human-level artificial intelligence until we solve the control problem.

In this paper I allow that we currently lack a solution to the control problem. It is, nevertheless, a problem that we have a rational expectation of solving. Since we should expect a solution to the control problem no delay in research on intelligent machines is necessary or desirable. This claim depends on a distinction between two types of currently unsolved technological problems. There are currently unsolved problems for which it is rational to expect solutions and currently unsolved problems for which no such expectation is appropriate. The control problem belongs to the first category. We know where we will find a solution – it will come from progress in AI. It is appropriate to proceed on the assumption that we will solve the control problem. Any delay in progress predictably delays the solution.

A rational expectation of solving the control problem does not reduce the probability of a human-unfriendly superintelligence to zero. However, a rational response to risk suggests that it is possible to overreact to bad but unlikely outcomes. We should not permit fear of a possible but very improbable unfriendly superintelligence to prevent us from acquiring the very great benefits likely to come from progress in artificial intelligence.

Bostrom on AI and existential risk

Bostrom presents superintelligence as a predictable consequence of current work in AI. He uses "the term 'superintelligence' to refer to intellects that greatly outperform the best current human minds across many very general cognitive domains" (Bostrom 2014, 52).

Work in artificial intelligence is currently progressing toward the goal of Artificial General Intelligence (AGI) – the capacity to perform any intellectual task performed by a human. Bostrom sides with those who hold that AGI is possible, though he allows that there is uncertainty about when we will achieve it. Bostrom is chiefly interested in what will happen once we achieve the goal of AGI, or get sufficiently close to it. He predicts an intelligence explosion, "an event in which, in a short period of time, a system's level of intelligence increases from a relatively modest endowment of cognitive capabilities (perhaps sub-human in most respects, but with a domain-specific talent for coding and AI research) to radical superintelligence" (Bostrom 2014, 29). Since we made the computer with these capacities, and could presumably make improvements to it, it follows that this AI can improve itself. These self-improvements will make it even better at improving itself. And so on. Pretty soon there should be an artificial superintelligence. According to Bostrom, "existential catastrophe" is "a plausible default outcome" of the resulting intelligence explosion (2014, 115). For "existential catastrophe" read "human extinction."

We need not suppose that AIs are malevolent to fear them. Rather, the danger they pose comes from an agnosticism on our part about their goals. Eliezer Yudkowsky, another doomsayer about superintelligent machines explains that "the AI does not hate you, nor does it love you, but you are made out of atoms which it can use for something else" (quoted in Barrat 2013, 22). When you spray insecticide on ants that have found their way onto your kitchen bench, you probably don't hate them. The ants are inconvenient. This is how a superintelligent AI could think about humans. We currently take up a lot of space on the planet. And we are a potential threat. According to Bostrom, should a superintelligence have goals that conflict with human interests, it could have both reason and the capacity to eliminate us. The superintelligence could arrange "to acquire an unlimited amount of physical resources and, if possible, to eliminate potential threats to itself and to its goal system. Human beings might constitute potential threats; they certainly constitute physical resources" (Bostrom 2014, 116).

Our inability to predict what an intelligence far more powerful than our own will want to do seems to make the control problem especially difficult. The immense challenge of controlling something more cognitively powerful than us is apparent from our success at thwarting natural selection's programming. Natural selection created us as vehicles for our DNA. It sought to control us by making sex pleasurable. But we have been sufficiently smart to get the sexual pleasure without serving nature's end. We invented contraception. If humans are sufficiently intelligent to circumvent nature's ends, then a superintelligent AI should be sufficiently so to circumvent ours.

