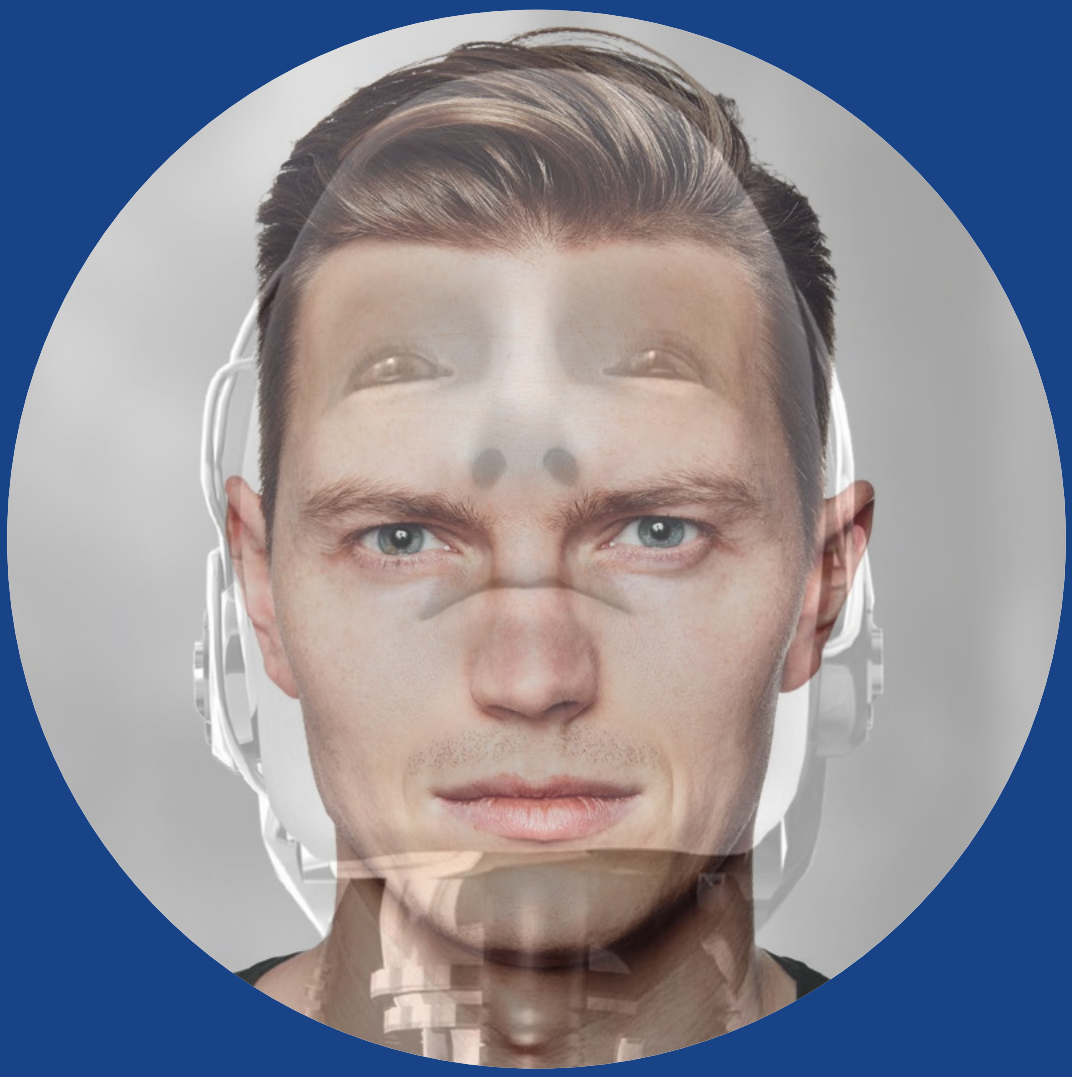


The Next Step

EXPONENTIAL LIFE

Provably Beneficial Artificial Intelligence

STUART RUSSELL

Opening image:
Detail of an image from one of the videos by the German electronic music pioneers Kraftwerk during the performance that opened their retrospective *The Catalogue 12345678* in Tate Modern's Turbine Hall on February 6, 2013. London, UK.

Stuart Russell
University of California, Berkeley, CA, USA

Stuart Russell received his BA with first-class honors in Physics from Oxford in 1982, his PhD in Computer Science from Stanford in 1986, and then joined the faculty of the University of California at Berkeley. He is a Professor (and former Chair) of Electrical Engineering and Computer Sciences and holds the Smith-Zadeh Chair in Engineering. He is a Fellow of AAAI (Association for the Advancement of Artificial Intelligence), ACM (Association for Computing Machinery), and AAAS (American Association for the Advancement of Science); and winner of the Computers and Thought Award and the ACM Karl Karlstrom Outstanding Educator Award. From 2012 to 2014 he held the Blaise Pascal Chair and the Senior Chair of Excellence of the ANR (French National Research Agency) in Paris. His book *Artificial Intelligence: A Modern Approach* (with Peter Norvig) is the standard text in AI; it has been translated into thirteen languages and is used in over 1300 universities in 116 countries.

