

ACCELERATING AI

*John O. McGinnis**

Recently, Artificial Intelligence (AI) has become a subject of major media interest. For instance, last May the *New York Times* devoted an article to the prospect of the time at which AI equals and then surpasses human intelligence.¹ The article speculated on the dangers that such an event and its “strong AI” might bring.² Then in July, 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, such as the predator drones that the United States now uses to kill terrorists.³

These articles encapsulate the twin fears about AI that may impel regulation in this area—the existential dread of machines that become uncontrollable by humans and the political anxiety about machines’ destructive power on a revolutionized battlefield. Both fears are overblown. 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. In any event, given the centrality of increases in computer power to military technology, it would be impossible to regulate research into AI without empowering the worst nations on earth.

Instead of prohibiting or heavily regulating artificial intelligence, the United States should support civilian research into a kind of AI that will not endanger humans—a so-called “friendly AI.”⁴ First, such support is the

* Stanford Clinton Sr. Professor of Law, Northwestern University. I thank Mark Movsesian, Michael Rappaport, and Ardith Spence for their helpful comments. I am also very grateful to Michael Abramowicz for our discussions on this subject in our joint seminar, Law and Accelerating Technology.

¹ See John Markoff, *The Coming Superbrain*, NYTIMES.COM, May 23, 2009, <http://www.nytimes.com/2009/05/24/weekinreview/24markoff.html> (link).

² *Id.*

³ See John Markoff, *Scientists Worry Machines May Outsmart Man*, NYTIMES.COM, July 25, 2009, <http://www.nytimes.com/2009/07/26/science/26robot.html> (link); NYTIMES.COM, Predator Drones and Unmanned Aerial Vehicles, http://topics.nytimes.com/top/reference/timestopics/subjects/u/unmanned_aerial_vehicles/index.html (last visited Mar. 29, 2010) (link).

⁴ For a definition of “friendly AI,” see SINGULARITY INSTITUTE FOR ARTIFICIAL INTELLIGENCE, CREATING FRIENDLY AI § 1, <http://www.singinst.org/upload/CFAI.html> (last visited Mar. 29, 2010), which summarizes the goals of friendly AI as assuring that AI seeks the elimination of “involuntary pain, death, coercion, and stupidity.” (link). I might suggest an even weaker definition as simply assuring that AI does not create harm to humans or limit their freedom through either malevolence or stupidity.

best way to make sure that computers do not turn out to be an existential threat. It would provide incentives for researchers in the most technologically advanced nation in the world to research and develop AI that is friendly to man.

Second, such support is justified because of the positive spillovers that computational advances will likely provide in collective decisionmaking. The acceleration of technology creates the need for quicker government reaction to the potentially huge effects of disruptive innovations. For instance, at the dawn of the era in which the invention of energy-intensive machines may have started to warm up the earth, few recognized any risk from higher temperatures that such machines might cause.⁵ Yet as I will describe below, current developments in technology make the rise of energy-intensive machines seem slow-moving. Assuming that man-made atmospheric warming is occurring,⁶ it likely presents only the first of a number of possible catastrophes generated by accelerating technological change—dangers that may be prevented or at least ameliorated through earlier objective analysis and warning. But it is no less important to recognize that other technological advances may create a cascade of benefits for society—benefits that false perceptions of risk may retard or even preclude. As a result, gathering and analyzing information quickly is more important than ever to democratic decisionmaking because the stakes of such regulatory decisions have never been higher.

Given that AI has substantial potential to help society formulate the correct policies about all other accelerating technologies with transformational capacity, such as nanotechnology and biotechnology, the most important policy for technological change is that for AI itself. Strong AI would help analyze the data about all aspects of the world—data that is growing at an exponential rate.⁷ AI then may help make connections between policies and consequences that would otherwise go overlooked by humans, acting as a fire alarm against dangers from new technologies whose chain of effects may be hard to assess even if they are quite imminent in historical terms.

Such analysis is not only useful to avoiding disaster but also to take advantage of the cornucopia of benefits from accelerating technology. Bet-

⁵ It is true that a Swiss chemist argued in the 1890s—still well after the beginning of industrialization—that human action could cause global warming. See Bradford C. Mank, *Standing and Global Warming: Is Injury To All Injury to None?*, 35 ENVTL. L. 1, 12 (2005). Nevertheless, even in the 1930s a scientist who reiterated this concern was thought eccentric. See Maxine Burkett, *Just Solutions to Climate Change: A Climate Justice Proposal for a Domestic Clean Development Mechanism*, 56 BUFF. L. REV. 169, 173 n.3 (2008) (link).

⁶ It is not necessary to be confident that man-made atmospheric change is occurring to recognize that it would be useful to evaluate the risk of such an event. As I suggest later, one of the advantages of AI is that it will help evaluate such risks.