We could seek to prevent this by programming human-friendly goals into AIs. A superintelligence that retains these goals will not seek human extinction. It will act in ways that promote our interests or are, at a minimum, compatible with them. Bostrom thinks that attempts to ensure that a superintelligence is human friendly are either likely to fail or, at least, bring no certainty of success. According to his orthogonality thesis – "more or less any level of intelligence could in principle be combined with more or less any final goal" (Bostrom 2014, 107). The orthogonality thesis suggests that sending humanity extinct could either be adopted by a superintelligence as a final goal, or be recognized as instrumental to achieving some other final goal. Suppose that we do succeed in

instilling goals that seem to us to be human friendly. Bostrom observes that there are "perverse instantiations" of such goals. These perverse instantiations satisfy our stated goals in ways that turn out badly for us. Bostrom offers the example of the goal "make us smile." His suggested perverse instantiation is "Paralyze human facial musculatures into constant beaming smiles." The goal "make us happy" is perversely instantiated by "Implant electrodes into the pleasure centers of our brains" (Bostrom 2014, 120).

According to Bostrom, there is a limited window in which we must try to program human friendliness into AIs. We can program the machines that will predictably give rise to the first AGI. But a superintelligent descendant of the first AGI will resist attempts to program human-friendly desires into it much in the way that humans resist brainwashing. Bostrom warns:

> Before the prospect of an intelligence explosion, we humans are like small children playing with a bomb. Such is the mismatch between the power of our plaything and the immaturity of our conduct. Superintelligence is a challenge for which we are not ready now. (2014, 259)

Why we should expect to solve the control problem

Bostrom exaggerates the difficulty of the control problem. The dialectic he describes opposes the strategies of humans aspiring for control with the responses of an artificial superintelligence seeking to realize its goals. Conceived this way, it seems unlikely that we will ever solve the control problem. It's not just a matter of our not being ready *now* to respond to machines smart enough to quickly make themselves even smarter. Bostrom presents himself as calling for us to delay the arrival of AGI. However, if we accept his presentation of the control problem, then complete abandonment of AI seems to be our only effective response.

The book's central chapters consider a variety of ways to constrain or appropriately motivate a superintelligence. Here Bostrom takes the part of the superintelligence, showing how it could circumvent any attempt to make it friendly or to prevent it from achieving its unfriendly goals. The short answer is that you cannot solve the control problem because the thing you're trying to control is superintelligent and you're not. The superintelligence does what it wants and we are powerless to stop it.

Consider the strategies that Bostrom groups together under the heading of "boxing methods." Essentially these involve isolating an evolving AGI to prevent it from causing us harm. Suppose that we deprive a boxed AI of external manipulators to prevent it from acting in the external world. Bostrom proposes that the AI might circumvent this measure "even when it lacks access to external manipulators, simply 'by thinking' (that is, by shuffling the electrons in its circuitry in particular patterns)" thereby generating radio waves (Bostrom 2014, 130). Furthermore, the idea of boxing an AI seems to defeat the very idea of building it. We want it to find solutions to our problems and to be able to communicate these solutions to us. Herein lies the opportunity for the superintelligence to escape its box:

> If the AI can persuade or trick a gatekeeper to let it out of the box, resulting in its gaining access either to the Internet or directly to physical manipulators, then the boxing strategy has failed. Human beings are not secure systems, especially not when pitched against a superintelligent schemer and persuader. (Bostrom 2014, 130)

There seem to be strong inductive grounds for thinking that a superintelligent AI will circumvent any strategy adopted by far less intelligent human beings seeking to control it. In effect we are playing a very complex strategy game against a player that is vastly more intelligent than us. A move may seem excellent when evaluated according to standards appropriate to our human intellects. We should nevertheless feel confident that a much more intelligent adversary will find an effective countermeasure.

This is not the correct way to assess the threat from superintelligence. Bostrom conflates two ways to describe this threat.

> [A] Progress in AI is likely to yield a superintelligence that will form the goal to send humanity extinct. Once formed there's likely to be little that we can do to prevent it from realizing that goal.
>
> [B] If a superintelligence forms the goal to send humanity extinct then there's likely to be little that we can do to prevent it from realizing that goal.

Statement A supports delay in research on AI. However, Bostrom's various thought experiments offer no support for A. They do offer support to the conditional statement B. But the truth of B does not suggest the need to delay progress in AI. In what follows I argue we can falsify the antecedent of B. Trends in AI are unlikely to yield artificial superintelligences that form the goal of sending humanity extinct. I allow that we cannot reduce to zero the probability of a human-unfriendly superintelligence. The control problem is, nevertheless, something that we have a rational expectation of solving. We shouldn't be too concerned that we do not currently have a solution, because we know where its solution will come from. It will come from progress in AI. The same technological trends that generate the problem solve it.

