CHAPTER SIX

Accelerating AI

Many different kinds of technologies, from nanotechnology to biotechnology, promise to dramatically change human life. But of all these potentially revolutionizing technologies, the most important for social governance is artificial intelligence (AI), because AI is an information technology. As a result, the development of machine intelligence can directly improve governance, because progress in AI can help in assessing policy consequences. More substantial machine intelligence can process data, generate hypotheses about the effects of past policy, and simulate the world to predict the effects of future policy. Thus, it is more important to formulate a correct policy toward AI than toward any other rapidly advancing technology, because that policy will help advance beneficial policies in all other areas.

The holy grail of AI is so-called strong AI, defined as a general purpose intelligence that approximates that of humans. Strong AI has the capacity to improve itself, leading rapidly to machines of greater-than-human intelligence. More concretely, it raises dramatic possibilities of very substantial benefits and dangers, from the prospect of a machine-enabled utopia to that of a machine-ruled despotism.

Fortunately, the correct policy for AI—substantial government support for Friendly AI—both promotes AI as an instrument of collective decision making and helps prevent the risk of machine takeover. Friendly AI can be defined broadly as the category of AI that will not ultimately prove dangerous to humans.[1] The benefits of supporting Friendly AI are twofold. First, the creation of Friendly AI is the best and probably only way of forestalling unfriendly AI, because Friendly AI can prove a crucial ally to humanity in its efforts to prevent the rise of dangerous machines. Second, government support is justified on the wholly independent grounds of improving social decision making. Even if strong AI is not realized for decades, progress in AI can aid in the gathering and analysis of data for evaluating the consequences of social policy, including policy toward other transformative technologies.

The Possible Coming of Strong AI

The idea of artificial intelligence powerful enough to intervene in human affairs has been the stuff of science fiction from HAL in *2001: A Space Odyssey* to the robots in *Wall-E*.[2] The notion of computers that rival and indeed surpass human intelligence might at first seem to be speculative fantasy rather than a topic that should become a salient item on the agenda of social analysis and policy.[3] But travel to the moon was itself once a staple of science fiction in the nineteenth and twentieth centuries.[4] Yet because of a single government program, exploration of the moon is now a historical event of more than forty years' standing. Technology has likely been accelerating since then, so events that seem even a greater stretch of imagination today may be the staple of the news of tomorrow. And unlike a lunar landing, even incremental advances in AI have direct implications for social governance.

Strong AI entails the creation of machines with the general human capacity for abstract thought and problem solving. It is generally conceded that if such machines were possible, they would soon surpass human cognitive abilities, because the same processes that gave rise to them could rapidly improve them. The machines themselves could aid in this process with their greater-than-human capacity to share information among themselves.[5]

The success of strong AI depends on the truth of three premises. The first is functionalism. Functionalism turns on the proposition that cognition is separate from the system in which cognition is realized.[6] Thus, abstract thinking can be equally realized in a biological system like the brain or in an electronic one like a computer. Under this hypothesis a system of symbols, when properly actualized by a physical process, regardless of whether that process is biologically based, is "capable of intelligent action."[7]

The philosopher John Searle is most prominent among scholars who challenge the notion that a machine manipulating abstract symbols can become the equivalent of a human mind. Searle provides the analogy of a Chinese room.[8] If someone is put in a room and asked questions in Chinese, he can be given written directions on how to manipulate Chinese characters so as to give answers to the questions in Chinese.[9] Yet because he himself understands nothing of Chinese, this manipulation of symbols is a poor simulacrum of human understanding.[10]

One powerful objection to Searle's analogy is that the entire system—the written directions plus the human manipulator—does understand Chinese.[11] Searle thus unfairly anthropomorphizes the subject of understanding. Confusing the proposition that AI may soon gain human capabilities with the proposition that AI may soon partake of human nature

is the single greatest systemic mistake in thinking about computational intelligence—an error that science fiction itself has encouraged and perpetuated. But regardless of whether Searle is wrong or right about the conceptual possibility of strong AI, greater machine intelligence could still help humans in managing information and even formulating hypotheses about the social world. The metamorphosis of AI into a conscious agent may bear on the capacity of humans to control it. But continued progress in AI can bring very substantial social benefits even if the consciousness barrier is never reached.