⁷ See ECONOMIST.COM, *The Data Deluge*, Feb. 25, 2010, http://www.economist.com/opinion/displaystory.cfm?story_id=15579717 (describing the rapid growth of data available in the modern world) (link).

ter analysis of future consequences may help the government craft the best policy toward nurturing such beneficent technologies, including providing appropriate prizes and support for their development. Perhaps more importantly, better analysis about the effects of technological advances will tamp down on the fears often sparked by technological change. The better our analysis of the future consequences of current technology, the less likely it is that such fears will smother beneficial innovations before they can deliver Promethean progress.

In this brief Essay, I first describe why strong AI has a substantial possibility of becoming a reality and then sketch the two threats that some ascribe to AI. I show that relinquishing or effectively regulating AI in a world of competing sovereign states cannot respond effectively to such threats, given that sovereign states can gain a military advantage from AI, and that even within states, it would be very difficult to prevent individuals from conducting research into AI. Moreover, I suggest that AI-driven robots on the battlefield may actually lead to less destruction, becoming a civilizing force in wars as well as an aid to civilization in its fight against terrorism. Finally, I offer reasons that friendly artificial intelligence can be developed to help rather than harm humanity, thus eliminating the existential threat.

I conclude by showing that, in contrast to a regime of prohibition or heavy regulation, a policy of government support for AI that follows principles of friendliness is the best approach to artificial intelligence. If friendly AI emerges, it may aid in preventing the emergence of less friendly versions of strong AI, as well as distinguish the real threats from the many potential benefits inherent in other forms of accelerating technology.

I. THE COMING OF 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*.⁸ The notion of computers that rival and indeed surpass human intelligence might first seem to be speculative fantasy, rather than a topic that should become a salient item on the agenda of legal analysis and policy.⁹ But travel to the moon was itself once a staple of science fiction in the nineteenth and twentieth centuries.¹⁰ Yet because of a single government program, man's exploration of the moon is now a historical event of more than forty years standing. And unlike a lunar landing,

⁸ See *2001: A SPACE ODYSSEY* (Metro-Goldwyn-Mayer 1968); *WALL-E* (Pixar Animation Studios 2008).

⁹ Consideration of artificial intelligence has not bulked large in legal scholarship. One interesting article analyzes whether artificial intelligence can play the role of a trustee. See Lawrence B. Solum, *Legal Personhood for Artificial Intelligences*, 70 N.C. L. REV. 1231 (1992).

¹⁰ See, e.g., JULES VERNE, *FROM THE EARTH TO THE MOON* (1865).

the development of artificial intelligence has direct implications for social governance.

Strong artificial intelligence is the creation of machines with the general human capacity for abstract thought and problem solving. It is generally conceded that if such machines are 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.¹¹

The success of strong AI depends on the truth of three premises. The first premise is functionalism. Functionalism turns on the proposition that cognition is separate from the system in which cognition is realized.¹² 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, is “capable of intelligent action.”¹³

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.¹⁴ 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.¹⁵ Yet he himself understands nothing of Chinese and, as a result, this manipulation of symbols is a poor simulacrum of human understanding.¹⁶ One powerful objection to Searle’s analogy is that the entire system—the written directions plus the human manipulator—does understand Chinese.¹⁷ Searle thus unfairly anthropomorphizes the subject of understanding. As I discuss below, 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 made in thinking about computational intelligence—an error that science fiction has perpetuated.¹⁸

The second claim undergirding strong AI is that computers will have the hardware capacity to mimic human thought. Raw computer power has been growing exponentially according to Moore’s law. Moore’s law,

¹¹ See Irving John Good, *Speculations Concerning the First Ultrainelligent Machine*, 6 *ADVANCES IN COMPUTERS* 31, 31–36 (1965) (link).

¹² See HENRY BRIGHTON & HOWARD SELINA, *INTRODUCING ARTIFICIAL INTELLIGENCE* 42 (2003).

¹³ See Allen Newell & Herbert A. Simon, *Computer Science as Empirical Inquiry: Symbols and Search*, 19 *COMM. OF THE ACM* 113, 118 (1976) (link).

¹⁴ John R. Searle, *Minds, Brains, and Programs*, 3 *BEHAV. & BRAIN SCI.* 417, 417–18 (1980).

¹⁵ See *id.* at 418.

¹⁶ See *id.*

¹⁷ See, e.g., DANIEL C. DENNETT, *CONSCIOUSNESS EXPLAINED* 439 (1991).