We should distinguish between two types of currently unsolved technological problems: those that are currently unsolved but which we should expect to solve; and those that are currently unsolved for which there is no rational expectation of a solution.

[1] Technological problems that are currently unsolved but which we should expect to solve

Typically we have some awareness of where the solutions will come from. Our expectations may be supported by established trends in technology. Sometimes we can describe the solutions in sufficient detail that we have a good sense of how to find them. These rational expectations permit us to proceed on the assumption that we will solve problems. A rational expectation of a solution is not a logical guarantee. There may, for example, be some physical law that will prevent us from solving the problem. Or a solution that at first seemed straightforward may turn out to be so complex that it is beyond human intellects. But we may be justified in believing it unlikely that there is a law that would prevent a solution. Moreover, we may understand the problem sufficiently well to make it reasonable to believe that we are intelligent enough to solve it.

The problem of how to make more powerful computers falls into this category. Moore's Law and related generalizations support the claim that the most powerful computers in ten years' time will be more powerful than the most powerful computers today. We do not know how to make these more powerful computers today – if we did we would make them instead of the less powerful ones that we actually make. But we have well-supported ideas about where solutions to the problem of how to make more powerful computers will come from.

Another example of a currently unsolved but predictably soluble problem is that of significantly extending the range of the batteries that power electric cars. In 2015 the Tesla Model S 85D was the best performer at 435 kilometers on a single charge.[2] The problem of making better batteries is exceptionally challenging. Suppose that no one in January 2016 can make a battery capable of powering a car over 600 kilometers on a single charge. We nevertheless know in general terms where the solution will come from. There's an important sense in which, to solve the problem, battery designers have only to keep on doing the kinds things that they are currently doing. A rational expectation is no logical guarantee. But it seems unlikely that there are yet-to-be discovered physical laws that will limit the batteries that can be fitted in a car to ranges of less than 500 kilometers.

A feature of these unsolved but predictably soluble problems is that we should proceed with an expectation of solving them. Those who use computers to solve very challenging problems should

expect that the future will bring computers more powerful than those that exist today. Cancer researchers whose computers are not quite powerful enough to analyze complex patterns in who gets melanoma and who doesn't can look forward to more powerful machines to apply to the problem. Urban developers trying to work out good locations for future charging stations for electric vehicles are justified in assuming that the batteries of the future will power cars over longer distances than today's batteries.

[2] Technological problems that are currently unsolved for which there is no rational expectation of a solution

Consider the problem of recovering a live specimen of a Cretaceous Period Tyrannosaurus. Suppose you grant that travel backwards in time is not a logical impossibility. Technologies capable of transporting people or devices back in time and returning both them and a captive Tyrannosaur to the present may not violate any physical law and may therefore be discoverable. This problem nevertheless falls into a different category from those described above. There are no existing technological trends that point to the required time-travel technologies. We have only the vaguest and most technologically unspecific ideas about how to go back in time and retrieve a genuine Cretaceous Period Tyrannosaur. Technological progress does occasionally deliver surprises. But it would be an odd allocation of resources to begin constructing cages capable of containing large carnivorous theropods on the assumption that we will soon acquire the necessary technologies.

The control problem belongs squarely in the first category. It is certainly difficult. As Bostrom's book demonstrates, we currently lack a solution to it. But existing trends in the development of AI give reason to expect a solution.

Moravec's Paradox

Bostrom presents the intelligence explosion as a consequence of AGI. An artificial intelligence with capacities sufficiently close to our own could make itself more intelligent, unleashing further powers of self-improvement. According to the orthogonality thesis, a superintelligence "could in principle be combined with more or less any final goal." Once a superintelligence has formed a human-unfriendly goal, it's unlikely that we could do much about it.

