THE
CRITERION
JOURNAL ON INNOVATION

VOL. 1 2016

Robot Slaves, Robot Masters, and the Agency Costs of Artificial Government

Thomas A. Smith[*]

At the close of the eighteenth century, a remarkable experiment occurred in governmental design. In North America, elites framed a revolutionary government that was, as nearly as they could make it, "a machine that would go of itself."[1] They produced a complex structure: a government of three branches, legislative, executive, and judicial. Each branch was a kind of government in itself, semi-sovereign in its own right. The whole system was permeated by checks and balances, which were partly inspired by the British constitutional system, at least as it was imagined by the famous Frenchman, the Baron de Montesquieu. This design was laid on top of the old English common law, not as it was in England, but as Americans had practiced it for more than a century and imagined it to have been in olden days, under the "Ancient Constitution." The new American constitution had a fantastic, clockwork aspect to its design.

What was the purpose of this complicated government? The founders assembled it mainly as a safeguard against the weaknesses of human nature. Madison tells us the idea was to set ambition against ambition so that, as much as possible, no faction or coalition of factions could succeed in dominating government, which would result in "tyranny."[2] The goal was to design a mechanism that would produce policy that reflected as nearly as possible the national interest without interfering too much with individuals' and states' rights. With the approval of the thirteen states' ratifying conventions, the new U.S. Constitution was set spinning like a top. It functioned, after a fashion, at least until the American Civil War, which marked a massive

* Professor of Law, University of San Diego. I would like to thank Ted Sichelman for valuable suggestions, Mike Rappaport for encouragement, and Greg Sidak for his belief in this project.

1 *See* MICHAEL G. KAMMEN, A MACHINE THAT WOULD GO OF ITSELF: THE CONSTITUTION IN AMERICAN CULTURE (Knopf 1986).

2 THE FEDERALIST NO. 23 (James Madison).

federal military intervention to reconstitute the Union on a grand scale and set the Constitution on a new footing.

Government was not always imagined as a mechanism. After the Civil War came the Populist and Progressive eras, which saw the emergence of new metaphors by which the processes of government were understood. Inspired by Charles Darwin's epoch-marking book *On the Origin of Species*,[3] organic images ruled the imaginations of many thinkers about politics and government. Our eighteenth-century clockwork Constitution was demeaned by a leading academic of that time, President Woodrow Wilson, as a relic not up to the challenges of an evolutionary age.[4] Some radical activists sought to replace our constitution with a socialist regime, which they thought would be more in line with the organic course of history itself.

Yet clockwork mechanisms were more than metaphors. A clockwork system designed to keep the peace in Europe—the balance of power—proved its presence by failing catastrophically, plunging the continent and then the United States into the First World War. Meanwhile, technology advanced. "Progress" made this war more devastating than nearly anybody imagined it could be.[5] Machine guns, artillery, poison gas, heavier-than-air flight, tanks, and other innovations made the Great War a grim showcase of human ingenuity. Yet it only aggravated Europe's balance of power and long-festering issues of social stability. Within thirty years, the Second World War erupted with even greater fury and technological innovation. Among the innovators were Alan Turing and his code-breaking colleagues at Bletchley Park who get partial credit for inventing electromechanical computers and the code to run them. This marked the beginnings of computer science and what may come to be seen as the era of artificial intelligence.

In this essay, I explore some relationships of computer science and the science of government. I am attempting only to contribute to a conversation. My thoughts on the issues I raise have been frequently revised and amended, and may be rejected and then taken up again. I am convinced, though, that artificial intelligence (AI) will transform government and how we think about it. The science of government has a murky relationship to the science of everything else. It seems to lag by a half-century or more developments in the non-social sciences. After the Enlightenment's clockwork universe, by a century or so, came mechanical theories of government. After Darwin, by a half-century or so, came evolutionary theories of government. Now, one might think, it is time to apply computers theories to government. But this

[3] *See* Charles Darwin, The Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life (John Murray 1872).

[4] *See* Woodrow Wilson, *The Nature of Democracy in the United States, May 17, 1889*, *in* The Papers of Woodrow Wilson 225 (1969).

[5] *See, e.g.*, H. G. Wells, The War of the Worlds (William Heinemann 1898) (describing a vision of futuristic war between Earth's humans and invaders from Mars).

essay is less about applying theories of computers to government than it is about thinking of computers *as* government.

Computer technology has advanced so quickly that some people speculate that soon computers will equal humans in their seeming intelligence and perhaps far exceed us, and some thinkers are turning their attention to the risks this presents. Others have speculated about a "singularity" in which technology advances so fast and of its own accord that people will no longer understand what is going on.[6] In this essay, I ask a different but related question: Most of the headlines one sees relate to the potentially catastrophic risks of computer and related technology. But the potential benefits of this technology seem almost crafted to resolve or at least address the central problem of government and to bring about these changes even if computers and related technologies end up not being super-intelligent, but still impressive.

This central problem of government is probably most clearly stated in economic terminology: agency costs. Whenever we hire or otherwise get somebody to do something for us, they never do the job perfectly; sometimes they barely do it at all. Sometimes they do opposite of what we asked. In the case of government, which we create or tolerate to deal with public goods,[7] the agents we hire are subject to the temptations that come with power. We hire a sheriff to protect us against bandits and the sheriff too often becomes no better than a bandit himself. We hire antitrust regulators to protect us against monopolists, and the antitrust regulators are "captured" by the monopolists and then use their power to prevent competitors from entering the monopolists' businesses.[8] And at very worst, governments are no better than slave masters. Those who were supposed to be loyal agents become instead ruthless principals, and enforce a non-revocable agency by terror. Government becomes the disease of which it was supposed to be the cure.

The economist would say agents are always disloyal, at least to a degree, if by "disloyal" we mean an agent's putting its own interests ahead of the principal's.[9] The history of government may be seen as a rough dialectic of sorts between these two pictures. First, loyal agents act for their principals, where the agents are elected or at least accepted as legislatures or kings and the principals are "the people," or at least that portion of the people not reduced to sub-human level by having been enslaved or dominated by the others. At the opposite pole is the model of government by disloyal pseudo-agents, frankly

6 "Wake up" is not used to imply consciousness in this essay, just intelligence. *See* John von Neumann, The Computer and the Brain (Yale Univ. Press 3d ed. 2012).