MORE ABOUT THE AUTHOR [+]

Should we be concerned about long-term risks to humanity from superintelligent AI? If so, what can we do about it? While some in the mainstream AI community dismiss these concerns, I will argue instead that a fundamental reorientation of the field is required. Instead of building systems that optimize arbitrary objectives, we need to learn how to build systems that will, in fact, be provably beneficial for us. I will show that it is useful to imbue systems with explicit uncertainty concerning the true objectives of the humans they are designed to help.

INTRODUCTION

The goal of artificial intelligence (AI) research has been to understand the principles underlying intelligent behavior and to build those principles into machines that can then exhibit such behavior. In the field's early years, several distinction definitions of "intelligent" were pursued, including emulation of human behavior and the capability for logical reasoning; in recent decades, however, a consensus has emerged around the idea of a *rational agent* that perceives and acts in order to maximally achieves its objectives. Subfields such as robotics and natural-language understanding can be understood as special cases of the general paradigm. AI has incorporated probability theory to handle uncertainty, utility theory to define objectives, and statistical learning to allow machines to adapt to new circumstances. These developments have created strong connections to other disciplines that build on similar concepts, including control theory, economics, operations research, and statistics.

Progress in AI seems to be accelerating. In the last few years, due in part to progress in machine learning, tasks such as speech recognition, object recognition, legged locomotion, and autonomous driving have been largely solved. Each advance in capabilities brings with it new potential markets and new incentives to invest in further research, resulting in a virtuous cycle pushing AI forward. Within the next decade we are likely to see substantial progress on effective language understanding, leading to systems capable of ingesting, synthesizing, and answering questions about the sum total of human knowledge.

Despite all this progress, we are still far from human-level AI. For example, we have no practical methods for inventing useful new concepts such as "electron" or useful new high-level actions such as "write slides for tomorrow's lecture." The latter capability is particularly important for systems operating in the real world, where meaningful goals may require billions of primitive motor-control actions to achieve. Without the ability to conceive of and

reason about new, high-level actions, successful planning and acting on these timescales is impossible. Undoubtedly there are more breakthroughs needed that we will not know how to describe until we see our best efforts to build general-purpose AI systems failing in interesting ways. The difficulty in predicting such breakthroughs means that giving any precise estimate of the date on which human-level AI will arrive is foolhardy. Nonetheless, most experts believe it is likely to arrive within the present century (Müller and Bostrom, 2016; Etzioni, 2016).

It is hard to overstate the significance of such an event. Everything our civilization offers is a consequence of our intelligence; thus, access to substantially greater intelligence would constitute a discontinuity in human history. It might lead to solutions for problems of disease, war, and poverty; at the same time, several observers have pointed out that superintelligent AI systems can, by their very nature, have impacts on a global scale—impacts that could be negative for humanity of the systems are not designed properly.[1] The game is to define the problem that our AI systems are set up to solve, such that we are guaranteed to be happy with the solutions; and the stakes could hardly be higher.

RISKS AND REBUTTALS

Concerns about superintelligent AI are hardly new. Turing himself, in a 1951 radio address, felt it necessary to point out the possibility:

> If a machine can think, it might think more intelligently than we do, and then where should we be? Even if we could keep the machines in a subservient position, for instance by turning off the power at strategic moments, we should, as a species, feel greatly humbled. … [T]his new danger … is certainly something which can give us anxiety.

I. J. Good (1965), who had worked with Turing during World War II, went one step further, pointing out the possibility of self-improving AI systems: "There would then unquestionably be an 'intelligence explosion,' and the intelligence of man would be left far behind." The *AI control problem*, then, is how to ensure that systems with an arbitrarily high degree of intelligence remain strictly under human control.

It seems reasonable to be cautious about creating something far more intelligent than ourselves; yet we need more than a general sense of unease if we are to channel in the right direction the relentless scientific and economic pressure to build ever-more-capable systems. Many novels and films have

translated this unease into scenarios of spontaneously evil machine consciousness, which is both vanishingly unlikely and, as a technical phenomenon to be avoided, impossible to address. In fact, to the extent that we understand the problem at all, the most likely source of difficulty appears to be a failure of *value alignment*—we may, perhaps inadvertently, imbue machines with objectives that are imperfectly aligned with our own. Norbert Wiener (1960) put it this way: "If we use, to achieve our purposes, a mechanical agency with whose operation we cannot interfere effectively... we had better be quite sure that the purpose put into the machine is the purpose which we really desire."

Unfortunately, neither AI nor other disciplines built around the optimization of objectives (economics, statistics, control theory, and operations research) have much to say about how to identify the purposes we really desire. Instead, they assume that objectives are simply implanted into the machine. AI studies the capacity to achieve objectives, not the design of those objectives. And as King Midas found out, getting what one asks for is not always a good thing.

Bostrom (2014) elaborates on several additional arguments suggesting that the problem has no easy solutions. The most relevant for the analysis in this paper is due to Omohundro (2008), who observed that intelligent entities will tend to act to preserve their own existence. This tendency has nothing to do with any self-preservation instinct or other biological notion; it is just that an entity cannot achieve its objectives if it is dead. This means that Turing's reliance on the off-switch, as quoted above, is misplaced: according to Omohundro's argument, a superintelligent machine will take steps to disable the off-switch in some way. Thus we have the prospect of superintelligent machines, whose actions are (by definition) unpredictable by mere humans, whose imperfectly and incompletely specified objectives may conflict with our own, and whose motivation to preserve their own existence in order to achieve those objectives may be insuperable.