The second premise undergirding strong AI is that computers will have the hardware capacity to mimic human thought. As described in chapter 2, raw computer power has been growing exponentially according to Moore's law. Assuming the computational capacity of computers continues to grow as Moore's law predicts, the hardware capacity of a computer is likely to achieve equality with a human brain between 2025 and 2030.[12] Even if this pace does not continue, it seems hard to believe that such capacity will not be reached by the midpoint of this century.

The third premise is that programmers will be able to provide the software to convert the gains in hardware into advances in AI. I have already suggested that software is also making very substantial improvements over time. Nevertheless, capturing the fluidity of human intelligence presents the greatest challenge.[13] In fact, some have argued that despite the previous growth in computational capacity and progress in software, AI has been largely a failure with little to show for fifty years of work.[14] This assessment seems far too harsh. For over a decade we have seen how computers have been able to defeat the greatest chess players in the world.[15] In 1997 IBM built Deep Blue, the machine that defeated world chess champion Garry Kasparov. While Deep Blue was a supercomputer, programs run on personal computers can now beat the best players in the world. Thus, what was once the frontier of AI has become the stuff of everyday computing.

This kind of progress on complex games is continuing. Although the rules of the famous Chinese game Go are simple and few, the nature of the game and the size of the board require any AI using the same computational strategy as Deep Blue to make approximately 1 million to the 10th power assessments of the game as it progresses. If each calculation lasted one microsecond, Deep Blue, using 1997 AI to play Go, would still require more time per turn than the age of the universe.[16] Yet researchers predict that ten more years of exponential growth in computer hardware and progress in computer software will make it possible to program an AI capable of calculating Go positions as thoroughly as Deep Blue analyzed chess moves.[17] Thus, AI is continuing to benefit not only from increased computational power but also from better ideas on how to make use of that power.

Of course, it is true that chess and Go are completely formal systems, mimicking only a narrow sliver of human capability. But it is hardly surprising that AI progresses from success in more formal and predictable environments to success in more informal and fluid ones. Software progress in tandem with the growing hardware capability continues to extend the kind of intelligence that AI can present.[18] A recent example is IBM's progress from developing Deep Blue in 1997 to developing Watson in 2011. This supercomputer specializes in playing *Jeopardy!*—a game of general knowledge. By 2011 Watson succeeded in beating the best *Jeopardy!* champions of all time.[19] This victory required the computer to react to natural language and to disentangle humor, recognize puns, and resolve ambiguity. In short, winning at *Jeopardy!* required Watson to operate in a much less formal world, one more like the chaotic world that human intelligence confronts on a daily basis.

While Watson used multiple formal rules in the form of algorithms to navigate this world, it is precisely its use of formal systems to compete in the world of more fluid human intelligence that makes its achievement a greater breakthrough than that of Deep Blue. Spinoffs from Watson will have implications for tasks of more general relevance to society than games. Already IBM is moving to sell Watson-like programs to help in medical diagnosis. These developments suggest that progress in AI falling short of strong AI will have large social effects over time. Some advances, such as a medical diagnosis program, are likely to complement a job done by a human, boosting productivity. But under other circumstances these programs could also substitute for a growing number of routine white-collar jobs. In my own field of law, the routine drafting of wills and contracts may soon be largely done by computers.

Continuing exponential growth in computational power is on the cusp of changing other key elements of social life. Cars run by computers now can autonomously navigate city traffic, another capacity that requires the ability to adapt to more fluid situations.[20] In fact, Google is confident enough that such systems can become routine in the near future that it has successfully petitioned the state of Nevada to make it legal for such autonomously driving cars to be on the road.[21]

Google's own core business—its search capacity—is itself an astonishing example of creating an intellectual function that far surpasses human capacities, and its search function is becoming more powerful. In part, the increased scope for search is simply a function of the increase in mobile computing. Mobile search opens up new search categories, like navigation and finding nearby services. But search will deepen as well as broaden in scope. It is deepening by the increasing capacity of the Web to categorize data. Search will aggregate the opinions of others from social media, adding to the capacity of our personal decision making. In the

future the search function will use evidence of our past searches to help put queries into the context of our larger objectives. The mechanisms of search thus gain a greater underlying intelligence, morphing into agents that can actively help us organize our lives.