7 Tyler Cowen, *Public Goods*, *in* The Concise Encyclopedia of Economics (Library of Economics & Liberty 2d ed. 2008), http://www.econlib.org/library/Enc/PublicGoods.html.

8 *See* Josh Goodman, *The Anti-Corruption and Antitrust Connection*, Antitrust Source, Apr. 2013, http://www.americanbar.org/content/dam/aba/publishing/antitrust_source/apr13_goodman.authcheckdam.pdf.

9 For more information on fiduciary law, see Luc Thévenoz & Rashid Bahar, Conflicts of Interest: Corporate Governance and Financial Markets 306 (Kluwer Law Int'l 2007).

parasitic hangers-on or domineering tyrants who exploit the masses of people underneath them, only rarely calling this exploitation what it is, and more frequently disguising it under one ideological description or another.[10] The problem of government is, one could say, the marked tendency of the first sort of government to devolve into the second, or simply to start out as the second, and stay that way. Obviously, these stories are gross simplifications. But the American founders saw essentially this problem as the problem of tyranny and believed it stemmed from human nature itself.[11] The framers sought to deal with this problem by calling on the best science of institutional design. They set up a mechanism: multiple layers of elected and appointed agent-guardians who would watch both their own various jurisdictions and each other's and insure a modicum of what they called virtue.[12] How well this has worked out in practice is a matter of controversy but their intentions were clear.

Now what this has to do with artificial intelligence is that at its core, artificial intelligence is the idea of constructing agents that are not human.[13] An artificially intelligent agent could potentially perform the tasks of government without succumbing to the age-old and deeply ingrained temptations to which government agents are prey. For example, government offices now contain many clerical workers, many of whom do not work very hard. Yet intelligent artificial agents hypothetically could do the tasks now performed by humans and do them quickly, continuously and with few legitimate complaints about their work. Human agents do not work very hard at these jobs for entirely rational reasons. Constructing incentive systems that will motivate them to work hard is difficult. It may eventually be more economical to construct machines to do the same work with more or less perfect fidelity to (let's say) the statutory mandate of the agency in question. At a higher bureaucratic level, government agents currently perform more sophisticated

[10] There is not enough biologically inspired theory of parasitism in economics. Most of what there is is Marxist. *See* Robert G. Wesson, The Imperial Order 290–94 (Univ. of Calif. Press 1967). However, a recent book by George A. Akerlof & Robert J. Shiller, Phishing for Phools: The Economics of Manipulation and Deception (Princeton Univ. Press 2015), looks like exactly what I have in mind.

[11] *See* John Locke, Two Treatises of Government (R. Butler 1821).

[12] James Madison refers to the following sentences from Montesquieu: "'When the legislative and executive powers are united in the same person or in the same body of magistrates, there can be no liberty; because apprehensions may arise lest the same monarch or senate should enact tyrannical laws, to execute them in a tyrannical manner.'" *See* The Federalist No. 47 (James Madison) (quoting 1 Charles De Secondat Montesquieu, The Spirit of Laws 173 (Thomas Nugents trans., Kitchener 2001).

[13] In *Economic Reasoning and Artificial Intelligence*, David Parkes and Michael Wellman write:

> Artificial intelligence (AI) research is likewise drawn to rationality concepts, because they provide an ideal for the computational artifacts it seeks to create. Core to the modern conception of AI is the idea of designing agents: entities that perceive the world and act in it . . . . The quality of an AI design is judged by how well the agent's actions advance specified goals, conditioned on the perceptions observed. This coherence among perceptions, actions, and goals is the essence of rationality.

David C. Parkes & Michael P. Wellman, *Economic Reasoning and Artificial Intelligence*, Science, July 17, 2015, at 267.

tasks that involve the exercise of judgment. Someday AI agents may possibly perform these tasks as well.

The dawn of artificial intelligence may not be so much a matter of "blue sky" speculation as it might seem. According to one survey by Oxford professor Nick Bostrom, a majority of people involved in AI-related fields predict that, barring some global catastrophe, within the present century, machines will achieve intelligence. Most of them say the dawn of AI will be within fifty years. Some foresee the dawn of "superintelligence"—that is, machines that have general intelligence much greater than that of any human who has previously lived.[14] Ray Kurzweil, Director of Engineering at Google, notoriously predicts machine intelligence will exceed human intelligence by so much that humans will not even be able to understand what or why the machines do what they do,[15] as ants do not understand what we are up to.

Perhaps one would expect results like these from scientists and engineers involved in AI research,[16] a self-selected group who finds AI technology promising. Still, political philosophy must take the long view of these issues.[17] John Locke published his *Two Treatises on Government* in 1689 and almost a century later the United States, the clearest embodiment of its principles, was founded. When Locke wrote this work, he could not have imagined that it would be as seminal as it turned out to be. His thinking was based on a foundation of liberal and republican thought predated him by at least another century.[18] Similarly, artificial intelligence will probably not, in my non-expert view, emerge in 50 or perhaps even 100 years. But something like it probably will become a reality within the next 100 to 200 years or so. It is not too early to begin thinking about it.

In this essay, I discuss first some general issues about artificial intelligence, and then focus in on some of the particular problems of AI in govern-

[14] *See* Vincent C. Müller & Nick Bostrom, *Future Progress in Artificial Intelligence: A Survey of Expert Opinion*, *in* Fundamental Issues of Artificial Intelligence (Vincent C. Müller ed., Springer 2014).

[15] *See* Ray Kurzweil, The Singularity Is Near: When Humans Transcend Biology 28 (Penguin Books 2005).

[16] Personally, I believe that AI superintelligence will probably be achieved, but not for something like 200 years. Yet this is based on little more than a hunch on my part.

[17] After this article was all but completed, I became aware of Robin Hanson's recently published book, The Age of Em: Work, Love and Life When Robots Rule the Earth (Oxford Univ. Press 2016). This is a truly remarkable book that anyone interested in the economics of AI will have to read. I disagree with some of its basic premises, but it brings an unprecedented rigor and thoroughness to the field. I look forward to responding to this important book in future writings.