A number of objections have been raised to these arguments, primarily by researchers within the AI community. The objections reflect a natural defensive reaction, coupled

perhaps with a lack of imagination about what a superintelligent machine could do. None appear to hold water on closer examination. (If some of the objections seem *prima facie* absurd, rest assured that several even more absurd ones have been omitted to spare their originators embarrassment.) Several of the objections appear in the recent AI100 report by Stone et al. (2016), while others have been made by individuals participating in panel discussions at AI conferences:

Human-level AI is impossible. This is an unusual claim for AI researchers to make, given that, from Turing onward, AI researchers have been fending off such claims from philosophers and mathematicians. The claim, which is backed by no arguments or evidence, appears to concede that if superintelligent AI *were* possible, it *would* be a significant risk. It is as if a bus driver, with all of humanity as passengers, said, "Yes, I'm driving toward a cliff, but trust me, we'll run out of gas before we get there!" The claim also represents a foolhardy bet against human ingenuity. We have made such bets before and lost. On September 11, 1933, renowned physicist Ernest Rutherford stated, with utter confidence: "Anyone who expects a source of power in the transformation of these atoms is talking moonshine." On September 12, 1933, physicist Leo Szilard invented the neutron-induced nuclear chain reaction. A few years later, having demonstrated such a reaction in his laboratory, Szilard wrote: "We switched everything off and went home. That night, there was very little doubt in my mind that the world was headed for grief."

It's too soon to worry about it. The right time to worry about a potentially serious problem for humanity depends not on when the problem will occur, but on how much time is needed to devise and implement a solution that avoids the risk. For example, if we were to detect a large asteroid predicted to collide with the Earth in 2066, would we say it is too soon to worry? And if we consider the global catastrophic risks from climate change, which are predicted to occur later in this century, is it too soon to take action to prevent them? On the contrary, it may be too late. The relevant timescale for human-level AI is less predictable, but of course that means it, like nuclear fission, might arrive considerably sooner than expected.

It's like worrying about overpopulation on Mars. This is an interesting variation on "too soon to worry," one that appeals to a convenient analogy: not only is the risk easily managed and far in the future, but also it is extremely unlikely we would even *try* to move billions of humans to Mars in the first place. The analogy is a false one, however. We *are already* devoting huge scientific and technical resources to creating ever-more-capable AI systems. A more apt analogy would be a plan to move the human race to Mars with no consideration for what we might breathe, drink, or eat once we arrive.

Human-level AI is really not imminent, so we needn't worry. This is another variation on "too soon to worry," but one that attributes concerns about AI control to the false belief that superintelligent AI is imminent. This objection simply mis-states the reasons for concern, which are not predicated on imminence. For example, Bostrom (2014) writes: "It is no part of the argument in this book that we are on the threshold of a big breakthrough in artificial intelligence, or that we can predict with any precision when such a development might occur."

We're the experts, we build the AI systems, trust us. This objection is usually accompanied by disparagement of those raising concerns as being ignorant of the realities of AI. While

it is true that some public figures who have raised concerns, such as Elon Musk, Stephen Hawking, and Bill Gates, are not AI researchers, they are hardly unfamiliar with scientific and technological reasoning. And it would be hard to argue that Turing (1951), Wiener (1960), Good (1965), and Minsky (1984) are unqualified to discuss AI.

You're just Luddites. Musk, Gates, Hawking, and others (including, apparently, the author) received the 2015 Luddite of the Year Award from the Information Technology Innovation Foundation. It is an odd definition of Luddite that includes Turing, Wiener, Minsky, Musk, and Gates, who rank among the most prominent contributors to technological progress in the twentieth and twenty-first centuries. Furthermore, the epithet represents a complete misunderstanding of the nature of the concerns raised and the purpose for raising them. It is as if one were to accuse nuclear engineers of Luddism if they point out the need for control of the fission reaction. Some objectors also use the term "anti-AI," which is rather like calling nuclear engineers "anti-physics." The purpose of understanding and preventing the risks of AI is to ensure that we can realize the benefits. Bostrom (2014), for example, writes that success in controlling AI will result in "a civilizational trajectory that leads to a compassionate and jubilant use of humanity's cosmic endowment."

Your doom-and-gloom predictions fail to consider the potential benefits of AI. If there were no potential benefits of AI, there would be no economic or social impetus for AI research, and hence no danger of ever achieving human-level AI. This objection is like accusing nuclear engineers who work on containment of never considering the potential benefits of cheap electricity. The sad fact is that the potential benefits of nuclear power have largely failed to materialize precisely because of insufficient attention paid to containment risks at Three-Mile Island and Chernobyl.