Other human-like abilities in which computers are improving rapidly include translating languages in real time and flying planes. A new program can write newspaper articles, and its architects even predict that it will garner prizes for journalism in five years.[22] The growing list of examples of specific intelligences generated by AI underscores its progress.

While these feats are impressive and exceedingly useful emulations of human behavior, they do not yet approximate strong AI, a general intelligence that spans the many functions and capabilities of human intelligence, to say nothing of a self-conscious intelligence. Recent trends in AI have turned away from direct attempts to mimic human intelligence, focusing instead on the comparative advantage of computers in turning their enormous calculating power to analyzing huge masses of data.[23] Through the use of multiple algorithms, such systems can accomplish feats that surpass human intelligence without in any way replicating its mechanisms.[24]

Some theorists, however, have suggested that this dependence on computers' brute strength in calculating capacity underscores the large gap between what has been accomplished and what needs to be achieved for strong AI. One critique is that AI must be embodied to become strong, and thus research into robotics is an important component of progress. Other researchers suggest that emotion will become a necessary component of strong AI, particularly any AI that aspires to consciousness.

Because of the focus on such obviously human qualities as embodiment and emotion, as opposed to simply computational capacity, some AI researchers are looking to mimic aspects of the human brain. They will be aided in this project by mapping the human brain, permitting a kind of reverse engineering. In 2007 IBM simulated half of a mouse brain.[25] While we perceive the gap between mouse and humans as very wide, in evolutionary terms these animals are relatively close and share a computational architecture. Moreover, in a world of technological acceleration, it is a fundamental mistake to view progress as linear.[26] As a result of exponential computational power, far more progress can be expected on brain simulations in the next ten years than in the previous ten and even more in the ten years after that. This progress is likely to open up new lines of AI research.

But the point here is not to prove that AI will succeed in replicating and then in surpassing human intelligence; it is just to suggest that such a prospect is plausible. In any event, greater progress in useful artificial intelligence can be expected. Even if AI does not actually exceed human

intelligence, it may still offer useful insights that advance our collective decision making.

The Threats of AI

AI has become a subject of major media interest. In May 2010 the *New York Times* devoted an article to the prospect of the day when AI will equal and then surpass human intelligence.[27] The article speculated on the dangers that such developments in strong AI might create.[28] Then in July 2010 the *Times* discussed computer-driven warfare. Various experts expressed concern about the growing power of computers, particularly as they become the basis for new weapons like the Predator, a robotic drone that the United States now uses to kill terrorists.[29] These articles highlight the twin fears about growing machine intelligence: the existential dread of machines that become uncontrollable by humans and the political anxiety about machines' destructive power on a revolutionized battlefield.

Bill Joy, former chief technologist for Sun Microsystems, is most eloquent on the first concern. He does not disagree with Ray Kurzweil that we are entering an age of unprecedented technological acceleration in which AI will become vastly more powerful than it is today.[30] But he does not share the view that this development will lead to technological utopia. In his article "Why the Future Doesn't Need Us," he raises the alarm that human beings cannot ultimately control these machines.[31] The power of his critique lies precisely in his acknowledgment of the wealth of potential benefits from strong AI. But for Joy, however great the benefits of AI might be, the risk of losing control of the intelligence created is still greater. In his view, man resembles the sorcerer's apprentice—too weak and too ignorant to master the master machines. This concern represents the culmination of a particular kind of fear that goes back to the Romantic era and was first represented by Dr. Frankenstein's monster, who symbolized the idea that "all scientific progress is really a disguised form of destruction."[32]

Fears of artificial intelligence on the battlefield may be an even more immediate concern. Nations have always attempted to use technological innovation to gain advantages in warfare.[33] Computational advancement today is essential to national defense. The Defense Advanced Research Projects Agency spends billions of dollars developing more advanced military mechanisms that depend on increasingly more substantial computational capacity.[34]

It is hard to overstate the extent to which advances in robotics, itself driven by AI, are transforming the U.S. military. During the Afghanistan

and Iraq wars, more and more Unmanned Aerial Vehicles (UAVs) of different kinds have been used. Ten unmanned Predators were in use in 2001, and at the end of 2007 there were 180.[35] Unmanned aircraft, which depend on substantial computational capacity, form an important part of our military and may prove to be the majority of aircraft by 2020.[36] Under President Obama, the campaign against al-Qaeda is being conducted principally by drone strikers. It has been said that the F-35 now being built is the American military's last manned fighter plane and that the last American fighter pilot has already been born.[37]