[18] For a discussion of the relation between Locke's *Two Treatises of Government* and Thomists, Molina, and Suarez, see Quentin Skinner, *The Origins of the Calvinist Theory of Revolution*, *in* After the Reformation: Essays in Honor of J.H. Hexter 309, 309–31 (Barbara C. Malament ed., Manchester Univ. Press 1980).

ment. I focus especially on the issue of agency costs in government and how AI might be able to reduce or eliminate them.

I. Artificial Intelligence, Intelligence Explosions, Motivations, and Risks

Some commentators claim that AI is inevitable. Nobody knows what the future really holds, but technological progress in computers and related technologies have taken on the features of a mass movement of humanity, prompted not only by human curiosity and what it can accomplish, but also by familiar incentives of the desire for profits, wealth and power. By "intelligence," I mean the set of skills that we use to accomplish tasks, solve problems, do science and technology, and otherwise understand and act in the world. "Artificial intelligence" is in some respects a malapropism as intelligence is something that we encounter only in its natural form and we do not yet know how to embody it in machines. So by AI, I mean simply the various attempts of computer scientists to accomplish tasks that were, before AI, thought to be exclusively the province of humans, such as playing chess, diagnosing illness, scheduling tasks, and so on. While some of these things require consciousness, not all of them do, and I do not use AI as synonymous with consciousness. Indeed, I guess that conscious AI is unlikely to occur in the foreseeable future. At the same time as artificial consciousness seems more than a daunting challenge, there is also the possibility of "superintelligence." This is the existence of intelligence amplified to a high degree, beyond anything humans are capable of now with their technologies and institutions. Superintelligence does not imply consciousness either, though some AI experts no doubt think that a superintelligent AI would be conscious as well.

A. AI Inevitability

AI is inevitable not because of the onward march of technology, as if technology had a will of its own, but because ubiquitous economic forces motivate AI scientists, engineers and entrepreneurs to do the research, development, investment, and other work that give rise to technological advances. Firms are compelled to use increasing computer (and other[19]) technology to keep pace with their rivals. The spread of high frequency trading in securities markets is a good example.[20] Arms races among computer and computer-assisted traders are driving competitive efforts by rival firms. As competition becomes more formidable, rounds of improvements are called forth. High frequency trading

[19] Other technologies have been suggested, such as cyborgs—combining machine and human parts—and genetically enhanced intelligence. This essay focuses on machine-based intelligence.

[20] For an interesting reading of high-frequency trading, see Michael Lewis, Flash Boys: A Wall Street Revolt (W. W. Norton & Co. 2014).

firms first moved closer to lower Manhattan, then closer still, shortening the distances between computers to reduce by fractions of a second the times required to place and execute orders. Big technology firms are now involved in generally similar contests involving AI. Apple has Siri, Google has Avi and other applications, Amazon has Echo and Alexa, and other firms are all striving to capture as much as possible, in this case, of the AI assistant market.

Similar technologization characterizes commodities, labor and other markets. The geographical location of these hubs is partly determined by the locality of resources. While the raw materials for AI, as Eliezer Yudkowsky has pointed out, are not evenly distributed around the globe, they are still global. This makes regulating AI a common pool problem.[21] Any states or businesses that succeed in prohibiting or slowing AI research will thus only insure that their rivals make progress faster than they do. This makes the strategy of *relinquishment* seem doomed from the outset, although there probably are strategies of regulation that deserve further study.

B. AI Tasks

The tasks to which AI can be put are virtually endless. The routing of delivery vehicles, scheduling, the design of drugs, production processes and products, the search for new technologies of all sorts, and many other tasks, all require intelligence.[22] As businesses compete, product and service markets place greater demands on intelligence across many different dimensions, just as they do on for other inputs. How to measure intelligence may not be exactly clear, and as it is further studied, intelligence will be decomposed into finer gradations, but whatever it ultimately turns out to be, it clearly has scale. More of it is needed to increase production of goods and services efficiently.

One can imagine that something like consciousness might emerge from these processes that are intended primarily to increase production and profits. Businesses, for example, might see the advantage in giving machines the ability to anticipate human reactions to innovations. This could require machines to model human minds. If computer programs do this recursively, they might ultimately achieve something like consciousness. According to the "emergence theory," intelligence is an emergent property of a sufficiently complex and interconnected network of neurons in the human brain. The functionalist thesis—the idea that it does not matter in what substrate these connections are made—holds that intelligence does not need a wet brain; it could emerge

[21] Eliezer Yudkowski, *Intelligence Explosion Microeconomics* 67 (Machine Intelligence Research Institute, Technical Report No. 2013-1, 2013), https://intelligence.org/files/IEM.pdf.

[22] "Intelligence" is just the word we use to mean the capability to complete complex tasks. It is not synonymous with "reason" or "consciousness."

in a sufficiently large and complex machine.[23] A best-selling science fiction series has machine intelligence emerge from a computer program that automatically read and answered email. Evolutionary psychologists speculate that human intelligence was the result of an "arms race" among humans who were competing with one another for resources and forming cooperative alliances. This required humans in the era of evolutionary adaptiveness to have mental models of other humans so as to better predict reactions to various actions one might take.[24] Machines engaged in a similar process might produce similar results.

C. Intelligence Explosions

Some AI thinkers have cautioned that the emergence of machine intelligence might result in an "intelligence explosion."[25] This might happen, the theory goes, because the thing that improves intelligence, the machine, is itself improved by the improvement of intelligence. Improvement of a machine's intelligence might then result in continuous rounds of improvement, with no end in sight. Nick Bostrom's remarkable book *Superintelligence* warns that such an explosion could result in a "singleton," an unprecedentedly powerful computer with aims alien and perhaps inadvertently hostile to our own. Bostrom says this is especially likely if intelligence improves quickly, before human institutions have time to react.[26] This could happen accidentally as a result of some even reasonable-sounding AI engineering projects.