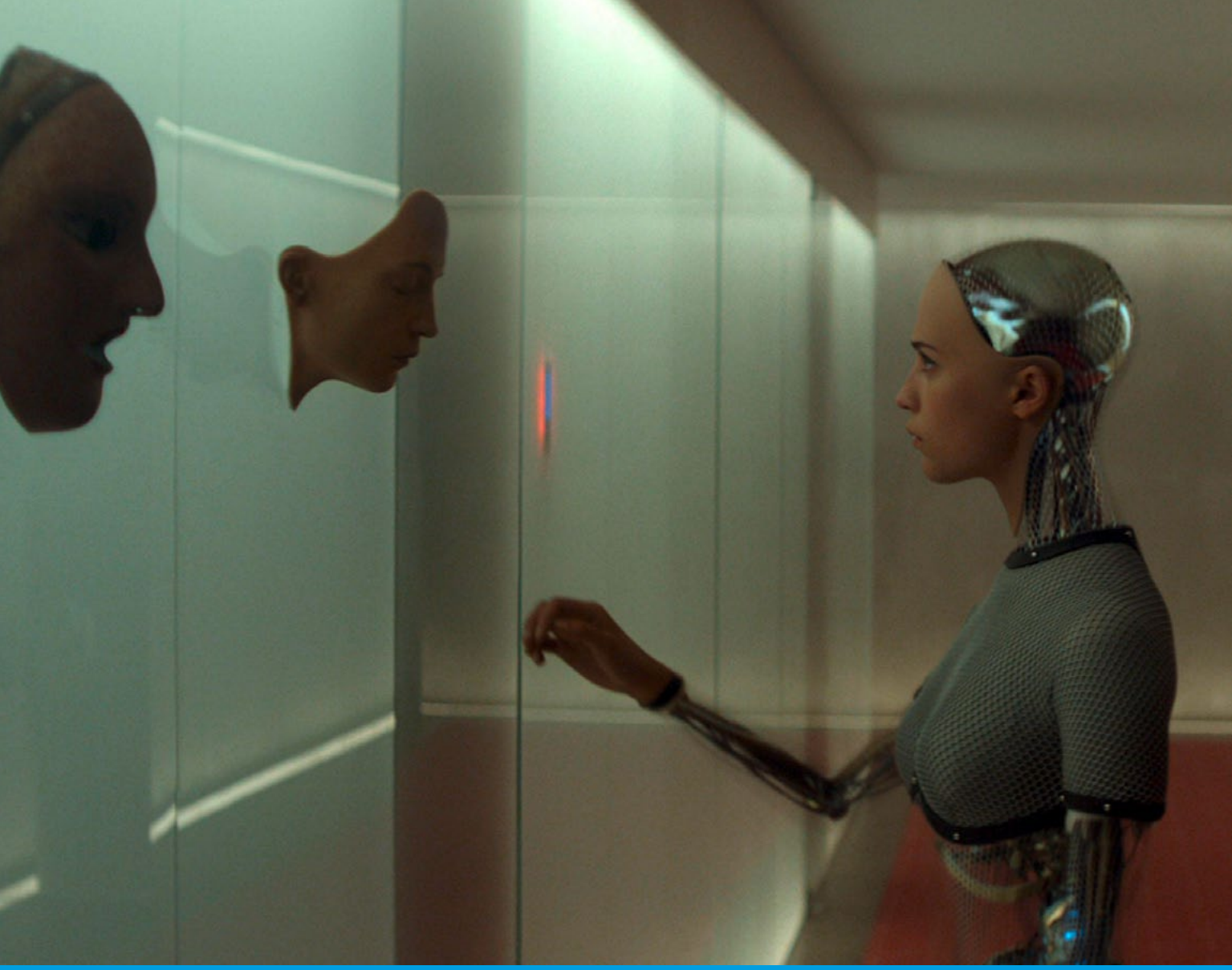
You can't control research. At present, no one is arguing that AI research be curtailed; merely that attention be paid to the issue of preventing negative consequences of poorly designed systems. But if necessary, we *can* control research: we do not genetically engineer humans because the molecular biology community decided, at a workshop at Asilomar in 1975, that it would be a bad idea, *even though "improving the human stock" had been a longstanding goal of many researchers in the biology community for several decades before that*.

Don't mention risks, it might be bad for funding. See nuclear power, tobacco, global warming.

In addition to these policy-level objections, there are also objections based on proposed simple solutions for avoiding negative consequences of superintelligent AI:

