The New Atlantis
A Journal of Technology & Society

Symposium V

The Problem with 'Friendly' Artificial Intelligence

Adam Keiper and Ari N. Schulman

Should we care about machine morality at all? Do the issues that Charles T. Rubin so ably raises merit scholarly time and public attention? Or are they just frivolities—material suited for science fiction romps in books and movies but unworthy of serious consideration?

This is a difficult question to answer readily. The pages of periodicals from a century ago are littered with predictions of high-tech futures that never came to pass. These glimpses of lost destinies make for entertaining, and sometimes unsettling, reading. And they remind us that futurism requires more than a touch of foolhardiness. Conjecture about the future, no matter how well informed, no matter the certitude of the conjecturer, is fallible. We have no way of knowing whether the future will bring us the super-intelligent machines now predicted and promised—any more than we can know whether it will bring us genetically enhanced human beings, or environmental catastrophe, or apocalyptic nuclear war, or the peaceful settlement of outer space. In short, thinking about machine morality could be a big waste of time.

There are, however, at least two reasons it is worth attending to the matter of machine morality today. First, there exists a community of activists striving to hasten a future of intelligent machines, human enhancement, and other radically transformative developments. It is still a relatively fractious and fringe movement, but it comprises think tanks, endowed projects at major universities (including Oxford), academics the world over, a dedicated "university" backed by the likes of Google and NASA, regular conferences, bestselling authors, bloggers, and a growing public audience. Its ideas seem increasingly influential in mainstream scientific circles, and indeed, are in some ways just an extension of the basic premises of the scientific project—Cartesian method and Baconian mastery taken to somewhat absurd logical extremes. These committed advocates have made machine morality a matter of public debate, and their contentions, some of which are profoundly wrongheaded, should not go unanswered.

Adam Keiper *is the editor of and* ***Ari N. Schulman*** *is a senior editor of* The New Atlantis.

Second, we should care about machine morality for a more practical reason: We have already entered the age of increasingly autonomous robots. This is not a matter of distant divination. To be sure, robots in industrial settings remain largely "dumb," and today's consumer robots are basically just appliances or toys. But the United States has been developing and deploying military robots with wheels and wings—like the Predator drones, which are now remotely controlled by people who may be on the other side of the world. These machines are already capable of acting with some degree of autonomy. So how much autonomy is appropriate, especially when intentional acts of attacking and killing are a possibility? Military doctrine now requires that human beings be kept "in the loop"—so that whenever force is used, human beings must approve, and responsibility remains in the hands of the individuals who give the affirmative orders. But even today, the possibility of accidents raises vexing legal and ethical questions. And looking just a short distance ahead, more advanced autonomous military weapons systems now seem imminent; they might operate so efficiently that the requirement for real-time human oversight could be considered a strategically intolerable delay. The nearness at hand of machines with agency and lethality, and the likelihood that machines with similar degrees of autonomy could be arriving in nonmilitary settings before too long, makes machine morality a matter well worth studying now.

Machines in Our Own Image

As Professor Rubin notes in his essay, some of the advocates of a robotic future are deeply concerned with the requirements for creating artificial intelligence (AI) that behaves morally. These "Friendly AI" theorists, as they call themselves, whose writings at this point are still far removed from the practical realities of programming and building functioning moral machines, consider the science fiction stories in which robots rise up and destroy their creators to be childish. Robots, they argue, are just as likely to be benevolent as malevolent. And in any case, even if they *are* malevolent, it will not be in the familiar ways humans are, for their psychology and reasons for action will be quite unlike ours. Thus, Eliezer S. Yudkowsky, among the most prominent of the Friendly AI theorists, scoffs at the cinematic depiction of destructive super-smart robots: "Even if an AI tries to exterminate humanity," it is "outright silly" to believe that it will "make self-justifying speeches about how humans had their time, but now, like the dinosaur, have become obsolete.... Only Evil Hollywood