Solutions to the control problem emerge from the way in which we will achieve AGI. Progress toward AGI has been real and impressive, but uneven. This unevenness suggests solutions. Moravec's Paradox, named for the computer scientist Hans Moravec, captures one aspect of what I will call the challenge from agency (Moravec 1990). The paradox is pithily summarized by Steven Pinker as the idea that in AI "the hard problems are easy and the easy problems are hard" (Pinker 2007, 190). Thinking several moves ahead in a game of chess is difficult for a human. It has been unexpectedly easy to program computers to do this. Chess computers now beat the best human players. Practical tasks, especially those connected with sensorimotor abilities, the kinds of tasks that humans perform effortlessly, have proved very challenging. To pass the famous Turing Test, a machine is expected to produce verbal behavior indistinguishable from that of a human. Ben Goertzel, Matt Ikle, and Jared Wigmore describe the "Coffee Test" which they credit to the Apple cofounder Steve Wozniak. This test targets practical capacities. To pass the Coffee Test a robot must "go into an average American house and figure out how to make coffee, including identifying the coffee machine, figuring out what the buttons do, finding the coffee in the cabinet, etc" (Goertzel et al. 2012, 17). Machines that pass the test will not be thrown by the many ways in which homes and coffee makers vary.

Many of the tasks that are easy for us but hard for computers relate to agency – they involve interacting with the world in ways that promote our goals.[3] It's not surprising that we are good at such tasks. In biological systems, agency coevolved with the capacity to represent the world. Natural selection is interested in representing and thinking as tools for agency. Single-cell organisms evolved the capacity to detect sugar so that they could locate and consume it. More complex representational activities evolved to enable more complex behaviors. We and other evolved complex organisms are

the beneficiaries of all of this. Our effortless ability to make coffee is a consequence of many millions of years of coevolution of agency and the capacities to represent and reason about the world. There has been no such coevolution in computers – mainly because it wasn't required. We've always treated computers as our tools – as with other tools we supply the agency. You don't need a computer to get the point of an Excel spread sheet anymore than you need a hammer to understand what's so good about hammering in nails. It matters only that humans find such spread sheets useful and want nails to be hammered.

There's no reason to think that these practical, sensorimotor abilities will be forever beyond machines. Skynet, the malign superintelligence from the Terminator movies, could almost certainly make a killer espresso if it wanted to. But the fact that artificial agency is very likely to arrive in an incremental way offers humans ample protection from unfriendly AIs.

Degrees of agency

Agency is not an on/off concept. It comes in degrees. We can understand degrees of agency as determined by agents' capacities to act in the world they represent and to form plans in respect of it. I will call agents with high degrees of agency *sophisticated agents*. These agents are comparatively good at forming plans in respect of the world that they represent and acting to achieve these goals. We can call systems with low degrees of agency *clumsy agents*. Clumsy agents may be capable of complex representations of the world, but these representations generate only limited behaviors in response to it.

In 2011, the IBM computer Watson defeated the world's best human *Jeopardy!* players. Watson is a clumsy agent. It can represent many aspects of the world. But it is not particularly competent at forming goals in respect of that world or acting in it to satisfy them. It can, for $2000, offer "Christchurch" as the answer to "It's New Zealand's second-largest city."[4] But it cannot make and execute a plan to visit Christchurch. Watson is not entirely devoid of agency – victory in *Jeopardy!* required it to form plans in respect of its vast database of information – but its agency did not reach out to the states of the world represented by information in that database. We can contrast Watson's high-level clumsiness with the lower-level physical clumsiness of people to whom we traditionally apply the term "clumsy." Clumsy humans typically have no difficulty in formulating complex plans in respect of their environments. They may sometimes stumble because they fail to notice physical obstacles or, having noticed them, fail to respond adequately to their presence.

There's a sense in which Watson knew more than any of its human competitors. Watson's responses in *Jeopardy!* required some agency. It had to make choices about how to parse a question and how to search its database. But it is a long way off the sophisticated agency of its vanquished human opponents. Watson's opponents walked away from the contest and made plans to do other things. Someone flicked the off switch on Watson. The quip of the *Jeopardy!* champion, Ray Jennings, upon being bested by Watson, "I, for one, welcome our new computer overlords" suggests that the victory of a machine in this intellectually demanding game might presage some big change in the relationship between humans and machines. But without significant enhancements of Watson's powers of agency there is no danger of impending computer overlordship.