¹⁸ See *infra* notes 52–57 and accompanying text.

named after Gordon Moore, one of Intel's founders, is the observation that the number of transistors fitting onto a computer chip doubles every eighteen months to two years.¹⁹ This prediction, which has been approximately accurate for the last forty years, means that almost every aspect of the digital world—from computational calculation power to computer memory—is growing in density at a similarly exponential rate.²⁰ Moore's law reflects the rapid rise of computers as the fundamental engine of mankind in the late twentieth and early twenty-first centuries.²¹

The power of exponential growth is hard to overstate. As Robert Lucas once said in the economic context, once you start thinking about exponential growth, "it is hard to think about anything else."²² The computational power in a cell phone today is a thousand times greater and a million times less expensive than all of the computing power housed at MIT in 1965.²³ Projecting forward, the computational power of computers thirty years from now is likely to prove a million times more powerful than that of computers today.²⁴

To be sure, some technology pundits have long been predicting the imminent death of Moore's law, but it has nevertheless continued to flourish. Intel, a company that has a substantial interest in accurately telling software makers what to expect, projects that Moore's law will continue until at least 2029.²⁵ Technology theorist and inventor Ray Kurzweil shows that Moore's law is actually part of a more general exponential computation growth that has been gaining force for over one hundred years.²⁶ Integrated circuits replaced transistors, which previously replaced vacuum tubes, which in their time had replaced electromechanical methods of computation. Through all these changes in the mechanisms of computation, its power has increased at an exponential rate.²⁷ This historical perspective suggests that new methods under research, from carbon nanotechnology to optical computing to quantum computing, will likely permit computational

¹⁹ See Moore's Law: Made Real by Intel Innovation, <http://www.intel.com/technology/mooreslaw> (last visited Mar. 29, 2010) (discussing Moore's law, which predicts that the number of transistors on a silicon chip will roughly double every eighteen to twenty-four months, thus increasing microprocessor speed on a regular basis) (link).

²⁰ See Dan L. Burk & Mark A. Lemley, *Policy Levers in Patent Law*, 89 VA. L. REV. 1575, 1620 n.147 (2003).

²¹ Cf. HENRY ADAMS, *THE EDUCATION OF HENRY ADAMS* 379–90 (1918) (discussing the Virgin as the symbol of the Middle Ages and the steam engine as that of the nineteenth century).

²² Robert E. Lucas, Jr., *On the Mechanics of Economic Development*, 22 J. OF MONETARY ECON. 3, 5 (1988) (link).

²³ Ray Kurzweil, *Making the World a Billion Times Better*, WASHINGTONPOST.COM, Apr. 13, 2008, <http://www.washingtonpost.com/wp-dyn/content/article/2008/04/11/AR2008041103326.html> (link).

²⁴ See HANS MORAVEC, *ROBOT: MERE MACHINE TO TRANSCENDENT MIND* 104–08 (1999).

²⁵ Jeremy Geelan, *Moore's Law: "We See No End in Sight," says Intel's Pat Gelsinger*, SOA WORLD MAGAZINE (May 1, 2008), available at <http://java.sys-con.com/read/557154.html> (link).

²⁶ RAY KURZWEIL, *THE SINGULARITY IS NEAR: WHEN HUMANS TRANSCEND BIOLOGY* 67 (2005).

²⁷ *Id.*

power to continue growing exponentially, even when silicon-based computing reaches its physical limits.²⁸ 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 2030.²⁹ Even if this pace does not continue, it seems hard to believe that this capacity will not be reached by the midpoint of this century.

The third issue is whether programmers will be able to provide the software to convert the gains in hardware to make advances in AI. No doubt there are daunting challenges ahead in creating software that can understand the complex realities captured by human thought. In fact, some have argued that despite the previous growth in computational capacity, AI has been largely a failure with little to show for fifty years of work.³⁰ This assessment seems far too harsh. Since 1997, computers have been able to defeat the greatest chess players in the world.³¹ Cars run by computers can autonomously navigate city traffic.³² These feats can hardly be dismissed as powerful examples of intelligent behavior. Of course, it is true that chess is a completely formal system and even driving has a limited—although more unpredictable—problem set to be solved. But it is hardly surprising that artificial intelligence proceeds from creating intelligence in more formal and predictable environments to doing so in more informal and fluid ones. In any event, software progress continues in tandem with the growing hardware capability.³³

My point here is not to prove that general AI will succeed in replicating and then surpassing human intelligence, but just to suggest that such a prospect is quite 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 on problems that remain unsolved. Such progress can be very useful for social decisionmaking 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 generate testable hypotheses; thus we will not have to de-

²⁸ For a good introduction to quantum computing, see GEORGE JOHNSON, *A SHORTCUT THROUGH TIME: THE PATH TO THE QUANTUM COMPUTER* (2003).

²⁹ See KURZWEIL, *supra* note 26, at 125–27.

³⁰ See BRIGHTON & SELINA, *supra* note 12, at 23.

³¹ See IBM Research: Deep Blue Overview, <http://www.research.ibm.com/deepblue/watch/html/c.shtml> (record of match between chess champion Gary Kasparov and computer Deep Blue) (link).

³² See *CMU Robot Car First in DARPA Urban Challenge*, SPIE.ORG, Nov. 5, 2007, <http://spie.org/x17538.xml> (describing and providing video of the Defense Advanced Research Projects Agency (DARPA) challenge race around urban setting) (link).