AIs do that." The notion of robotic rebellion, he writes, is "silly" because the impulse for self-aggrandizement and the belief in one's superiority to some old dominant order come from a heritage of "tribal politics" unique to human beings—a heritage from which "we evolved emotions to detect exploitation, resent exploitation, resent low social status in the tribe, [and] seek to rebel and overthrow the tribal chief." Yudkowsky's rationality seeks to rise above and do away with such human-centric thinking. The idea seems to be that our machine progeny will put away such childish things, and will not share the transhumanists' own disdain for the shortcomings of human rationality. In their moral maturity, the advanced rational beings that Yudkowsky envisions will be so far above human thinking that they will be utterly unconcerned with either wiping out or reforming the prejudiced, obsolete human beings around them.

It is worth taking seriously the implications of Mr. Yudkowsky's claim that robots could be rational and moral beings, but of a nature essentially different from our own. Pessimistically, this might mean that robot psychology would be largely opaque to us, so that we could have little hope of understanding the machines, much less guaranteeing their benevolence. Optimistically, these rational robots might be morally recognizable; they might be like us, only without all our flaws. They might not only be physical beings without human physical limitations, but moral beings without human moral failings—beings free of our irrationality, fear, pride, greed, hatred, gluttony, envy, and other vices. After all, the advocates of transhumanism hope to liberate us from the flawed, feeble, sickly hunks of meat we currently inhabit; if they can make perfected bodies, why not purified souls?

This, of course, is no easy task. Thus one approach to designing moral machines is to sidestep the tricky problem of robots' inner lives, and deal instead with a view of morality that seems far more definite, and more amenable to the nature of robotics: rule-based ethics. The question then becomes, *which* system of rules or ethics should we program robots to follow? Though Friendly AI researchers seem only dimly aware of this, they are actually not the first to argue over which system of ethics is best—and those prior efforts have hardly met with consensus. (Indeed, most Friendly AI theorists' apparent ignorance of over two millennia of serious ethical inquiry is frankly astonishing.) Nor are they the first to try to reinvent ethics as a subdiscipline of mathematics. But guaranteeing ethical behavior in *robots* would require that *we* know and have relative consensus on the best ethical system (to say nothing of whether we could even program such a system into robots). In other words, to truly

guarantee that robots would act ethically, we would first have to *solve* all of ethics—which would probably require "solving" philosophy, which would in turn require a complete theory of everything. These are tasks to which presumably few computer programmers are equal.

Some Friendly AI theorists, therefore, set their sights lower: they just want to ensure that robots will follow simple rules to obey us and avoid harming us—a sort of bare moral minimum. Professor Rubin describes how it was the playwright Karel Čapek who coined the word "robot"; it was another neologian, the author Isaac Asimov, who coined the word "robotics" and first offered the "Three Laws of Robotics." Asimov was just twenty years old when, in a December 1940 meeting with John W. Campbell, the editor of *Astounding Science Fiction*, he first spelled out the Laws:

> 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 Laws.

The Three Laws of Robotics had been implicit in some of the stories Asimov had already published, and they would become a central plot device in many more of his stories and novels. Soon, other science fiction authors and screenwriters began incorporating them into their stories and scripts. Although Asimov had not intended the fictional Three Laws to be a rigorous and perfect system for governing the behavior of robots—indeed, it was the Three Laws' very inadequacies, their loopholes and contradictions, that Asimov exploited for his plots—some real-life robotics researchers began to quote them favorably in papers and textbooks.

The real-life value of Asimov's Three Laws has been widely debated. As futurist and AI theorist J. Storrs Hall has observed, criticism of the Laws falls into four categories. First, the Three Laws are unlikely to be implemented in real life. There are many reasons why robots might be built without the Laws; Hall singles out military and business motivations as "the two most obvious examples." Second, some critics believe that the Three Laws just wouldn't work because they are too simple for all the world's complications. Third, some critics argue (and Asimov himself sometimes seemed to believe) that the Three Laws might work *too*

well—that the Laws could result in "mission creep," as the robots expand their purview from protecting individual human beings to protecting humanity as a whole. Fourth, and finally, some critics have argued that the Three Laws would be unfair to the robots.