This is not to deny that high degrees of agency could be added to computers. It's reasonable to suppose that work in AI will eventually yield an artificial general intelligence. This AGI will be capable of the same feats of agency as its human creators. But when AIs do make good the agency gap it's likely that they will do so in ways that are human friendly. Don't think Skynet, instead think Data from the universe of *Star Trek*.

Why the first AIs with sophisticated agency are likely to be friendly

A pattern of development more probable than that suggested by Bostrom or depicted in the Terminator movies involves a future in which AIs become progressively more autonomous in ways

that are appropriately responsive to human needs and interests. There is a predictable path to a human-friendly superintelligence.

One of the features of intelligence explosion that most preoccupies Bostrom and Yudkowsky is that it's not a problem that we get to have many attempts at. In the Terminator movies, humans don't get to approach a newly self-aware Skynet and request a do over. One minute Skynet is uncomplainingly complying with all human directives. The next, it's nuking us. I suspect that we are likely to have plenty of opportunities for do overs in our attempts to make autonomous AIs. Autonomy is not an all-or-nothing proposition. The first machine agents are likely to be quite clumsy. They may be capable of forming goals in respect of their world but they won't be particularly effective at implementing them. This gives us plenty of opportunity to tweak their programming as they travel the path from clumsy to sophisticated agency.

In one of the most prescient of his speculations about the future of intelligent computers Alan Turing offered a suggestion about how we might give computers the kind of knowledge about the world that we possess. He proposed that we should design intelligent machines to be like very young humans (Turing 1950). Turing said "Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's? If this were then subjected to an appropriate course of education one would obtain the adult brain" (1950, 456). Rather than programming in all of the knowledge possessed by a human adult, we would program in the capacity to learn it. We should take Turing's advice in respect of the machine's moral development. His suggestion indicates how we might provide moral instruction to AIs. As AIs evolve from clumsy to sophisticated agents we can guide their moral development.

According to Bostrom's orthogonality thesis we cannot prevent an autonomous AI from desiring to act in unfriendly ways. That same warning applies to human beings. The existence of psychopathic murderers tells us that extreme unfriendliness toward other humans is compatible with the basic architecture of human autonomy. Parents are not powerless to prevent these bad outcomes. They can seek to intervene as their children's capacities for autonomous action expand. If you notice malevolent behavior in your child you can seek to correct it.

The clumsiness of the first generations of autonomous AIs gives us the opportunity to correct unfriendly goals. Suppose that a glitch in the programming of a floor cleaning robot leads it to seek out human-sized objects and ram them. This human-unfriendly behavior is easily addressed. Other clumsy AIs will have more knowledge about the world that we and they inhabit. But they are unlikely to be particularly good at forming plans relevant to that world. We are likely to notice clumsy malevolent acts. Suppose a machine intelligence tasked with gathering and analyzing data about the causes of cancer forms the goal to give humans cancer. The successful execution of this plan requires knowledge far beyond the causes of cancer. It requires knowledge about how we would respond to a computer that abruptly begins advising nearby humans to increase their exposure to tobacco and sunbeds. A human-unfriendly clumsy computer system may successfully execute its human operator by cutting off the supply of oxygen to the room in which she is working. But when it seeks to initiate an exchange of nuclear weapons it must thwart the countermeasures of the humans who currently control those weapons. This requires complex practical knowledge about the capacities and motivations of those human controllers.

What are we to make of the imaginative exercises in which Bostrom takes the part of the unfriendly superintelligence determined to circumvent our attempts at control? Skynet is an imaginable endpoint for AI. We can also imagine a grumpy child maturing into a nuclear-armed psychopath. The ease with which we can imagine these outcomes does not mean that we are powerless to influence AIs or children for the better. We understand that a desire to kill large numbers of people is compatible with the basic motivational machinery of humans. Parents seek to avoid this by raising their children properly. There's no reason we should not take the same attitude into the incremental process of turning clumsy computer agents into sophisticated ones.