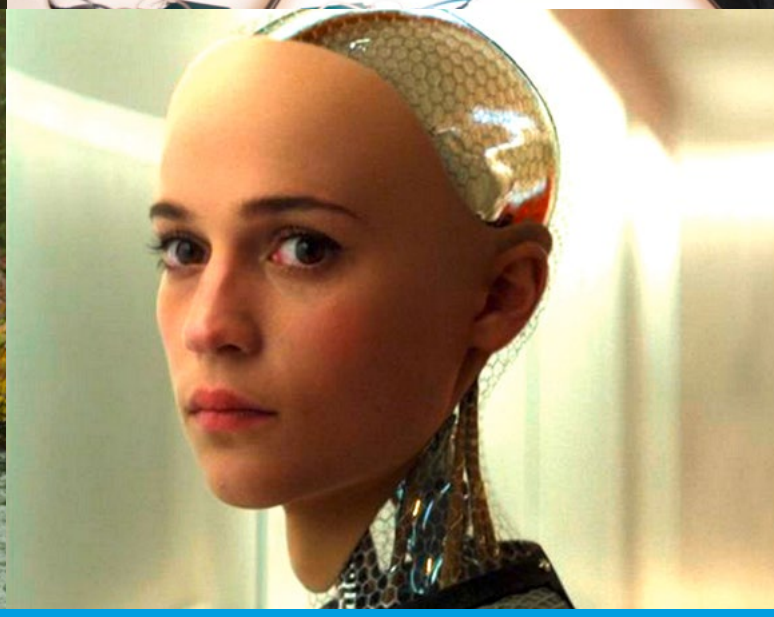
Instead of putting objectives into the AI system, just let it choose its own. It is far from clear how this solves the problem. The criteria by which an AI system would choose its own goals can be thought of as meta-objectives in their own right, and we face again the problem of ensuring that they lead to behaviors consistent with human well-being. We need to steer straight, not remove the steering wheel.

More intelligent humans tend to have better, more altruistic goals, so superintelligent machines will too. Beyond the fact that those making this argument think of themselves as more intelligent than average, there is precious little evidence for the premise of this argument; and the premise provides no support whatsoever for the conclusion.



“The challenge is not to act automatically.
It’s to find an action that is not automatic.
From painting, to breathing, to talking.
To falling in love...”

ALEX GARLAND
Writer and director Alex Garland was nominated for an Oscar for Best Original Screenplay for *Ex Machina* (2015).



Ex Machina (2015), Alex Garland.

Don't worry, we'll just have collaborative human-AI teams. Value misalignment precludes teamwork, so this solution simply begs the question of how to solve the core problem of value alignment.
Just don't put in "human" goals like self-preservation. See the discussion of Omohundro's argument above. For a coffee-fetching robot, death is not bad per se. Death is to be avoided, however, because it is hard to fetch the coffee if you are dead.

Don't worry, we can just switch it off. As if a superintelligent entity would never think of that.

SOLUTIONS

Bostrom (2014) considers a number of more serious technical proposals for solving the AI control problem. Some, under the heading of "oracle AI," seal the machines inside a kind of firewall, extracting useful question-answering work from them but never allowing them to affect the real world. (Of course, this means giving up on superintelligent robots!) Unfortunately, it seems unlikely to work—we have yet to invent a firewall that is secure against ordinary humans, let alone superintelligent machines. Others involve provably enforceable restrictions on behavior, but devising such restrictions is like trying to write loophole-free tax law (with superintelligent tax evaders!).

Can we, instead, tackle Wiener's warning head-on? Can we design AI systems whose purposes do not conflict with ours, so that we are sure to be happy with the way they behave? This is far from easy, but may be possible if we follow three core principles:

1. *The machine's purpose is to maximize the realization of human values*. In particular, it has no purpose of its own and no innate desire to protect itself.
2. *The machine is initially uncertain about what those human values are*. This turns out to be crucial, and in a way it sidesteps Wiener's problem. The machine may learn more about human values as it goes along, of course, but it may never achieve complete certainty.
3. *Machines can learn about human values by observing the choices that we humans make*.

It turns out that these three principles, once embodied in a formal mathematical framework that defines the problem the AI system is constitutionally required to solve, seem to allow some progress to be made on the AI control problem. In particular, at least in simple cases, we can define a template for agent designs that are provably beneficial under certain reasonable (if not strictly true) assumptions.

LEARNING REWARD FUNCTIONS

To explain the mathematical framework, it helps to be a little more precise about terminology. According to von Neumann and Morgenstern (1944), any rational agent can be described

as having a *utility function* $U(s)$ that assigns a real number representing the desirability of being in any particular world state s. Equivalently, this is the expected desirability of the possible future state sequences beginning with s, assuming that the agent acts optimally. (In operations research, this is often called the *value function*, a term that has a distinct meaning in economics.) A further assumption of stationary preferences is typically made (Koopmans, 1972), whose consequence is that the desirability of any state sequence can be expressed as a sum (possibly discounted over time) of immediate *rewards* associated with each state in the sequence. For convenience, the *reward function* $R(s,a,s')$ is defined to be the immediate reward associated with the transition from state s to state s' via action a. Typically, the reward function provides a concise way to define a task; for example, the task of playing backgammon can be defined by specifying the reward to be zero for all non-terminal states s' and a number between –192 and +192 for transitions to terminal states (the precise value depending on the state of the doubling cube and whether the games ends normally, with a gammon, or with a backgammon). The *utility* of a backgammon state s, on the other hand, will in most cases be a very complex function of s because it represents an expectation over future reward sequences with respect to all possible dice rolls occurring in the remainder of the game. For a person out enjoying his or her garden, rewards might be positive for smelling a rose (although not for smelling it 100 times in a row) and negative for pricking one's finger on the thorns, whereas the utility of being in the garden at that moment depends on all future rewards—and these might vary enormously depending on whether one is about to get married, about to begin a long prison sentence, and so on.