³³ See *Data, Data Everywhere*, ECONOMIST.COM, Feb. 25, 2010, http://www.economist.com/specialreports/displaystory.cfm?story_id=15557443 (suggesting that “improvements in the algorithms driving computer applications have played as important a part as Moore’s law for decades”) (link).

pend only on the ingenuity of researchers for testing the full range of explanations of the causes of matters.³⁴

One way to understand this development is to see 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 that which is objectively best supported. Computer simulations will also become more powerful and permit researchers to vary certain data from that which exists and see what results.³⁵ Such simulations will help enhance the robustness of modeling and empiricism that help in social analysis. The analytic capability of AI offers positive spillovers, because a sophisticated and quantitatively informed understanding of the current and future shape of society is a great social good, aiding citizens and politicians alike in coming to sound policy decisions.

To be clear, strong AI's social utility does not depend on predicting the future with precision. Given the randomness inherent in the world, that feat is not possible, no matter how great the increase in intelligence.³⁶ Even if AI only makes clear the possibility of unexpected future contingencies and offers some assessment of their likelihood with evaluation of possible solutions (including out-of-the-box ideas), AI will aid in planning for the contingencies.

II. THE THREATS OF AI

Many theorists of technology are very optimistic about the capacity of strong AI to substantially improve human life. Robots powered by AI can make life much easier, particularly for the disabled and elderly. AI has the potential to aid in discoveries that extend the human life span. At their extremes, some theorists of technology, like Kurzweil, believe that AI will lead, even within fifty years, to a kind of technological utopia where people enjoy very high incomes and very long or perhaps even indefinite life spans.³⁷

³⁴ See Stephen H. Muggleton, *Exceeding Human Limits*, 440 NATURE 409 (2006) (link).

³⁵ See JOSHUA M. EPSTEIN & ROBERT L. AXTELL, *GROWING ARTIFICIAL SOCIETIES: SOCIAL SCIENCE FROM THE BOTTOM UP* (1996). Social scientists already do this type of data varying in the social network arena, in fact. See Stephen P. Borgatti, *Identifying Sets of Key Players in a Social Network*, 12 COMPUTATIONAL & MATHEMATICAL ORG. THEORY 21, 33 (2006) (acknowledging that if 10% of the links in a particular social network may not actually exist, an analyst can randomly vary the dataset by 10% and re-apply the model to identify a set of the network's most important players that is "not necessarily optimal for the observed dataset, but will represent a high-quality solution for the neighborhood of the graph as a whole").

³⁶ See NASSIM NICHOLAS TALEB, *FOOLED BY RANDOMNESS: THE HIDDEN ROLE OF CHANCE IN LIFE AND IN THE MARKETS*, at xliii (discussing the tendency to underestimate the randomness of life) (Random House 2008) (2004).

³⁷ See KURZWEIL, *supra* note 26, at 122, 320–30.

Even those concerned that strong AI may threaten the existence of humanity premise their fears on the view that such artificial intelligence is possible and relatively imminent. For instance, Bill Joy, the former chief technologist for Sun Microsystems, does not disagree with Kurzweil that we are entering an age of unprecedented technological acceleration in which artificial intelligence will become vastly more powerful than it is today. But his outlook on this development is deeply pessimistic. In a widely discussed article, “Why The Future Doesn’t Need Us,” he raises the alarm that man cannot ultimately control these machines.³⁸ The power of his critique lies precisely in his acknowledgement 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. It appears that in his view, man resembles the sorcerer’s apprentice—too weak and too ignorant to master the master machines. Joy’s stance represents the culmination of a particular kind of fear that goes back to the Romantic Era and was first represented by the Frankenstein monster who symbolized the idea that “all scientific progress is really a disguised form of destruction.”³⁹

Fears of artificial intelligence on the battlefield may be an even more immediate concern raised by growing computational power. Nations have always attempted to use technological innovation to gain advantages in warfare.⁴⁰ Computational advance today is essential to national defense, or to put it more globally, sovereign military competition. The Defense Advanced Research Projects Agency (DARPA) spends billions of dollars developing more advanced military mechanisms that depend on ever more substantial computational capacity.⁴¹

It is hard to overstate the extent to which advances in robotics, which are driven by AI, are transforming the United States military. During the Afghanistan and Iraq wars, more and more Unmanned Aerial Vehicles (UAVs) of different kinds were used. For example, in 2001, there were ten unmanned “Predators” in use, and at the end of 2007, there were 180.⁴² Unmanned aircraft, which depend on substantial computational capacity, are an increasingly important part of our military and may prove to be the

³⁸ See Bill Joy, *Why The Future Doesn’t Need Us*, WIRED.COM, Apr. 2000, <http://www.wired.com/wired/archive/8.04/joy.html> (link). He is not alone in his concern. See KEVIN WARWICK, *MARCH OF THE MACHINES: THE BREAKTHROUGH IN ARTIFICIAL INTELLIGENCE* 280–303 (2004) (“Machines will then become the dominant life form on Earth.”).