Even below the earth, robots perform important tasks such as mine removal.[38] And already in development are robots that would wield lasers as a kind of special infantryman focused on killing snipers.[39] Others will act as paramedics.[40] It is not an exaggeration to predict that war twenty or twenty-five years from now may be fought predominantly by robots. The AI-driven battlefield gives rise to a set of fears that is different from those raised by the potential autonomy of AI. Here the concern is that human malevolence will lead to these increasingly capable machines wreaking more and more havoc and destruction.

The Futility of the Relinquishing AI or Prohibiting Battlefield Robots

Joy argues for "relinquishment"—that is, the abandonment of technologies that can lead to strong AI. Those who are concerned about the use of AI technology on the battlefield focus more specifically on prohibiting or regulating weapons powered by AI. But whether the objective is relinquishment or the constraint of new weaponry, any such program must be translated into a specific set of legal prohibitions. These prohibitions are certain to be ineffective. Nations are unlikely to unilaterally relinquish either the technology behind accelerating computational power or the research to further accelerate that technology.

Indeed, were the United States to relinquish such technology, the whole world would be the loser. The United States is a flourishing commercial republic that can uniquely supply the public goods of global peace and security. Because it gains a greater share of the peace dividend than other nations, it has incentives to shoulder the burdens to maintain a global peace that benefits not only the United States but the rest of the world as well.[41] By relinquishing the power of AI, the United States would encourage rogue nations to develop it.

Thus, the only realistic alternative to unilateral relinquishment would be a global agreement for relinquishing or regulating AI-driven weaponry. But such an agreement would face the same obstacles as nuclear

disarmament. As recent events with Iran and North Korea demonstrate,[42] it seems difficult, if not impossible, for such pariah nations to relinquish nuclear arms, because these weapons are a source for their geopolitical strength and prestige. Moreover, verifying any prohibition on their preparation and production is a task beyond the capability of international institutions.

The verification problems are far greater with respect to the technologies relating to artificial intelligence. Relatively few technologies are involved in building a nuclear bomb, but there are many routes to progress in AI. Moreover, building a nuclear bomb requires substantial infrastructure,[43] whereas AI research can be done in a garage. Constructing a nuclear bomb requires very substantial resources beyond the capacity of most groups other than nation-states.[44] AI research is done by institutions that are no richer than colleges and perhaps would require even less substantial resources.

Bill Joy recognizes the difficulties of relinquishment but offers no plausible means of achieving it. He suggests that computer scientists and engineers take a kind of Hippocratic oath that they will not engage in research with the potential to lead to AI that can displace the human race.[45] But many scientists would likely refuse to take the oath, because they would not agree with Joy's projections. Assuming some took the oath, many governments would not likely permit their scientists to respect it because of the importance of computational advances to national defense. Even without prompting from the government, many researchers would likely disregard the oath because of the substantial payoffs for advances in this area from private industry. All would have difficulty complying with such a directive because they could not easily predict what discoveries will propel AI forward in the long run.

For these reasons, prohibiting or substantially regulating research into AI is a nonstarter. Indeed, the relative ease of performing artificial intelligence research suggests that, at least at current levels of technology, it would be difficult for a nation to enforce a prohibition on research directed wholly against its own residents. Even a domestic prohibition would run up against the substantial incentives to pursue AI research, because the resulting inventions can provide lucrative applications across a wider range of areas than can research into nuclear weapons.

Exaggerated Fears of AI

The threats from strong AI—both the fear that it represents an existential threat to humanity and the fear that it will lead to greater loss of life in war—have been exaggerated because they rest on conceptual and empiri-

cal confusions. The existential fear is based on the mistaken notion that strong artificial intelligence will necessarily reflect human malevolence. The military fear rests on the mistaken notion that computer-driven weaponry will necessarily worsen, rather than temper, human malevolence.