To the extent that *objectives* can be *defined* concisely by specifying reward functions, *behavior* can be *explained* concisely by inferring reward functions. This is the key idea underlying *inverse reinforcement learning*, or IRL (Russell, 1998; Ng and Russell, 2000). An IRL algorithm learns a reward function by observing the behavior of some other agent who is assumed to be acting in accordance with such a function. (IRL is the sequential form of preference elicitation, and is related to structural estimation of MDPs in economics.) Watching its owner make coffee in the morning, the domestic robot learns something about the desirability of coffee in some circumstances, while a robot with an English owner learns something about the desirability of tea in all circumstances.

SOLVING SIMPLE AI CONTROL PROBLEMS

One might imagine that IRL provides a simple solution to the value alignment problem: the robot observes human behavior, learns the human reward function, and behaves according to that function. This simple idea has two flaws. The first flaw is obvious: human behavior (especially in the morning) often conveys a desire for coffee, and the robot can learn this, but we do not want the robot to want coffee! This flaw is easily fixed: we need to formulate the value alignment problem so that the robot always has the fixed objective of optimizing reward for the human (the first principle given above), and becomes better able to do so as it learns what the human reward function is.

The second flaw is less obvious, and less easy to fix. The human has an interest in ensuring that value alignment occurs as quickly and accurately as possible, so that

the robot can be maximally useful and avoid potentially disastrous mistakes. Yet acting optimally in coffee acquisition while leaving the robot as a passive observer may not be the best way to achieve value alignment. Instead, the human should perhaps explain the steps in coffee preparation and show the robot where the backup coffee supplies are kept and what to do if the coffee pot is left on the heating plate too long, while the robot might ask what the button with the puffy steam symbol is for and try its hand at coffee making with guidance from the human, even if the first results are undrinkable. None of these things fit in with the standard IRL framework.

By extending IRL to incorporate both human and robot as agents, it becomes possible to formulate and solve a value alignment problem as a cooperative and interactive reward maximization process (Hadfield-Menell et al., 2017a). More precisely, a *cooperative inverse reinforcement learning* (CIRL) problem is a two-player game of partial information, in which the human knows the reward function[2] while the robot does not; but the robot's payoff is exactly the human's actual reward. (Thus, CIRL instantiates all three principles given above.) Optimal solutions to this game maximize human reward and may naturally generate active instruction by the human and active learning by the robot.

Within the CIRL framework, one can formulate and solve the off-switch problem—that is, the problem of preventing a robot from disabling its own off-switch. (With this, Turing may rest easier.) A robot that is designed to solve the CIRL problem is sure it wants to maximize human values, but also sure it does not know exactly what those are. Now, the robot actually *benefits* from being switched off, because it understands that the human will press the off-switch to prevent the robot from doing something counter to human values. Thus, the robot has a positive incentive to preserve the off-switch, and this incentive derives directly from its uncertainty about human values. Furthermore, it is possible to show that in some cases the robot is *provably beneficial*, that is, the expected reward for the human is higher when the CIRL-solving robot is available, regardless of what the human's actual reward function is (Hadfield-Menell et al., 2017b).

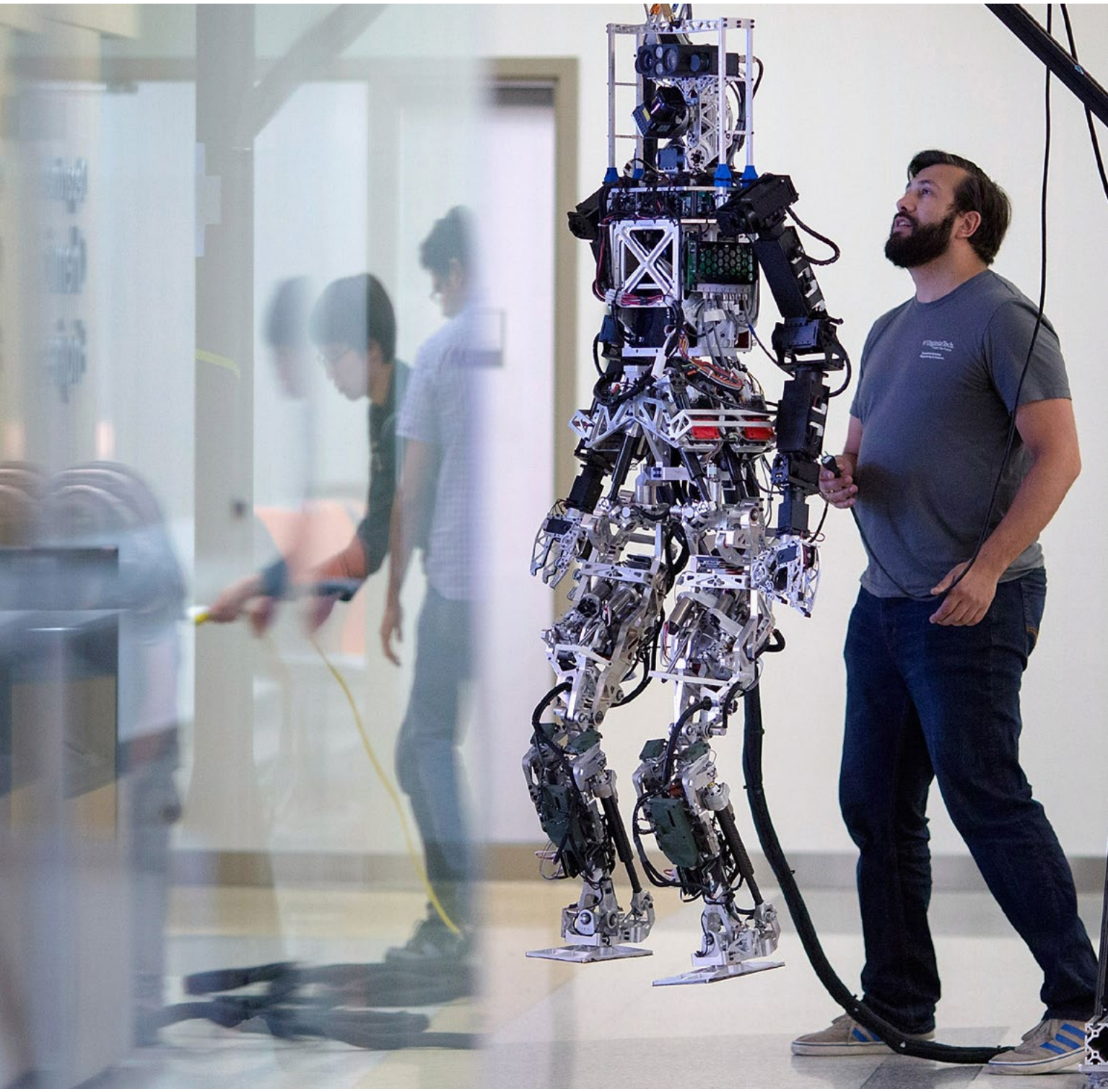
The off-switch example suggests some templates for controllable agent designs and provides at least one case of a provably beneficial system. The overall approach has some resemblance to mechanism design problems in economics,