³⁹ See RICHARD HOLMES, *THE AGE OF WONDER: HOW THE ROMANTIC GENERATION DISCOVERED THE BEAUTY AND TERROR OF SCIENCE* 94 n. (Vintage Books 2010) (2008).

⁴⁰ See ROBERT FRIEDEL, *A CULTURE OF IMPROVEMENT: TECHNOLOGY AND THE WESTERN MILLENNIUM* 85, 113, 131, 374 (2007) (discussing the relationship between technological innovation and war).

⁴¹ See Noah Shachtman, *DARPA Chief Speaks*, DANGER ROOM, Feb. 20, 2007, http://blog.wired.com/defense/2007/02/tony_tether_has_1.html (talking about defense developments, some of which depend on computational innovations) (link).

⁴² See P.W. SINGER, *WIRED FOR WAR: THE ROBOTICS REVOLUTION AND CONFLICT IN THE 21ST CENTURY* 35 (2009).

majority of aircraft by 2020.⁴³ Even below the skies, robots perform important tasks such as mine removal.⁴⁴ Already in development are robots that would wield lasers as a kind of special infantryman focused on killing snipers.⁴⁵ Others will act as paramedics.⁴⁶ 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 different set of fears than those raised by the potential autonomy of AI. Here, the concern is that human malevolence will lead to these ever more capable machines wreaking ever more havoc and destruction.

III. THE FUTILITY OF THE RELINQUISHMENT OF AI AND THE PROHIBITION OF BATTLEFIELD ROBOTS

Joy argues for “relinquishment”—i.e., the abandonment of technologies that can lead to strong AI. Those who are concerned about the use of AI technology on the battlefield would focus more specifically on 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, at least under current technology and current geopolitics, are certain to be ineffective. Thus, nations are unlikely to unilaterally relinquish 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 both a flourishing commercial republic that benefits from global peace and prosperity, and the world’s hegemon, capable of supplying the public goods of global peace and security. Because it gains a greater share of the prosperity that is afforded by peace than do 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.⁴⁷ By relinquishing the power of AI, the United States would in fact be giving greater incentives to rogue nations to develop it.

Thus, the only realistic alternative to unilateral relinquishment would be a global agreement for relinquishment or regulation of AI-driven weaponry. But such an agreement would face the same insuperable obstacles nuclear disarmament has faced. As recent events with Iran and North Korea demonstrate,⁴⁸ it seems difficult if not impossible to persuade rogue nations

⁴³ See Rowan Scarborough, *Unmanned Warfare*, WASH. TIMES, May 8, 2005, at A1.

⁴⁴ For a popular account in a major film, see *THE HURT LOCKER* (First Light Production 2009).

⁴⁵ See Singer, *supra* note 42, at 111.

⁴⁶ *Id.* at 112.

⁴⁷ See John O. McGinnis & Ilya Somin, *Should International Law Be Part of Our Law?*, 59 STAN. L. REV. 1175, 1236–38 (2007).

⁴⁸ See Gustavo R. Zlauvinen, *Nuclear Non-Proliferation and Unique Issues of Compliance*, 12 ILSA J. INT’L & COMP. L. 593, 595 (2006).

to relinquish nuclear arms. Not only are these weapons a source of geopolitical strength and prestige for such nations, but verifying any prohibition on the preparation and production of these weapons 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 arriving at strong artificial intelligence has many routes and still more that are likely to be discovered. Moreover, building a nuclear bomb requires substantial infrastructure.⁴⁹ Artificial intelligence research can be done in a garage. Constructing a nuclear bomb requires very substantial resources beyond that of most groups other than nation-states.⁵⁰ Researching artificial intelligence is done by institutions no richer than colleges and perhaps would require even less substantial resources.

Joy recognizes these difficulties, but offers no plausible solution. Indeed, his principal idea for implementing relinquishment shows that his objective is impossible to achieve. He suggests that computer scientists and engineers take a kind of Hippocratic Oath that they will not engage in AI research with the potential to lead to AI that can displace the human race.⁵¹ 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 likely not permit their scientists to respect it because of the importance of computational advances to national defense. Even left on their own, scientists would likely disregard the oath because of the substantial payoffs for advances in this area from private industry. Finally, scientists would have difficulty complying with such a directive even should they want to because of the difficulty of predicting what discoveries will propel AI forward in the long-run.

For all these reasons, verifying a global relinquishment treaty, or even one limited to AI-related weapons development, 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 such a prohibition on AI research directed wholly against its own residents. Even a domestic prohibition would run up against the substantial incentives to pursue such research because the resulting inventions can provide lucrative applications across a wider range of areas than can research into nuclear weapons.

⁴⁹ See William J. Perry, *Proliferation on the Peninsula; Five North Korean Nuclear Crises*, 607 ANNALS AM. ACAD. POL. & SOC. SCI. 78, 78–79 (2006) (link).

⁵⁰ See *id.*

⁵¹ See Joy, *supra* note 38 (link).