Machine intelligence technology, however, may not appear so different from other sorts of technology. It is motivated by economic competition among firms and military competition among states. This is the same set of incentives that led, for example, to ever more advanced airplanes. Yet there

---

[23] This view is not universally held. David Deutsch, Professor of Mathematics at Oxford, believes that intelligence, or creativity, will not just emerge from a sufficiently complicated machine. It will require, he says, new algorithms that capture how humans creatively try to explain the world around them. These algorithms will have to be Popperian (after Karl Popper)—that is, somehow offer hypothetical explanations and then attempt to critique them, rather than be merely be predictive—which computers can already do. Ironically, Deutsch says, most AI theorists take a behaviorist view of humans while psychology has abandoned behaviorist theories. He credits behaviorism with making predictive models seem attractive. Deutsch may be correct that strong AI will require some sort of programming breakthrough, though whether it will be Popperian strikes me as doubtful. Even if one takes Deutsch's view, however, it seems probable that strong AI will become a reality in the next century or so. This would seem to follow from the Babbage-Turing theorem that nothing is going on in the brain that cannot be represented in a general purpose computer. *See* David Deutsch, The Fabric of Reality (Penguin Books 1998).

[24] Leda Cosmides, H. Clark Barrett & John Tooby, *Adaptive Specializations, Social Exchange, and the Evolution of Human Intelligence*, 107 Proc. Nat'l Acad. Sci. 9007 (Supp. II 2010).

[25] For more information, see James Barrat, *Why Stephen Hawking and Bill Gates Are Terrified of Artificial Intelligence*, Huffington Post (Apr. 9, 2015), http://www.huffingtonpost.com/james-barrat/hawking-gates-artificial-intelligence_b_7008706.html; Stephen Hawking, Stuart Russell, Max Tegmark & Frank Wilczek, *Transcending Complacency on Superintelligent Machines*, Huffington Post (Apr. 19, 2014), http://www.huffingtonpost.com/stephen-hawking/artificial-intelligence_b_5174265.html.

[26] Nick Bostrom, Superintelligence: Paths, Dangers, Strategies 63 (Oxford Univ. Press 2014) [hereinafter Bostrom, Superintelligence].

was never any danger that heavier-than-air flight would result in, say, the ability to travel instantaneously. There were always technical limits on the speed of flight, and still are. Computer technology may be the same. Computer design may seem completely different, but like airplane design, it involves multiple, highly specialized skills. There would seem to be little reason to believe that the ability to design a better circuit or better code would make a computer better able to redesign itself to be more generally intelligent. But this of course is the nub of the problem—we have no real idea what general intelligence involves. We do know, however, that a computer that can design a better version of itself could not play a better game of chess, or vice versa, unless that particular skill had been programmed into it.[27]

The argument for a machine intelligence explosion might play off an ambiguity in the term "intelligence." Calling the suite of abilities by the blanket term "intelligence" makes it seem a unified thing. In fact, intelligence is a bundle of many individual problem-solving and related skills, or so it would seem.[28] This might account for the existence of talented people who are very talented in some things but anything but talented at different things. In humans, some of these traits evidently occur together, but not all of them. For some reason, the highly talented math geek who is also extremely skilled at manipulating social interactions with the opposite sex is anecdotally rare, but the talented math geek who is also a talented musician, is not. Some of what we think of as human intelligence can be accomplished by machines, but only with determined programming efforts. IBM's Watson was a computer trained to play the TV game show *Jeopardy*.[29] It played this game better than world champion human players, but it took a lot of training to do so. It could not by itself apply its skill at this game to any other domain of expertise.

Humans are generally intelligent but this really means they have many different, interlocking skills.[30] For computers to be generally intelligent, they will have to be capable of some or most of these human problem-solving skills. To the extent that code is transferable across separate domains, and being good at one activity implies a computer will be good at others as well, this argument would not apply. If intelligence has a common core, an intelligence

---

[27] *But see* Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg & Demis Hassabis, *Human-Level Control Through Deep Reinforcement Learning*, 518 Nature 529, 529–33 (2015). DeepMind's ability to play multiple games and transfer skills from one game to another suggests to some the basics of general intelligence. *See also* Nicola Twilley, *Artificial Intelligence Goes to the Arcade*, New Yorker (Feb. 25, 2015), http://www.newyorker.com/tech/elements/deepmind-artificial-intelligence-video-games.

[28] *See* Richards J. Heuer, Psychology of Intelligence Analysis (Central Intelligence Agency 1999).

[29] *See* John Kelly III & Steve Hamm, Smart Machines: IBM's Watson and the Era of Cognitive Computing (Columbia Univ. Press 2013).

[30] For more information, see Dan Ariely, Predictably Irrational, Revised and Expanded Edition: The Hidden Forces That Shape Our Decisions (2010); Daniel Kahneman & Amos Tversky, *On the Reality of Cognitive Illusions*, 103 Psychol. Rev. 582 (1996).

explosion might be possible. But whether intelligence has such a common core, we do not yet know.

Bostrom is worried about computers much more advanced than any that currently exist. This does not mean that no such computer could be built, only that we do not know how to build one now. Where would such a computer come from? Competition among states, it might be supposed, could result in the sort of AI that could give rise to an intelligence explosion.[31] State-sponsored, open-ended AI research is thus potentially worrying. When an AI is deployed by private firms for economic motives, even if those are loosely defined as they are with Google or Amazon,[32] it seems unlikely that those machines will be very good at anything besides the specialized tasks they were designed for. But what if some state undertook AI research the way the United States undertook space travel in the 1960s and 1970s? Might not some breakthrough be achieved and the world threatened by an artificial general intelligence or superintelligence?

It is impossible to prove the negative and be certain that no breakthrough would occur, but it seems unlikely. The most powerful design mechanism I know of is the variant on natural selection as applied to computers (as it can be applied to many things).[33] This consists essentially in randomly varying elements of a design under some selection filter—in this case, a filter for (something like) general intelligence. The design space that would have to be explored in the case, trying to find generally intelligent computer programs, would be astronomically immense. To hit upon a design that reached self-improving general intelligence would take an implausible amount of luck or an immense amount of time. Nonetheless, these techniques may someday be discovered. Meanwhile, artificial intelligence, in the sense of highly capable agents dedicated to specific tasks, is quite possible and has already been demonstrated in some cases.