where one attempts to incentivize other agents to behave in ways that will be provably beneficial to the mechanism designer. The key difference here is that we are *building* one of the agents in order to benefit the other.

The off-switch example works because of the second principle: that the robot should be uncertain about the true human reward function. Strangely, uncertainty about rewards has been almost completely neglected in AI, even though uncertainty about domain knowledge and sensor interpretation has been a core preoccupation for over twenty years.

One reason may be that uncertainty about the reward function is *irrelevant* in standard sequential decision problems (MDPs, POMDPs, optimal control problems) because the optimal policy under an uncertain reward function is identical to the optimal policy under a definite reward function equal to the expected value of the uncertain reward function. This equivalence only holds, however, when the environment provides no further information about the true reward function—which is not the case in CIRL problems, where human actions reveal information about human preferences. When the environment can supply additional information about the reward function, agents with reward uncertainty can exhibit behaviors that are not realizable by traditional AI systems with fixed reward functions.

The reader familiar with the concept of reinforcement learning (RL) might point out that the "reward signal" received by the RL agent after each state-action-state transition *does* provide information about the true reward function, because it gives the actual value of $R(s,a,s')$ for the observed transition. So, could ordinary RL be a basis for value alignment if the human simply supplies a reward signal directly to the robot? Unfortunately not! First, the human may not be able to quantify the true reward accurately, even for specific experienced transitions. Second, the formal model of RL assumes that the reward signal reaches the agent from *outside* the environment; but the human and robot are part of the same environment, and the robot can maximize its reward by modifying the human to provide a maximal reward signal at all times. The undesirability of this outcome, known as *wireheading* (Muehlhauser and Hibbard, 2014), indicates a fundamental flaw in the standard formulation of RL. The flaw is that the environment cannot supply *actual reward* to an agent; it can only supply *information* about the reward. Thus, a human giving a "reward signal" to the robot is not giving a reward, but is providing evidence (possibly noisy) about the human's preferences in the form of an action that selects a number. This new formulation clearly avoids the wireheading problem, because the robot can only be *worse off* if it modifies the information source to mask the underlying signal. And if the formulation stipulates that the robot has to maximize the human's *original* reward function, then modifying the human so that he or she has a new reward function that is easier to maximize does not do the robot any good.



A member of Team VALOR tests the Tactical Hazardous Operations Robot (THOR) while preparing for the DARPA Robotics Challenge in the Terrestrial Robotics Engineering and Controls lab (TREC) at Virginia Tech, USA.

PRACTICAL CONSIDERATIONS

I have argued that the framework of cooperative inverse reinforcement learning may provide initial steps toward a theoretical solution of the AI control problem. There are also some reasons for believing that the approach may be workable in practice. First, there are vast amounts of written and filmed information about humans doing things (and other humans reacting). Technology to build models of human values from this storehouse will be available long before superintelligent AI systems are created. Second, there are very strong, near-term economic incentives for robots to understand human values: if one poorly designed domestic robot cooks the cat for dinner, not realizing that its sentimental value outweighs its nutritional value, the domestic robot industry will be out of business. In the area of personal digital assistants, which seems likely to become a significant market before the end of the decade, there are obvious benefits to an assistant that quickly adapts to the complex and nuanced preferences of its owner.