IV. CONCEPTUAL ERRORS IN FEARS ABOUT 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 empirical confusions.

A. *The Existential Threat*

The existential threat can be dissolved if there is a substantial possibility of constructing friendly AI.⁵² 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 argument 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.⁵³ 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.⁵⁴ 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.⁵⁵ 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. One might think that such a process would be more likely to give rise to existential threats. Further, one cannot rule out that a property of malevolence, or at a least a will to power, could be an emergent property of a particular line of AI research.

Moreover, even a non-anthropomorphic human intelligence still could pose threats to mankind, but they are probably manageable threats. The greatest problem is that such artificial intelligence may be indifferent to

⁵² See SINGULARITY INSTITUTE FOR ARTIFICIAL INTELLIGENCE, *supra* note 4, § 2 (offering a detailed program for friendly AI). In this Essay, I abstract from any detailed program for realizing friendly AI. Instead, I would define friendly AI in terms of a goal—creating human-like cognitive capacity which will not autonomously decide to harm humans. As discussed below, deciding what steps will improve the likelihood of realizing this goal may be a difficult task, probably best undertaken by the kind of grant-making now administered by the National Institutes of Health in the medical field.

⁵³ See *id.*

⁵⁴ See ROBERT WRIGHT, *THE MORAL ANIMAL: EVOLUTIONARY PSYCHOLOGY AND EVERYDAY LIFE* 336–37 (1994). See also John O. McGinnis, *The Human Constitution and Constitutive Law: A Prolegomenon*, 8 J. CONTEMP. LEGAL ISSUES 211, 213 (1997) (describing the evolutionary drive to preserve and spread one's genes).

⁵⁵ WRIGHT, *supra* note 54, at 336–37.

human welfare.⁵⁶ Thus, for instance, unless otherwise programmed, it could solve problems in ways that could lead to harm against humans. But indifference, rather than innate malevolence, is much more easily cured. Artificial intelligence can be programmed to weigh human values in its decisionmaking.⁵⁷ The key will be to assure such programming.

The realistic prospect of wholly friendly AI is the first reason that government should support rather than regulate AI. Such support can provide strong incentives for developers of AI to take this issue seriously. If the mechanisms of artificial intelligence developed through such a program maintain a head start in calculating power, they can in turn be useful in finding 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.⁵⁸ The best way to support 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 that are designed both to advance AI and assure that such advances would be accompanied by appropriate safeguards.⁵⁹ 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.⁶⁰

⁵⁶ See SINGULARITY INSTITUTE FOR ARTIFICIAL INTELLIGENCE, *supra* note 4, § 3.2.3.

⁵⁷ *Id.* § 2.

⁵⁸ See Jonathan H. Adler, *Eyes on a Climate Prize: Rewarding Energy Innovation to Achieve Climate Stabilization*, 20–21 (Case Research Paper Series in Legal Studies, Working Paper No. 2010-15, 2010) available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1576699 (offering criteria for when prizes will work better than grants) (link); William A. Masters & Benoit Delbecq, *Accelerating Innovation with Prize Rewards* 10 (International Food Policy Research Institute Discussion Paper No. 835, 2008), available at <http://www.ifpri.org/sites/default/files/publications/ifpridp00835.pdf> (suggesting that prizes can be useful ways to elicit desired projects when the prize competition sets “a feasible but difficult objective [and] a clearly measurable goal, and [disperses] prizes in a predictable manner by an impartial authority”) (link).

⁵⁹ This use of peer review rather than regulation to address technical issues that are not possible to capture through bureaucratic mandates is something that should be considered more generally in the regulatory process. See Stephen P. Croley, *The Administrative Procedure Act and Regulatory Reform: A Reconciliation*, 10 ADMIN. L.J. 35, 47 (1996).

⁶⁰ Thus the policy recommendations here differ somewhat from the Singularity Institute. It considers relinquishment infeasible. See SINGULARITY INSTITUTE FOR ARTIFICIAL INTELLIGENCE, *supra* note 4, § 4.2.2. But the Institute seems to suggest a hands-off approach by the government. *Id.* § 4.2.1. While I agree that regulatory prohibitions are imprudent, government support for friendly AI is useful to give it the momentum to triumph over potentially less friendly versions, the possibility of

B. The Concern about Battlefield AI

It is not as if in the absence of AI wars or weapons will cease to exist. The way to think about the effects of AI on war is to think of the consequences of substituting technologically advanced robots for humans on the battlefield. In at least three ways, that substitution is likely to be beneficial to humans.

First, robots make 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, even if marginally so, must be counted as a benefit.

Second, one of the reasons that conventional armies deploy lethal force is to protect the human soldiers against death or serious injury. If only robots are at stake in a battle, a nation is more likely to use non-lethal force, such as stun guns and the like. The United States is in fact considering outfitting some of its robotic forces with non-lethal weaponry.