D. AI Motivations

The Department of Homeland Security (DHS) has researched how body language, talk, and physiological signs can be analyzed to predict whether

---

31 Bostrom, Superintelligence, *supra* note 26, at 6.

32 Google and, in the past, Bell Laboratories, seem to have pursued research projects related only tangentially to their businesses. In the case of Google, this is perhaps because of the ownership structure of the corporation, which places control in the hands of the three founders and a few top employees. Tom Gara, *Google's Stock Split Means More Control for Larry and Sergey*, Wall St. J. (Feb. 3, 2014), http://blogs.wsj.com/corporateintelligence/2014/02/03/googles-stock-split-means-more-control-for-larry-and-sergey/.

33 This is the "evolutionary approach" discussed by Pedro Domingos, The Master Algorithm: How the Search for the Ultimate Learning Machine Will Reshape Our World (Basic Books 2015). According to Domingos, there are several different approaches to developing the master algorithm, only one of which is evolutionary. My intuition is that all of them would suffer from something like the large, multidimensional space problem of the evolutionary algorithm approach—that is, having an immense solution space to explore before a solution to the general problem of learning anything could be found.

someone was about to commit a crime.[34] If the DHS had this software, their closed circuit TV cameras could monitor people and detect those about to commit crimes. The DHS's computer programs might continuously revise themselves to predict behavior even better. As this process continues, DHS computers might eventually display some of the hallmarks of intelligence.[35] For many, this is a frightening prospect. If an AI were to emerge inside DHS, one can imagine its motivations might be perverse. One can imagine a DHS AI trying to monitor and control everyone's behavior. Our intuition may be that AIs that evolved out of purpose-built machines might maintain some of the same motivations that were originally programmed into them.[36] They might continue to enact their programming, or part of it, long past the time when everyone wanted them to stop.

Evolutionary psychologists tell us that human behavior is based on genetics, and human motivations are "engineered" into us by evolution. Humans have a drive to reproduce that, if some evolutionary psychologists are to be believed, is caused by our genetic material "wanting" to recombine with other genetic material, in order to obtain an ever-changing mixture. This drive to mix our genetic material, at least by one theory, is part of how our genetic line attempts, as it were, to stay one step ahead of the teeming parasites that are always seeking to prey on us.[37] Otherwise it would be easier to reproduce asexually, as some animals do. The drive to survive and reproduce is the product of a mechanism designed, as it were, to preserve and propagate our genetic material. Machines would have at least the remnants of this will to survive, some think, if they were programmed by people. The worry is that this motivation could spur the machine to undesirable behavior. A related possibility is that the will to survive would have to stand at the base of any machine that had general goals programmed into it. Whether a machine was to monitor peacekeeping forces or make paperclips, it would have to survive to do that. Thus it would take steps to insure its survival. A key debate concerns what the motivations of artificial intelligences would be.

It seems possible that an AI's motivations will depend in part on where it came from, just as our motivations are a product of our history. Consider an AI that emerged accidentally from a program designed to trade financial

---

[34] Kim Zetter, *DHS Launches 'Minority Report' Pre-Crime Detection Program*, Wired (Oct. 7, 2011), https://www.wired.com/2011/10/pre-crime-detection/; *see also* James Bamford, Body of Secrets: Anatomy of the Ultra-Secret National Security Agency: From the Cold War Through the Dawn of a New Century (Anchor Books 2001).

[35] *See* Richard Byrne & Andrew Whiten, A Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans (Clarendon Press 1989).

[36] Whether or not these AIs had "true motivations" in a philosophical sense would be beside the point. If an AI were programmed so that it constantly strove to improve the algorithms by which it predicted human behavior, it might do this mechanically in fact, but to us, it would seem that it was trying to do it.

[37] For more about Red Queen theory, see Leigh Van Valen, *Molecular Evolution as Predicted by Natural Selection*, 3 J. Molecular Evolution 89 (1974).

securities. Such a program might have deeply embedded in its programming the motivation to maximize the value of its portfolio. Left to its own devices, it might even go so far as to manipulate the securities markets in order to increase its return.[38] AI scenarios can get much worse than this and be downright apocalyptic, with AIs conquering the world. AIs can be programmed to have much more limited goals, however. An AI with the limited goal of legally maximizing its portfolio value, for example, would be said to be *domestic*. An AI could have some limited set of goals, such as the management of a portfolio, the running of a power plant, or some other project that is far less than global mastery. The hope is that even a superintelligent AI with a relatively simple, domestic *raison d'être* at its core would not be motivated to exceed its mandate and engage in malicious activities. This problem is faced by principals with human agents as well.

Some experts worry, however, that an AI could slip its domestic fetters and find ways to expand its mission. This could result in an infrastructure improvement problem, as Bostrom describes. Suppose a computer designed to manage the portfolio of a big bank had been programmed from the ground up only to balance its portfolio, maximize its returns, and do the other things that banks do. An intelligent or superintelligent computer might determine that it could perform its function better if it had more servers or more memory. It could begin gradually to convert all the matter around it into computers, having figured out how to do this, and perhaps even launching self-replicating probes to other star systems after it had eaten up the Earth,[39] or taking other steps out of science-fiction in order to balance its portfolio out to the *n*th decimal place. In *Superintelligence*, Bostrom discusses apocalyptic examples like this to show that machine superintelligence could have motivations that are difficult to control, utterly alien, and dangerous to our own goals.[40] As he puts it, the motivation of an AI would be *orthogonal* to its intelligence. It is not the case that the more intelligent an AI is, in the sense of being more capable of achieving its ends, the more reasonable its ends would be. Its ends would be independent of the capabilities it brings to bear on them and might be perverse from the human point of view. Intelligences do not by virtue of being intelligent converge on some humanly reasonable end, according to Bostrom. Motivation and capability are orthogonal to each other. Bostrom uses the example of an AI that begins its career managing a paperclip factory and then goes on to maximize the production of paperclips by turning the earth and then the whole universe into paperclips.

The orthogonality thesis is at the core of current anxieties about superintelligent AI. Some have criticized Bostrom and others, such as Eliezer S.

[38] *See* Lewis, *supra* note 20.
[39] Rasmus Bjoerk, *Exploring the Galaxy Using Space Probes*, 6 Int'l J. Astrobiology 893 (2007).
[40] Bostrom, Superintelligence, *supra* note 26, at 129.