The Existential Threat

The existential threat can be dissolved if there is a substantial possibility of constructing Friendly AI.[46] As defined earlier, Friendly AI is artificial intelligence that will not use its autonomy to become a danger to mankind. The argument for Friendly AI begins by rejecting the proposition that advanced artificial intelligence will necessarily have the kind of willpower that could drive it to replace humanity. The basic error in such thinking is the tendency to anthropomorphize AI.[47] Humans, like other animals, are genetically programmed in many instances to regard their welfare (and those of their relatives) as more important than the welfare of any other living thing.[48] But the reason for this motivation lies in the history of evolution: those animals that put their own welfare first were more likely to succeed in distributing their genes to subsequent generations.[49] Artificial intelligence will not be the direct product of biological evolution nor necessarily of any process resembling it. Thus, it is a mistake to think of AI as necessarily having the all-too-human qualities that seek to evade constraints and take power.

This is not to say that one cannot imagine strong AI capable of malevolence. One way to create AI, for instance, may be to replicate some aspects of an evolutionary process so that versions of AI progress by defeating other versions, a kind of tournament of creation. Such a process would be more likely to give rise to existential threats. Further, one cannot rule out that malevolence, or at a least a will to power, could be an emergent property of a particular line of AI research.

It is also true that humans may be able to enhance themselves by fusing themselves to AI devices to improve their intelligence and other capabilities. Such cyborgs might well have the will to power. But it is likely easier to regulate the acquisition of dangerous capabilities embedded in humans than to regulate research into AI in general. Monitoring AI research in general seems a hopeless task; monitoring individual citizens for such enhanced capabilities is more plausible.

Even a non-anthropomorphic human intelligence could pose threats to mankind. The greatest problem is that such artificial intelligence may be indifferent to human welfare.[50] Thus, for instance, unless otherwise programmed, AI could solve problems in ways that harm humans. For instance, computers with autonomy may naturally try to use resources

efficiently and pursue their own objectives to the maximum. Such behavior could harm humanity, because an unrelenting focus on efficiency and maximal pursuit of any objective could easily bump up against human values other than efficiency and those objectives.[51] The solution likely lies in assuring that AI develops in an environment that cultivates the objectives of helpfulness and care.

In any event, the dangers from any variety of unfriendly or even indifferent AI provide strong reasons to develop a program of government support for Friendly AI. If Friendly AI maintains a head start in calculating power, its computational capacity can help discover ways to prevent the possible dangers that could emerge from other kinds of artificial intelligence. To be sure, this approach is not a guaranteed route to success, but it seems much more fruitful and practicable than relinquishment.

The question of how to support Friendly AI is a subtle one. The government lacks the knowledge to issue a set of clear requirements that a Friendly AI project would have to fulfill. It also lacks a sufficiently clear definition of what the end state of a Friendly AI looks like. This ignorance may inhibit establishing a prize for reaching Friendly AI or even any intermediate objective that makes progress toward this ultimate goal.[52]

The best way to support Friendly AI may be instead to treat it as a research project, like those funded by the National Institutes of Health. Peer review panels of computer and cognitive scientists would sift through projects and choose those designed both to advance AI and to assure that such advances would be accompanied by appropriate safeguards.[53] At first such a program should be quite modest and inexpensive. Once shown to actually advance the goals of Friendly AI, the program could be expanded—if necessary to the scale of the Manhattan Project during World War II.[54]

The Concern about Battlefield AI

The concern about robots on the battlefield is likely misplaced, because there are three ways that the movement to robotic forces could be beneficial. First, robots render conventional forces more effective and less vulnerable to certain weapons of mass destruction, like chemical and biological weapons. Rebalancing the world to make such weapons less effective must be counted as a benefit.

Second, the use of robots should reshape the law of war in a way that saves human lives. One of the reasons that conventional armies deploy lethal force is to protect their human soldiers against death or serious injury. When only robots are at stake in a battle, a nation is more likely to use nonlethal force, such as stun guns. In fact, the United States is

considering outfitting some of its robotic forces with nonlethal weaponry. The law of war's prohibitions against intentionally inflicting damage on civilians can be tightened in a world where the combatants are robots. In the long run, robots, whether autonomous or not, will be able to better discriminate among targets than among other kinds of weapons. As a result, the law of war can impose an effectively higher standard on avoiding collateral damage to humans and property that robotic forces would have to meet.[55] The need to protect robots from injury is less urgent than the need to protect humans, particularly as robots will, like other computational devices, become less expensive to make. Thus, the force authorized under international law for protection of robotic weaponry should be proportionately less as well.