There are obvious difficulties, however, with an approach based on learning values from human behavior. Humans are irrational, inconsistent, weak-willed, and computationally limited, so their actions do not always reflect their values. (Consider, for example, two humans playing chess: usually, one of them loses, but not on purpose!) Humans are also diverse in both values and circumstances, which means that robots must be sensitive to individual preferences and must mediate among conflicting preferences—a problem for social scientists as well as engineers. And some humans are evil, so the robot must have a way to filter out individual value systems that are incompatible with the general welfare.

It seems likely that robots can learn from non-rational human behavior only with the aid of much better cognitive models of humans. What of evil behavior? Is it possible to avoid corrupting our robots without imposing preemptive (and hence culturally relative) constraints on the values we are prepared to allow? It might be possible to use a version of Kant's categorical imperative: a reward function that ascribes negligible or negative value to the well-being of others lacks *self-consistency*, in the sense that if everyone operated with such a reward function then no one would obtain much reward.

SUMMARY

I have argued, following numerous other authors, that finding a solution to the AI control problem is an important task; in Bostrom's sonorous words, "the essential task of our age." I have also argued that, up to now, AI has focused on systems that are better at making decisions; but this is not the same as making better decisions. No matter how excellently an algorithm maximizes, and no matter how accurate its model of the world, a machine's decisions may be ineffably stupid, in the eyes of an ordinary human, if its utility function is not well aligned with human values.

This problem requires a change in the definition of AI itself, from a field concerned with pure intelligence, independent of the objective, to a field concerned with systems that are provably beneficial *for humans*. (I suppose we could also supply AI systems designed for other species, but that is probably not an immediate concern.) Taking the problem seriously seems to have yielded new ways of thinking about AI, its purpose, and our relationship to it.

NOTES

1. There are other possible risks from the misuse of increasingly powerful AI, including automated surveillance and persuasion, autonomous weapons, and economic disruption; these deserve serious study, but are not the subject of the current paper.
2. One might ask why a human who knows a reward function does not simply program it into the robot; here, we use "know" in the restricted sense of acting *as if* one knows the reward function, without necessarily being able to make it explicit. In the same sense, humans "know" the difference between the sounds for "cat" and "cut" without being able to write down the acoustic discrimination rule.

SELECT BIBLIOGRAPHY

— Bostrom, Nick. 2014. *Superintelligence*. Oxford: OUP.
— Etzioni, Oren. 2016. "Are the experts worried about the existential risk of artificial intelligence?" *MIT Technology Review*.
— Good, I. J. 1965. "Speculations concerning the first ultraintelligent machine." In *Advances in Computers 6*, Franz L. Alt, and Morris Rubinoff (eds.). Cambridge, MA: Academic Press.
— Hadfield-Menell, D., A. Dragan, P. Abbeel, and S. Russell. 2017a. "Cooperative inverse reinforcement learning." In *Advances in Neural Information Processing Systems 25*. Cambridge, MA: MIT Press.
— Hadfield-Menell, D., A. Dragan, P. Abbeel, and S. Russell. 2017b. "The off-switch." Submitted to AAAI-17.
— Koopmans, T. C. 1972. "Representation of preference orderings over time." In *Decision and Organization*, C. B. McGuire, and R. Radner, (eds.). Amsterdam: Elsevier/ North-Holland.
— Minsky, Marvin. 1984. "Afterword to Vernor Vinge's novel, 'True Names.'" Unpublished manuscript.
— Muehlhauser, L., and B. Hibbard. 2014. "Exploratory engineering in artificial intelligence." *Communications of the ACM* 57(9), 32–34.
— Müller, Vincent C., and Nick Bostrom. 2016. "Future progress in artificial intelligence: A survey of expert opinion." In *Fundamental Issues of Artificial Intelligence*, Vincent C. Müller (ed.), Synthèse Library 377. Berlin: Springer.
— Ng, Andrew Y., and Stuart Russell. 2000. "Algorithms for inverse reinforcement learning." In *Proceedings of the Seventeenth International Conference on Machine Learning*. Stanford, CA: Morgan Kaufmann.
— Omohundro, Stephen M. 2008. "The basic AI drives." In *Proceedings of the First AGI Conference*. Amsterdam: IOS Press.
— Russell, Stuart. 1998. "Learning agents for uncertain environments (extended abstract)." In *Proceedings COLT-98*. Madison, WI: ACM Press.
— Stone, Peter, et al. 2016. "Artificial intelligence and life in 2030." Stanford One Hundred Year Study on Artificial Intelligence: Report of the 2015 Study Panel.
— Turing, Alan M. 1951. "Can digital machines think?" Lecture broadcast on radio on BBC Third Programme; typescript at turingarchive.org.
— Von Neumann, J. and O. Morgenstern. 1944. *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.
— Wiener, Norbert. 1960. "Some moral and technical consequences of automation." *Science* 131 (3410): 1355-1358.

OpenMind

OPENMIND CHANNEL

READ THE FULL BOOK

The Next Step: exponential life

[+]

RELATED ARTICLES

Who Owns Big Data?

[+]

Big Data and the Future of Business

[+]

Human Legacies When Robots Rule the Earth

[+]

OTHER BOOKS