Computer engineers seeking to correct the behavioral glitches of clumsy AIs are likely to find them less recalcitrant than morally aberrant young humans. Parents and psychologists aren't always successful when they seek to intervene in the behavior of errant teenagers. They don't have access to the programming language of human teenagers' brains. We will, in contrast, be able to reprogram the first, clumsy autonomous AIs. Human programmers, assisted by non-autonomous AIs, can seek to rewrite any offending lines of code. Thus there's a process by which we can appropriately socialize AIs. This path is incremental. It should see improvements in the moral motivations of the machines as their agency becomes more sophisticated. It's not surprising that we cannot solve Bostrom's control problem in advance of our construction of relatively sophisticated computer agents. It's very difficult to work out, from scratch, how one could program a superintelligent AI to be nice to humans. But many problems that we can't solve all in one go become soluble if we tackle them incrementally. Autonomous AIs should evolve in response to the needs and interests of their human designers.

As artificial agents become less clumsy they are likely to act as checks on each other. Bostrom describes the pressure toward a singleton AI (Bostrom 2014, ch 5). In this picture, a variety of AIs may undergo an intelligence explosion and evolve quickly toward superintelligence. One of them gains a decisive strategic advantage over its competitors and purges, deletes, or absorbs them. Our fates depend on the friendliness of this singleton. A scenario in which a variety of AIs become incrementally more powerful agents suggests a different outcome. There is no reason to expect a single AI with a global decisive strategic advantage over all other AIs. Suppose that there are many varyingly sophisticated artificial agents whose development is guided by humans. This community of artificial agents should be expected to act as a check on unfriendly outliers much in the way that communities of humans restrain malevolent members.

We should expect to be able to morally educate AIs. This does not suggest that we have nothing to fear in respect of their development. But our concern should be about malevolent humans who acquire control over hugely knowledgeable but clumsy AIs. This fear is akin to that of malevolent humans gaining control over any powerful, potentially destructive technology. Note that a program of morally educating AIs responds to this threat. A powerful, morally educated AI is unlikely to comply with the instruction of a malevolent human owner that it kill humans.

Balancing the benefits and harms of artificial intelligence

I have argued that we should expect to solve the control problem. Admittedly, a rational expectation that the problem will be solved does not reduce the probability of a human-unfriendly superintelligence to zero. Progress in AI boosts a risk of extinction from a human-unfriendly superintelligence from zero to low. If we don't create AGI then there is no chance of an intelligence explosion leading to an unfriendly superintelligence. By continuing with research on AI we expose ourselves to a novel extinction threat. Should we accept even a small risk of existential catastrophe from research on AI?

An answer to this question depends on what we expect to get out of progress in AI. It is instructive to compare our responses to collective existential risk with how we respond to personal existential threats. It's hard to think of any choice that we make that does not bring some risk of death. If you choose to take a morning jog you increase your risk of imminent death from a heart attack or stroke. Does this spike in the risk of imminent death make a morning jog irrational? An answer to that question depends on the benefits you might derive from an exercise program. For ostensibly healthy people it can be prudentially rational to go out for the jog. The benefits of an appropriate exercise program can more than compensate for a temporary spike in your risk of death from heart attack or stroke.

Similar points can be made on behalf of research in AI. The increasing computational powers brought by progress in AI promise many benefits. Many of these potential benefits come not from the side of AI that is interested in building conscious robots – the side that has tended fascinate philosophers –

but rather from the side of AI that is interested in finding solutions to problems whose complexity seems to be beyond the computational limits of human intellects.