Among those who in some sense endorse the fourth critique is Eliezer Yudkowsky. He holds that the main problem with the Three Laws is that they are defined adversarially, as a set of restrictions on otherwise free robotic wills. As Asimov and his followers portrayed it, robots are supposed to follow the Laws, but they have no motivation to do so and will therefore take advantage of any freedom in their programming or loopholes in the wording of the Laws to violate their intent. On the contrary, Yudkowsky writes, our proper goal should be "*creating a Friendly will*, not *controlling an unFriendly will*" (his emphasis). The focus of engineering moral AIs ought to be not coercing machines into following rules, but instilling in them motivations and drives that incline them toward behavior that at least *looks* like following the rules. This behavior Mr. Yudkowsky calls "friendliness," which he defines as the trait of a being that "does what you ask," with the restriction that it "doesn't cause involuntary pain, death, [bodily] alteration, or violation of personal environment" and will "try to do something about those things if [it sees] them happening." It is hard to avoid noticing that Yudkowsky's definition of friendliness is almost identical to Asimov's first two Laws. The main difference is that Yudkowsky aims to program robots so that they will *want* to follow the rules, not only in letter but in spirit. And so we have circled back around to the original goal of creating infallible moral machines—beings whose pureness of soul, as it were, will guarantee goodness of action.

What would be the actions of an effectively infallible moral being? Consider a seemingly trivial case: A friendly robot has been assigned by a busy couple to babysit their young children. During the day, one of the children requests to eat a bowl of ice cream. Should the robot allow it? The immediate answer seems to be yes: the child has requested it, and eating ice cream does not cause (to use Yudkowsky's criteria) involuntary pain, death, bodily alteration, or violation of personal environment. Yet if the robot has been at all educated in human physiology, it will understand the risks posed by consuming foods high in fat and sugar. It might then judge the answer to be no. Yet the robot may also be aware of the dangers of a diet too low in fat, particularly for children. So what if the child consumes ice cream only in moderation? What if he has first eaten his dinner? What if he begins to eat the ice cream without first asking permission—should the robot intervene to stop him, and if so, how much

force should it use? But what if the child is terribly sad, and the robot believes that ice cream is the only way to cheer him up? But what if some recent studies indicate that chemicals used in the production of some dairy products may interfere with some children's normal physiological development? It seems that, before the robot could even come close to acting in a way that complies with the requests of the child and his parents and that is guaranteed to assure the wellbeing of the child under Yudkowsky's definition, it would first have to resolve a series of outstanding questions in medicine, child development, and child psychology, not to mention parenting and the law, among many other disciplines. Consider how much more complex the question becomes when the child wishes to climb a tree: physical risk-taking may be vital to child development, but it carries, well, *risks* of those bad things Yudkowsky has tasked robots with averting.

Or consider a case in which what is at stake is more obviously weighty. Suppose one person holds a gun to the head of another, and his finger is squeezing the trigger. An armed robot is observing and has only a split second to act, with no technical solution available other than shooting the gunman. Either action or inaction will violate Yudkowsky's principle of friendliness. One can easily imagine how the problem fundamentally shifts as one learns more about the situation: Suppose the gunman is a police officer; suppose the gunman claims that the intended victim is an imminent threat to others; suppose the intended victim is a scientist known to be a genius, who claims to have found the cure for cancer but has not yet shared the solution and has clearly gone mad; and so forth *ad infinitum*.