Third, advanced computerized weaponry benefits the developed world, particularly the United States, because of its highly developed capability in technological innovation. Robotic weapons have been among the most successful in the fight against al-Qaeda and other groups waging asymmetrical warfare against the United States. The Predator has been successfully targeting terrorists throughout Afghanistan and Pakistan, and more technologically advanced versions are being rapidly developed. Moreover, the Predator is able to find its targets without the need to launch large-scale wars to hold territory—a process that would almost certainly result in more collateral damage.[56] If the United States is generally the best enforcer of rules of conduct that make for a peaceful and prosperous world, it is beneficial to the world that weaponry powered by AI will dominate the battlefield, because that is an area in which the United States is likely to retain an advantage. Thus, for the United States to relinquish computer-driven weaponry would be a grave mistake.[57]

It might be thought that any military exploitation of greater computational capacity is in tension with the development of Friendly AI. But the destructive powers unleashed by computer-driven weaponry do not necessarily entail the creation of strong AI that would lead to computers displacing humanity as a whole. Military activity by culture and design strongly reflects a system of hierarchy and subordination. Thus, nations will focus on integrating robots into their systems of command and control, putting an emphasis on assuring that orders are obeyed and making it unlikely that research by the military will be the origins of a strong AI that would seek to displace humans.

In any event, the United States and other advanced industrial nations can be better trusted than less developed nations to take account of these dangers, particularly if they have an ongoing Friendly AI program. The combination of support for civilian research into Friendly AI and continued deployment of computerized weaponry by the United States remains a better policy than the alternatives of relinquishment or robotic disarmament.

The Benefits of AI in an Age of Accelerating Technology

Artificial intelligence differs from other accelerating technologies like biotechnology and nanotechnology in that it may help social governance by more successfully evaluating the consequences of various policies, including those toward AI itself. Insofar as artificial intelligence remains beneficent, it will facilitate the gathering and analysis of information that helps collective decision making about other technologies as well as about more quotidian matters of social policy, from welfare reform to tax policy.

In particular, improving AI will help with three important tasks in addressing social policy. First, it will help gather and organize the burgeoning amount of socially relevant data that technology continues to generate. Second, it will help form hypotheses about the data—hypotheses that human researchers may not come up with on their own. Finally, it will help improve our knowledge of social consequences by running simulations, and testing whether explanations of the effects of social policy are robust when small changes are made in conditions or assumptions.

Accelerating computer power is creating the phenomenon of "big data" as ubiquitous monitors track both material events, like the temperature at discrete localities or the paths of automobiles, and social events, like blog posts and tweets. But organizing such data will be an increasing challenge in a world of technological acceleration. The accelerating accumulation of information in the world makes mechanisms drawn from AI all the more necessary in order to sort and analyze that information.[58] Already the U.S. military is having trouble analyzing all the information acquired from its drones and other surveillance devices, because it lacks sufficient sorting and analytic capacity.[59]

The military's data problem captures future dilemmas for social decision making. As accelerating technology creates new complexity more rapidly in areas like nanotechnology, biotechnology, and robotics, social decision making must struggle to keep up with analyzing the wealth of new data. Societies prosper if they can use all the information available to make the best decisions possible. The problem now is that the information available to be processed may be swelling beyond our human capacity to achieve sound social decision making without the aid of AI.

The amount of data generated will in turn provide the basis for more conclusions about the effects of various social policies. Progress in AI can also be very useful by allowing computers to assist humans in explaining social phenomena and predicting the trajectory and effects of social trends. By 2020 computers are expected to be able to generate testable hypotheses. We will then no longer have to depend on the ingenuity of human researchers alone for formulating the full range of explanatory causes of social behavior and the effects of government policy.[60]

One way to understand this development is to see the increase in computational power as allowing more hypotheses to emerge from data rather than being imposed on the data. Greater computational power may allow computers to create competition between such emerging hypotheses, with the winner being the one that is objectively best supported. Ian Ayres, in his book *Supercrunchers*, has noted, "Trolling through databases can reveal underlying causes that traditional experts never considered."[61] In the future, however, computers may do the trolling autonomously, even constructing databases that will be optimally trolled.