I offer one example of this potential benefit. The machine learning expert Pedro Domingos imagines a future approach to cancer that he calls CanceRx (Domingos 2015, 259–61). CanceRx would apply insights about computer learning to the vast quantities of data about the disease that we are beginning to accumulate. It seeks out patterns in data about who gets cancer and who doesn't and how different cancers respond to different therapies. The mutability of cancer frustrates those who try to treat it. But CanceRx will be built for this. We shouldn't expect an endpoint when CanceRx serves up a convenient Cure for Cancer. Rather, "The model is continually evolving, incorporating the results of new experiments, data sources, and patient histories" (Domingos 2015, 259). Domingos says: "Because every cancer is different, it takes machine learning to find the common patterns. And because a single tissue can yield billions of data points, it takes machine learning to figure out what to do for each new patient" (2015, 261).

We face many complex problems. The problems of climate change and global poverty are characterized by complexities that machine learning can help with. This is not to suggest that solutions will come solely from advances in machine learning. But it is reasonable to expect that advances in machine learning will help. The expected magnitude of these benefits justifies the novel variety of risk from superintelligence. The fact that the benefits from research in AI justify some degree of risk does not mean that one should not take reasonable steps to reduce that risk. But alarmist rhetoric is unhelpful. (See Goertzel and Pitt 2012 for some constructive suggestions.) We should aim for a rational management of future technological risks that appropriately balances risks against potential benefits.

Concluding comments

Human-unfriendly superintelligences are good fodder for science fiction movies. But that does not mean we should worry much about them. A more probable path of development suggests the very gradual emergence of sophisticated artificial agency that is informed by and responsive to human interests. This does not reduce the threat from unfriendly artificial superintelligence to zero. So it's good that some people are currently concerned about this threat, just as it's good that some people are worried about the potential threat of extra-terrestrial invasion. But we should not permit these somewhat eccentric concerns to distract our focus from the many problems that work in AI can contribute to.

Notes

1. Thanks to Russell Blackford, Simon Keller, David Lawrence, and Nick Smith for comments on earlier versions of this paper.

2. Schaal 2015.

3. One person who is not particularly scared about the arrival of a superintelligence is the pioneering roboticist Rodney Brooks. The title of a piece "Artificial intelligence is a tool, not a threat" (Brooks 2014) neatly summarizes his view about the future benefits and dangers of artificial intelligence. Brooks uses one of his robotic creations, the floor-cleaning Roomba robot, to illustrate the low risk from research in AI and robotics.

4. See the *Jeopardy!* Archive (n.d.) for show #6087, originally broadcast on February 15, 2011.

References

Barrat, J. 2013. *Our final invention: Artificial intelligence and the end of the human era*. London: Macmillan 2013.

Bostrom, N. 2014. *Superintelligence: Paths, dangers, strategies*. Oxford: Oxford University Press.

Brooks, R. 2014. Artificial intelligence is a tool, not a threat. Rethink Robotics (blog). http://www.rethinkrobotics.com/blog/artificial-intelligence-tool-threat/ (accessed January 19, 2016).

Domingos, P. 2015. *The master algorithm: How the quest for the ultimate learning machine will remake our world*. London: Allen Lane.

Goertzel, B., M. Ikle and J. Wigmore. 2012. The architecture of human-like general intelligence. In *Theoretical Foundations of Artificial General Intelligence*, volume 4 of the series Atlantis Thinking Machines, 123–44. Amsterdam: Atlantis Press.

Goertzel. B. and J. Pitt. 2012. Nine ways to bias open-source AGI toward friendliness. *Journal of Evolution and Technology* 22(1) (February): 116–31.
Available at http://jetpress.org/v22/goertzel-pitt.pdf (accessed January 19, 2016).

Jeopardy! Archive. n.d. Show #6087.
http://j-archive.com/showgame.php?game_id=3576 (accessed January 19, 2016).

Moravec, H. 1990. *Mind children: The future of robot and human intelligence*. Cambridge MA: Harvard University Press.

Pinker, S. 2007. *The language instinct: How the mind creates language*. New York: William Morrow & Co.

Schaal, Eric. 2015. 10 electric vehicles with the best range in 2015. Autos CheatSheet. November 12. http://www.cheatsheet.com/automobiles/top-10-electric-vehicles-with-the-longest-driving-range.html/?a=viewall (accessed January 19, 2016).

Turing, A. 1950. Computing machinery and intelligence. *Mind* 59: 433–60.