Third, AI-driven weaponry gives an advantage to the developed world and particularly to the United States, because of its advanced 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, a robotic airplane, has been successfully targeting terrorists throughout Afghanistan and Pakistan, and more technologically advanced versions are being rapidly developed. Moreover, it does so in a targeted manner without the need to launch large-scale wars to hold territory—a process that would almost certainly result in more collateral damage.⁶¹ If one believes that the United States is on the whole the best enforcer of rules of conduct that make for a peaceful and prosperous world, this development must also be counted as a benefit.

Importantly, the law of war can be adapted to the use of robots. The basic requirements of the prohibition against intentionally inflicting damage on civilians should have no less force when the inflictors of damage are robots. In the long run, robots, whether autonomous or not, may be able to discriminate better than other kinds of weapons, thus allowing a higher standard for avoidance of civilian deaths.⁶² The requirement of proportio-

which I do not think can be ruled out a priori. AI is potentially friendly but not necessarily so. Moreover, as I discuss in Part V, the positive externalities of AI for collective decisionmaking suggest that such support is warranted.

⁶¹ As discussed by Professor Kenneth Anderson, see Kenneth Anderson's Law of War and Just War Theory Blog, *Why Targeted Killing? And Why Is Robotics so Crucial an Issue in Targeted Killing?*, <http://kennethandersonlawofwar.blogspot.com/2009/03/why-targeted-killing-and-why-is.html#links> (Mar. 27, 2009, 12:36 EST) (link).

⁶² Cf. *Rise of the Drones: Unmanned Systems and the Future of War: Hearing Before the Subcomm. on National Security and Foreign Affairs of the H. Comm. on Oversight and Government Reform*, 111th Cong. 11 (2010), available at http://oversight.house.gov/images/stories/subcommittees/NS_Subcommittee/3.23.10_Drones/Anderson.pdf (written testimony of Kenneth Anderson) (comparing the collateral damage caused by drones and an artillery barrage) (link).

nality in war may even have more bite, since the need to protect robots from all injury may be less than the need to protect humans, so the force effectively authorized under international law for troop protection may be proportionately less. Thus, relinquishment by the United States would be a grave mistake if it were substantially possible that artificial intelligence consistent with continued human flourishing could be constructed.⁶³ It might be thought that any military exploitation of artificial intelligence is in tension with development of friendly AI. But the destructive powers unleashed by computer-driven weaponry does not necessarily entail the creation of strong AI that would lead to computers displacing humanity as a whole. Moreover, even if ultimately competition among nation-states leads to pressure to develop the use of strong AI in military matters, there will be powerful incentives for the United States to constrain the AI that drives such weaponry from engaging in the kind of behavior that Joy fears. In any event, it is likely that the United States and other advanced industrial nations can be better trusted than other nations to take account of these dangers, particularly if they have an ongoing friendly AI program. Thus, the combination of support for friendly AI and research into and deployment of computerized weaponry by the United States remains a better policy than the alternatives of relinquishment or disarmament.

V. THE BENEFITS OF AI IN AN AGE OF ACCELERATING TECHNOLOGY

We live in an age of accelerating technology. Because of the exponential growth symbolized and exemplified by Moore's law, technological innovation progresses faster than ever before. Some theorists of technology believe that such acceleration has been going on throughout human history.⁶⁴ It took much less time for the industrial age to replace the era of agriculture than it did for that era to succeed the long eon of hunter-gathering of our distant past. The post-industrial age is following on the heels of industrialization more quickly still.

Even within our own lifetimes, the quickening of the pace of change is palpable. A visit to an electronics store, or even a grocery store, would find a whole new line of products within two years, whereas someone visiting a store between 1910 and 1920—let alone 1810 and 1820—would not have noticed much difference. Even cultural generations move faster. Facebook, for instance, has in a few years completely changed the way college students relate,⁶⁵ whereas the tenor of college life in 1960 would have been recognizable to a student in 1970.⁶⁶

⁶³ See SINGULARITY INSTITUTE FOR ARTIFICIAL INTELLIGENCE, *supra* note 4, §§ 1, 2, 4.

⁶⁴ See KURZWEIL, *supra* note 26, at 17–19.

⁶⁵ Maria Tess Shier, *The Way Technology Changes How We Do What We Do*, 2005 NEW DIRECTIONS FOR STUDENT SERVS. 77, 83 (2005) (link). See also Matthew Robert Vanden Boogart, *Uncovering the Social Impacts of Facebook on a College Campus* (2006) (unpublished M.S. dissertation,

NORTHWESTERN UNIVERSITY LAW REVIEW COLLOQUY

Most importantly for society, accelerating technology is spawning a multitude of innovations—even fields of innovations unknown three decades ago. Biotechnology, for instance, raises the possibility of radical life extension, and with it, very important changes in demography.⁶⁷ Nanotechnology, which concerns the control and fabrication of matter smaller than one billionth of a meter, is also proceeding on a wide front and has already been incorporated into products from sunscreen to pesticides.⁶⁸ While the field has enormous promise, it also raises potential threats. These include relatively mundane threats, such as the bodily damage that tiny particles can cause, and much wilder scenarios in which nanomachines replicate themselves until the earth is destroyed.⁶⁹ Beyond these specific fields, the reality of accelerating technology will result in wholly new disruptive innovations that cannot now be predicted. But these too may rapidly shape the structure of society for good or ill.