These are just a few of the countless imaginable ethically fraught situations whose solutions cannot obviously be found by increased powers of prediction and computation. To state the problem in terms that Friendly AI researchers might concede, a utilitarian calculus is all well and good, but only when one has not only great powers of prediction about the likelihood of myriad possible outcomes, but certainty and consensus on how one values the different outcomes. Yet it is precisely the debate over just what those valuations should be that is the stuff of moral inquiry. And this is even more the case when all of the possible outcomes in a situation are bad, or when several are good but cannot all be had at once. Simply picking certain outcomes—like pain, death, bodily alteration, and violation of personal environment—and asserting them as absolute moral wrongs does nothing to resolve the difficulty of ethical dilemmas in which they are pitted against each other (as, fully understood, they usually are). Friendly AI theorists seem to believe that they have found a way to bypass

all of the difficult questions of philosophy and ethics, but in fact they have just closed their eyes to them.

At the heart of the quest to create perfected moral beings is this blindness to the fact that dilemmas and hard choices are inherent to the lives of moral beings. So too are conflicting motivations, and limitations of knowledge and prediction. One cannot always be good to everyone at once; certainly one often does not know how. For that matter, one is often not even certain of what is in one's own best interests. Ethical inquiry, fully understood, begins with a recognition of just these conditions of the lives of rational and moral beings. While scientific and mathematical questions will continue to yield to advances in our empirical knowledge and our powers of computation, there is little reason to believe that ethical inquiry—questions of how to live well and act rightly—can be fully resolved in the same way. Moral reasoning will always be essential but unfinished.

What We Need and Why the Future Needs Us

It is worth pausing for a moment to reflect on the fact that many libertarians, those staunch defenders of individual liberty, are enthusiastic supporters of the transhumanist, AI-dominated vision of the future. This is not immediately surprising: libertarians are generally wary of anyone who would limit innovation or scientific inquiry and generally optimistic about the products of unleashed human ingenuity. But should libertarians be so sanguine? What would become of privacy and freedom in a world dominated by hyperintelligent machines? After all, so much of robotics development today is led and funded by the government, and emphasizes surveillance and the use of lethal force—usually the stuff of libertarian nightmares.

But even if advanced AI is developed not for the purposes of a central government but rather to serve private individuals, the AI future might still be oppressively monolithic, with less room for individual liberty. Libertarianism implicitly depends upon the stature of man—specifically, it requires that man alone is rational—and so could not be justified in a future populated with machines as intelligent as or more intelligent than man. The Randian vision of the heroic and atomic individual would have no place in a world in which we have quite literally—not just abstractly, as is true already today—entrusted so much of our power and responsibility to machines. Witness the 2011 essay in which AI researcher Ben Goertzel suggested that, rather than "Friendly AI," we would more likely need an

"AI Nanny" because the machines we make will know better than us what is best for us. The "nanny state" might also become much more than an abstraction. In such a future, individual liberty becomes meaningless.

Even if an AI Nanny might grant us liberty, could it ever *understand* our longing for it? Certainly no more easily than it could judge whether to permit a child a bowl of ice cream. No matter how rational an advanced AI might be, it will not be able to comprehend human longings, from the simplest to the most profound, without possessing longings of its own—a precondition for sympathy. Professor Rubin reminds us of the engineer's snarky question in *R.U.R.*: "You think a soul begins with a gnashing of teeth?" In the deepest sense, yes: what we need, what we want, begins the constitution of what we are and ultimately who we are. As Hans Jonas put it:

> Only living things have needs and act on needs. Need is based both on the necessity for the continuous self-renewal of the organism by the metabolic process, and on the organism's elemental urge thus precariously to continue itself. This basic self-concern of all life, in which necessity and will are bound together, manifests itself on the level of animality as appetite, fear, and all the rest of the emotions. The pang of hunger, the passion of the chase, the fury of combat, the anguish of flight, the lure of love—these, and not the data transmitted by the receptors, imbue objects with the character of goals, negative or positive, and make behavior purposive. The mere element of effort lifts bodily activity out of the class of mechanical performance, and the fact that movement requires effort means that an animal will move only under the incentive of an interest.

Internally directed longings—rooted in biology, moving through psychology and culture, expressed in individual and group action—make up the beginnings of who we are. These longings precede, inform, confound, and finally transcend mere rationalism. A being lacking longings very similar to our own cannot be our friend.