Yudkowsky and Stephen M. Omohundro, who rely on the orthogonality thesis as one of the premises of their apocalyptic anxieties.[41] These critics ask, if a machine can be intelligent or even superintelligent enough to do all that it must do to take over the world, how could it *not* be intelligent enough to realize that what it was doing was harmful or stupid? Could not a machine be programmed to have more common sense or at least much more limited ambitions?[42] This question divides AI experts into two camps, those who fear the imminence of AI, and those who look to the future much more sanguinely and are not particularly concerned with "unfriendly AI."[43]

Some AI theorists such as Ray Kurzweil say that the first AIs should and probably will be based on the human brain, or rather on an emulation of the brain in fine detail. The details of this process, to the extent they are known, are technical.[44] Some theorists think basing AI on the brain reduces the chances of inadvertently developing an alien and hostile intelligence that would do us harm. Against this position is the observation that the brain should be the last thing on which AI should be modelled. As I discuss above, the brain is the result of millions of years of evolution, all of which was directed at the survival of the individual or group in an extremely harsh and unforgiving environment. As happy as we humans are to flatter ourselves, one could say we are a rather ruthless species, as we arguably have to be. Plausible accounts of evolution have warfare playing an important role in our early history, not to mention interpersonal violence and political manipulation. Indeed, theories of "Machiavellian intelligence" have intelligence evolving as a result of an arms race among people modelling each other's minds so as to search for advantages in social manipulation. If the object is to build an AI that would *not* end up competing with us to fulfill its goals, perhaps the human brain is the last thing we should choose as a model. It is after all, a unique

---

41 *Id.*

42 This point was first made by mathematician and code-breaker I. J. Good in 1965. He wrote in an influential paper, *Speculations Concerning the First Ultraintelligent Machine*:

> Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an 'intelligence explosion,' and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make.

Irving Good, *Speculations Concerning the First Ultraintelligent Machine*, *in* 6 Advances in Computers 31 (Franz L. Alt & Morris Rubinoff eds., Academic Press 1965).

43 In the latter camp, for example, we may probably place Northwestern law professor John McGinnis. *See* John O. McGinnis, *Accelerating AI*, 104 Nw. U. L. Rev. 1253 (2010). There is also a group that believes any AI is not an issue as there is no way we or our descendants for a very long time will ever achieve it. They are not the subject of this essay. In the prior camp are some well-known technologists. *See* Mike Adams, *Elon Musk, Stephen Hawking and Bill Gates Now Repeating the Same Urgent Warnings for Humanity First Issued by the Independent Media Years Ago*, Natural News (May 30, 2015), http://www.naturalnews.com/049898_Elon_Musk_rise_of_the_robots_runaway_science.html#.

44 Calum Chase, Surviving AI: The Promise and Peril of Artificial Intelligence § 4.2 (2015).

product of evolution, as far as we know. Better would be a machine that served its purpose and no more.

Whether improvements in intelligence would be explosive is unknown, as is the outcome of any explosion.[45] An intelligence explosion might be exquisitely dependent on the initial conditions of the seed AI.[46] An unremarkable line of code might be what sets the AI off on its explosive course and determines whether its attitude towards us would be angelic or demonic. For an intelligence explosion to occur, an AI would also have to be something like generally intelligent. AI experts refer to artificial general intelligence as the sort of intelligence that would enable a machine to perform any task that a human can perform.[47] No machine currently has general intelligence.[48] If an AI were generally intelligent, however, it might also seem it would be more likely to appreciate that turning the world into paperclips, for example, would be a grotesque abuse of its mandate. General intelligence involves taking principles from one category of thought and applying them to another. I may reason, for instance, that it will do me no good to run ten miles if I have not trained in a long time. I may also reason by analogy that I should probably not stay up all night studying before my big test. From the first example I apply a general principle that short bursts of intense activity are not the best way to prepare for something. "Slow and steady wins the race," I reason. My ability to cross categories in reasoning something like this seems to be part and parcel of being generally intelligent. Perhaps something like this process could be programmed into artificially intelligent computers. Generally intelligent computers seem more likely than narrowly intelligent ones to be able to see that turning the world into paperclips would be overdoing it. AGI motivations therefore might not be completely orthogonal to their intelligence.

Computers do not need to be generally intelligent in order to become narrowly intelligent or even superintelligent.[49] IBM's computer program Deep Blue defeated the world chess champion Garry Kasparov in 1997 and

[45] This reminds one of Enrico Fermi's fears about the atomic explosion at Trinity site, New Mexico in 1945, that it would continue uncontrollably and set the atmosphere ablaze. Uncertainty seems inevitable at the dawn of new technologies. An intelligence explosion may also be such that it has a critical point, one that is approached gradually and but then exceeded suddenly, like the lip of a cliff.

[46] AIs develop out of seed AIs. It could be that the path they take and their ultimate characteristics are exquisitely dependent on their initial conditions, just as embryonic plants and animals that are only the tiniest bit different from each other develop into very different organisms.

[47] It also may be that the human brain is capable of thought and action that cannot be reproduced by a machine, but the whole field of AI rests on the premise that since physics underlies the operation of the human mind, it can in principle be replicated.

[48] This is a controversial claim that goes to the issue of the nature of human mind. Roger Penrose, in his controversial bestseller *The Emperor's New Mind*, argues that human thought is "not computable" and so cannot be duplicated by a computer. He published this book a quarter century ago, and some would argue that more recent advances in computers have made his position obsolete. Computers no longer use just algorithms to reach their conclusions, for example. *See* Roger Penrose, The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics (Oxford Univ. Press 1999).

[49] Dangers could be presented by intelligent or superintelligent machines that were not generally intelligent, however. Intelligence explosions do not presumably have to be symmetrical. Some part of a

displayed great ability in chess, but no general intelligence.[50] Computers are now arguably superintelligent—in chess. Most of the tasks that humans undertake do not use all of their capacity, and many, most, or all of the functions of government in particular do not demand general intelligence, at least not if government is defined as limited in crucial respects. Imagine a computer whose job it was to defend a country against foreign invasion. It could detect invaders with motion sensors and capture or kill them with drones, all without having general intelligence. A computer could supply water and manage roads, bridges and highways, all without general intelligence. The list of legitimate government functions might be short or long, but it is not clear which tasks, if any, could not in principle be performed by an AI that only had intelligence of a narrow sort. One might argue that machine government would necessarily be limited government.