Experts will also integrate their skills with computers, improving predictions within their fields. As Garry Kasparov has noted, humans plus computers remain better at chess than computers alone, even if the computer on its own is more powerful than any person on his own.[62] Human skills complement those of computers in hypothesis creation as well. Intuition and common sense help balance the computer's relentless application of formal algorithms.

The process of integrating human expertise and AI power may well be a recursive one. Different experts will use AI programs to look at the data and try to perfect models of an issue, like climate change. Then computers can use algorithms that discount certain experts because of past mistakes and weigh the combination of their predictions to form a result that is likely better than any single prediction.[63] This combined prediction then becomes grist for human evaluation through information markets. Similarly, regulatory experiments can begin through randomized trials, but computer algorithms can then tell us the rate at which to move toward the more successful regulation and ultimately when to end the experiment and choose the best regulation. The process is like using an algorithm to direct how money should be placed in different slot machines with unknown payout formulas. As different payouts are observed, it becomes rational to change the proportions of the slot machines on which bets are placed. The more general point here is that synthesizing a variety of methods—experiments, the analysis of experts, AI assistance, and prediction markets—may often yield better predictions than a single method can provide. Government can thus improve policy analysis by encouraging the applications of many different kinds of information technology.

Computer simulations will also become more powerful, permitting researchers to vary certain data from what exists and see what results.[64] Simulations will help enhance the robustness of modeling and empiricism that test the likely effects of social policy. Almost all social science analysis is complicated by statistical noise. Such noise can be the result of errors in measurements of the behavior to be explained or the result of the difficulty of capturing all the causes of human behavior. Simulation

allows the researcher to vary some of the conditions and see if the results stay relatively similar. If they do, the empirical results are more robust and more likely to provide a map of the actual world.

One should not object to using simulations because of the fact that they are not themselves direct representations. Since measurement and other knowledge are limited, all data are themselves approximations of the world rather than direct representations. Computer simulations, when used in conjunction with empirical methods, are likely to make our approximations for social science more accurate and our conclusions less likely to be changed by the perturbations of the actual world.

A particularly promising area for simulations is so-called agent-based modeling, which allows a researcher to specify the behavior of virtual actors or agents.[65] These agents can then interact with another agent through repeated rounds of action and can then change their behavior through the rounds depending on what they have learned. These simulations permit researchers to find "emergent properties, that is properties arising from the interactions of the agents that cannot be deduced simply by aggregating the properties of the agents. When the interaction of the agents is contingent on past experiences . . . [agent-based modeling] might be the only practical method" of capturing a world with social learning.[66] Such modeling permits us to use the best evidence of regularities of human behavior and to refine our predictions by recognizing that actors may learn and thus change their behavior. In this sense, agent-based modeling has some of the advantages of prediction markets, which consider empirical analysis of past behavior but also force bettors to consider whether learning from past behavior will change future behavior.

Considered at their most general, these latter uses of AI expand our knowledge of the world by looking at the adjacent possible worlds, thereby helping society to see more clearly the likely results of government interventions. Increasing the capacity of AI to formulate hypotheses and to structure simulations could potentially render better decisions in the high stakes created by accelerating technology. That prospect itself justifies government subsidization of Friendly AI.

The greater capacity to process information should also better predict natural catastrophes, providing us with advance warning to prevent them or take preemptive measures to avoid their worst consequences. The more sophisticated the simulations and modeling of earthquakes, weather, and asteroids, and the better aggregation of massive amounts of data on those phenomena, the more accurate such measures are likely to be.[67] Moreover, by estimating the risks of various catastrophes, society is better able to make collective decisions to use its limited resources to focus on the most serious ones.

To be clear, AI's social utility does not depend on predicting the future with complete precision. Given the randomness inherent in our world, that feat is impossible, no matter how great the increase in intelligence.[68] Even if AI only makes clear the possibility of unexpected future contingencies, offers some solutions that would not otherwise be contemplated, and provides some assessment of their likelihood of success, it will help improve policy.

Government policy toward AI should be rooted in two complementary rationales: the need to deploy AI as an aid in collective decision making and the need to forestall dangers from strong AI. Each is individually sufficient to justify a program of support for Friendly AI. Together they make a most compelling case. The question of what degree and form of support is warranted will be subtle and difficult. But that is the right question to ask, not whether we should retard AI's development with complex regulations or, still worse, relinquish it.