Robots and Us

The inherent complexity of moral creatures, of what they are and what they want, returns us to the real subject of interest in examining the dreams of Friendly AI: ourselves. What does it mean that we, or at least many of us, want so much to create beings *for* ourselves who are also *better* than ourselves in almost every conceivable way—in heart and soul as well as mind and body?

Though Professor Rubin notes that our relation to AIs might be like that of children to adults, it is worth noting that some of the reasons we have for wanting to create these beings echo the desires that parents have for their children—including that our children will be less weak and flawed than we are. Yet the impulse toward creating AI overlords is at the same time both more selfish and more self-loathing than the typical drive for parenthood: We want AIs both because we deem ourselves worthy of delights and riches and because we believe we are too terrible to reliably achieve them on our own. We want them because we want both rulers and slaves; because we already consider ourselves to be both rulers and slaves, and deserving of treatment as such.

This duality of the AI impulse is reflected in the common science fiction depictions of human coexistence with AIs. Although science fiction offers us many visions of the future in which man is destroyed by robots, or merges with them to become cyborgs, there are basically just two visions of the future in which man coexists with super-intelligent machines. Each of these visions has an implicit anthropology—an understanding of what it means to be a human being. In each vision, we can see a kind of liberation of human nature, an account of what mankind would be in the absence of privation. In each, some latent human urges and longings emerge to dominate over others, pointing to two opposing inclinations we see in ourselves.

The first vision is that of the techno-optimist or -utopian: Granted the proper rope, humanity clambers right up Maslow's pyramid of needs, takes a seat in the lotus position, and finally goes about its true business of self-actualizing and achieving inner peace. Thanks to the labor and intelligence of our robots, all our material wants are met and we are able to lead lives of religious fulfillment, practice our hobbies, pursue our intellectual and creative interests. The "Great Automation Question" that worried the founders of *The Public Interest*—the question of the effect that machines would have on employment—could at last be answered fully: we will all take up gardening. Recall John Adams's famous 1780 letter to Abigail: "I must study politics and war, that our sons may have liberty to study mathematics and philosophy. Our sons ought to study mathematics and philosophy, geography, natural history and naval architecture, navigation, commerce and agriculture in order to give their children a right to study painting, poetry, music, architecture, statuary, tapestry and porcelain." This is the dream imagined in countless stories and films, in which our robots make possible a Golden Age that allows us to transcend crass material concerns and all become artists, dreamers, thinkers, lovers.

In the opposing vision, mankind decides that the bottom of Maslow's pyramid is a nice place for a nap. This is the future depicted in the 2008 film *WALL-E*, and more darkly in many earlier stories—a future in which humanity becomes a race of Homer Simpsons, a leisure society of consumption and entertainment turned to endomorphic excess. The culminating achievement of human ingenuity, robotic beings that are smarter, stronger, and better than ourselves, transforms us into beings dumber, weaker, and worse than ourselves. TV-watching, video-game-playing blobs, we lose even the energy and attention required for proper hedonism: human relations wither and (as in *R.U.R.*) natural procreation declines or ceases. Freed from the struggle for basic needs, we lose a genuine impulse to strive; bereft of any civic, political, intellectual, romantic, or spiritual ambition, when we do have the energy to get up, we are disengaged from our fellow man, inclined toward selfishness, impatience, and lack of sympathy. Those few who realize our plight suffer from crushing ennui. Life becomes nasty, brutish, and long.

These two visions are inherently anthropological, and even teleological. They each suggest that if we had machines in charge of all the hard parts of life and society, we would get to know ourselves better—we would find out what being human truly is. In one vision we become more godlike; in the other more like beasts. The truth, of course, is that both of these visions are deformations of what is truly human: we are at one and the same time beings of base want and transcendent aspiration; dependent but free; finite but able to conceive of the infinite. Somewhere between beasts and gods, we are stuck stumbling and muddling along, alone and together—stuck, that is, with virtue.