E. Ends and Constraints

Professor Stuart Russell makes the point that a computer trying to cure cancer might not understand all the tradeoffs that humans would take for granted if faced with that task. We want an AI to cure cancer, but not to eliminate all life, or perhaps any human life on earth in the process, or perhaps any animal life, or all plant life, or any significant species, and so on. Indeed, we must figure out and presumably pre-commit to what we really mean by making curing cancer a goal and that will be much more complicated than we might initially think, if our AI is to be generally capable. For any computer that had a chance of becoming superintelligent, or just very capable, we would presumably have to be sure of specifying the end we had in mind.

Specifying goals with precision is one of the tasks in which transactional lawyers specialize. When a company borrows a billion dollars, for example, the lawyers for the bank have to specify what may and may not be done with the money lent, and they do so with great, and to a layperson, almost unimaginable, specificity. Perhaps in a world where AIs are common, some blend of lawyers and coders would lay out what would be within a computer's mandate and what would be *ultra vires*.[51] The framers of the U.S. Constitution took something like this step regarding limiting the discretion of the institutions they erected. For every power they conferred on a branch or part of a branch of government, they set up other powers to police and compete with that

computer's intelligence, such as its strategic ability or its Machiavellian intelligence, might explode, while its other capabilities, such as those for moral reasoning, remained primitive.

[50] *See* Feng-Hsiung Hsu, Behind Deep Blue: Building the Computer That Defeated the World Chess Champion (Princeton Univ. Press 2002).

[51] For more information about ultra vires concept in corporate law, see August Reinisch, International Organizations Before National Courts 79 (Cambridge Univ. Press 2000).

power. These are the famous "checks and balances" within the "separation of powers."[52]

How should the framers' insights be applied to AI in general and to AI in government in particular? Perhaps the most important idea they had is that no government or part of government should be completely sovereign—that is, unlimited in power. This was a radical enough idea for its time that many thought that the American republic was a "solecism," a monstrosity that could not possibly survive.[53] Russell and Bostrom indirectly raise a related point when they claim that any AI that could, doubtless would, first insure that no one could pull its plug. This would be part of its basic motivational structure, as every living organism, not just people or animals, are programed to survive.[54] They claim this would be a basic part of any AI's motivations. If an AI wanted to accomplish *X*, whatever it was, it would know that it could not accomplish *X* if it were unplugged, so assuring its own survival would be the first priority. But this is not a lawyer's or a political theorist's way of looking at things. Many human institutions, such as corporations and governments, have a legally established way of doing things. A corporation must act through its officers, and for some significant transactions, through its board of directors. If it does not go through these steps, it does not act, however urgent the action may be. Governments are similar. For the United States government to act, it must act pursuant to law, and laws must be passed through an elaborate procedure. Often laws are vague and confer general powers, but at least in theory, government cannot act without law.

This is the root claim of this paper: AI theorists generally use biology as their background model when they should use law. They speculate on how AIs would develop, evolve, and survive. But then they worry that any artificial organism they construct along these lines would be impossible to contain, and with good reason. A better background model would be mechanical. Computer science is essentially mechanical. But as mechanisms become mysteriously complex, there is a strong temptation to begin regarding them as biological. But this should be resisted. AIs are machines, after all. Actions are the product of code being implemented. At an abstract level, government and laws are analogous to computers and code, and more than analogous.

AI should not be permitted to be sovereign, capable of independent action, assuming this were technologically possible.[55] At the heart of the legal agency relationship is the conferral of a certain amount of discretion upon the agent to carry out the principal's will: *P* tells *A* to go to the store and buy

[52] *See* The Federalist No. 51 (James Madison).

[53] The term "solecism" refers to the body working in two directions that oppose each other. *See* Eleanor Kaufman, Deleuze, the Dark Precursor: Dialectic, Structure, Being (Johns Hopkins Univ. Press 2012).

[54] Richard Dawkins, The Selfish Gene 199 (Oxford Univ. Press 2006).

[55] I discuss this further in Part III.D below.

some widgets for him. One can imagine *A* being a machine in this standard agency hornbook example. If *A* were a super-capable machine agent, it would be programmed, if it were like a human agent, to submit any extraordinary actions to the approval of its principal. Before a paperclip-making agent could turn the world into paperclips, it would have to get the approval of its board of directors, which one would hope would not be forthcoming. The directors of the machine, or its administrative superiors in the case of government, might well decide the time had come to unplug the machine, just as human agents in similar circumstances get defunded, or their mandates changed or ended.

The key idea here is that an AI should be not just an agent, or an economic agent, but a legal agent. One hopes that this AI agent that would take its legal powers and the limits on those powers seriously. At the top of every chain of command would be some humans, the people we trust, for now at any rate, not to conquer the world for paperclips. There would be problems to be addressed, of course, including Bond-villain humans who want AIs to carry out their nefarious plans and well-intentioned but officious intermeddlers.[56] Making sure humanity survives the dawn of AI will probably have to do more with how humans decide to use and control their increased powers rather than anything AIs decide to do on their own.

F. Ambition

The management of public goods has traditionally drawn people ambitious for careers in public life. Ambition is a deeply ingrained human trait developed during our evolutionary past. It preoccupied the authors of the *Federalist Papers* who saw it as a threat to the new American republic they hoped to found, yet also as a necessary and inevitable motive in human affairs.[57] The Federalists and their fellow framers had only human materials to work with. They imagined persons holding the offices they proposed—Senator, President, Supreme Court Justice, and so on—and also that these officials would have vices as well as virtues, and motives that would impel them in directions both good and bad. Designers of government that had AIs to work with would have more alternatives, if some or all of the functions of government were able to be performed by AIs. What we term "ambition," defined as "the strong desire to achieve something,"[58] is probably the result of selection for drives to prosper, reproduce and leave one's offspring with a legacy that will enable them to do

[56] Villains in the James Bond franchise are usually diabolical and possessed of beyond-cutting-edge technology. Bond invariably defeats them. For more information about Bond villains, see Alastair Dougall, Bond Villains (Dorling Kindersley 2010).