Such technological changes may sometimes require new laws, regulations, and social norms to avoid their dangers and realize their promise. But these technologies do not themselves directly facilitate the better information gathering and analysis necessary to superintend these technological transformations. In other words, they do not themselves directly provide the information conducive to the regulation and integration of accelerating technologies into society. Artificial intelligence is fundamentally different in this respect. Insofar as artificial intelligence remains beneficent, it facilitates the gathering and analysis of information that helps the regulation of further advances not only in its own field, but in other fields of accelerating technology as well. It has the added advantage that its focus will be on the processing of objective information. While the growth of AI will not end ideological battles over the interpretation of data, more and more objectively analyzed facts provide ballast to deliberation, preventing extreme and unsupported claims and providing an anchor to democratic consensus.

Indeed, one way of thinking of the importance of AI is that the accelerating accumulation of information in the world makes mechanisms to sort

Kansas State University), available at <http://krex.k-state.edu/dspace/bitstream/2097/181/1/MatthewVandenBoogart2006.pdf> (link).

⁶⁶ More parochially, the publication of law review articles has changed more in the last six years than it did in the previous fifty. The rise of the Social Science Research Network has made drafts of scholarly articles available—often long before publication. Law reviews have responded by finding new ways to add value. For example, major law reviews have added online components that sometimes comment on print articles. See, e.g., *Northwestern Law Review Colloquy*, <http://colloquy.law.northwestern.edu> (last visited Apr. 12, 2010) (link).

⁶⁷ See GLENN REYNOLDS, *AN ARMY OF DAVIDS: HOW MARKETS AND TECHNOLOGY EMPOWER ORDINARY PEOPLE TO BEAT BIG MEDIA, BIG GOVERNMENT AND OTHER GOLIATHS* 175–93 (2006).

⁶⁸ See Diana M. Bowman & Graeme A. Hodge, *A Small Matter of Regulation: An International Review of Nanotechnology Regulation*, 8 COLUM. SCI. & TECH. L. REV. 1, 3 (2007), <http://www.stlr.org/cite.cgi?volume=8&article=1> (link).

⁶⁹ See Albert C. Lin, *Size Matters: Regulating Nanotechnology*, 31 HARV. ENVTL. L. REV. 349, 355 (2007) (link).

and analyze that information all the more necessary.⁷⁰ Already, the military is having trouble analyzing all the information it is getting from its drones because it lacks sufficient sorting and analytic capacity.⁷¹ This problem is a metaphor for social decisionmaking as a whole. As accelerating technology creates new complexity more rapidly than ever before in areas, such as nanotechnology, biotechnology, and robotics, that were not even known a few decades before, social decisionmaking must struggle to keep up with analyzing the wealth of new phenomena no less than the military has struggled to process the ever more detailed information it receives from its modern technology. Societies prosper if they can use all of the information available to make the best decisions possible. The problem now is that information available to be processed may be swelling beyond human capacity to achieve sound social decisionmaking without the aid of AI.

One side benefit of greater capacity to process information may be the ability to better predict natural catastrophes and either 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 better such measures are likely to be.⁷² Moreover, by estimating the risks of various catastrophes, society is better able to use its limited resources to focus on the most serious ones.

The acceleration of technology can create unparalleled cascades of benefits as well as new risks of catastrophe. This acceleration could potentially endanger the future of the human race, but could also potentially radically extend the life span of individual humans. If such acceleration is the fundamental phenomenon of our age, the assessment of the consequences of technology is an essential task for society. As a result, the government has a particular interest in accelerating the one technology that may analyze the rest of technological acceleration—AI. The question of what degree and what form of support is warranted to boost the acceleration of this technology to help us with decisionmaking about the rest of accelerating is subtle and difficult. But that is the right question to ask, not whether we should retard its development with complex regulations, or still worse, relinquish it.

⁷⁰ See ATUL GAWANDE, *THE CHECKLIST MANIFESTO: HOW TO GET THINGS RIGHT* (2010) (drawing attention to problems created by the greater sea of information). Gawande's solution in the medical field is to create checklists focused on the most important protocols for saving lives. *Id.* at 137–38. While a checklist approach may indeed be the best solution to information overload given present technology, one general problem with a checklist is that it creates a one-size-fits-all approach that is frozen in time. In contrast, AI would allow one to decide what is most important for a particular patient, taking up-to-date information into account.

⁷¹ See Christopher Drew, *Military is Awash in Data from Drones*, NYTIMES.COM, Jan. 10, 2010, <http://www.nytimes.com/2010/01/11/business/11drone.html> (link).

⁷² See FLORIN DIACU, *MEGADISASTERS: THE SCIENCE OF PREDICTING THE NEXT CATASTROPHE* (2010).