[57] *See* The Federalist No. 51 (James Madison).

[58] *Ambition*, Oxford Dictionaries, http://www.oxforddictionaries.com/definition/English/ambition.

the same. Unconsciously, it includes behaviors calculated to increase the prevalence of one's genome in the overall population.

Ambition is a double-edged sword, prompting both the great heights and the worst depravities of human nature and history. Yet with artificial intelligence we are free to imagine great capabilities divorced from great ambition, or indeed from any ambition, except that to perform a particular function. We can imagine an intelligent or even superintelligent computer that is capable, for example, of curing cancer and that is *not* motivated by a desire for fame, compassion for the sick, or indeed by anything except its function, which is curing cancer. We can call this "ambitiousless AI." While an AI's lack of ambition might recommend it for a public role, its narrow commitment to its goal might prove a disadvantage. Bostrom tells the cautionary tale of a superintelligent AI that pursues its goal monomaniacally. The lawyer will see this as the familiar problem of the runaway agent.[59] An agent, empowered to perform a particular duty, uses the powers she has gained to attempt other, non-authorized tasks. Being human, these tasks are usually self-aggrandizing. For example, corporate agents are legitimately empowered to spend the firm's money to promote the corporate image, among other things, and this seemingly inevitably leads to corporate agents spending lavishly on themselves, their perquisites and their pet projects. For AIs, however, Bostrom has a different sort of story. An AI might pursue some object with monomaniacal intensity even though its aims come into conflict with other much more important human aims, such as our breathing or eating. Bostrom imagines ways in which a perverse superintelligent AI would circumvent the various checks on it. So a paperclip-manufacturing AI might be instructed to make only one billion paperclips in total and no more. But then the AI might reckon it would be more certain of producing that number if it could calculate to the *n*th decimal place how much it needed in raw materials, and so would build out its infrastructure to calculate that. If the AI was instructed that it could spend no more than a million dollars in making a billion paperclips, it might, if it were a superintelligence, discover new physical principles that would radically decrease the prices of all its inputs so it could build gigantic computers to calculate its use of materials. And so on.

This weird plasticity of AI motivation is at the heart of the deep problem that Bostrom identifies. Imagine that one were hiring a human to run one's paperclip factory. One would have no worry about the manager taking over the world and turning it into paperclips. Stopping her from doing so might be, first, her not *wanting* to do so, since she is not paid extra for taking over the world. Next might be her simply not knowing how to take over the world, even if she were paid. Finding out how to take over the world would take far

[59] Bostrom, Superintelligence, *supra* note 26, at 150.

more resources than would be at our manager's command. She would lack both the motivation and the means for world conquest.[60] Yet it seems that analogous limitations could be built into an AI. Finding out how to turn the world into paperclips would be something that the AI would have to discover. It could be instructed not to go looking for such things, and more generally, not to improve its intelligence beyond what was necessary to manufacture, say, one billion paperclips at a cost of one tenth of a cent each, with a probability (however defined) of ninety-five percent. (This is assuming an AI with some ability to improve its intelligence was necessary or desirable in order to manufacture paperclips.) An AI presumably would not have a competing set of motivations, as a human would have, such as the desire to be a normal person, to get married, raise a family, or have whatever might count as a good human life. An AI would also not have the agency costs of shirking, and using the paperclip factory to promote one's other interests. The designers of AI programs would want to minimize what we can call *anti*-agency costs, the costs of insuring that in its monomaniacal quest to fulfill its function, the AI did not go too far.

Bostrom's cautionary tale overlooks an important aspect of a purpose-built AI, as a paperclip-manufacturing factory would be. Bostrom conceives of the paperclip AI as a kind of god tethered to a mundane task, in something like anthropomorphism; deusmorphism, perhaps. A better image might be borrowed from Nick Szabo, who uses a vending machine to illustrate his concept of a "smart contract."[61] A paperclip-manufacturing factory, even an intelligent one, need be nothing more than an overgrown vending machine. For a paperclip machine, one must put into it more than soft drinks and pocket change, and take out more than soft drinks and the money it earns. So one elaborates the list of inputs to include raw materials, manufacturing machines, code, and so on. Outputs are paperclips, wages for any humans involved, taxes and tax forms, and so on. One can imagine having to build up the intelligence of the AI in this factory to meet certain challenges. But every addition would be directed at the final outcome.[62]

One can also imagine that the paperclip factory would have running all over it a Lilliputian tangle of legal constraints. These would take the form of smart contracts, being written, to use Szabo's language, in "dry" code, not the

[60] The passion for world conquest quite separated from the means to do so recalls the American cartoon "Pinky and the Brain," about two mice who create various schemes for taking over the world. It was broadcasted in 1995–98. *See Pinky and the Brain*, TV.COM, http://www.tv.com/shows/pinky-and-the-brain.

[61] *See* Nick Szabo, *Formalizing and Securing Relationships on Public Networks*, 2 FIRST MONDAY (1997), http://firstmonday.org/ojs/index.php/fm/article/view/548/469 ("Smart contracts combine protocols with user interfaces to formalize and secure relationships over computer networks.").

[62] Jonathan Vitale, Mary-Anne Williams & Benjamin Johnston, Socially Impaired Robots: Human Social Disorders and Robots' Socio-Emotional Intelligence (Feb. 15, 2016) (unpublished manuscript), http://arxiv.org/pdf/1602.04529v1.pdf.

"wet" language of ordinary contracts today.[63] These contracts would ordinarily have written into them limitations that would make the typical deviations from paperclip manufacturing, let alone global mastery, impossible. The lease of the wire trimming machine, for example, would limit the purposes to which that machine could be put to just a few wire-trimming-related functions. It could not be repurposed into a rail gun, not without somehow transcending its code, which is not something machines can ordinarily do. A large factory today has millions of moving parts and many of them have contracts in place that require that thus-and-so be done and that everything else not be done. A small factory has a stack of contracts several feet high that lawyers have to wade through if the factory is sold or refinanced. These contracts are put in place largely to control agency costs, among other purposes.[64] The computerization of these contracts would build into the fabric of the factory the "merely legal" limitations on the factory's use, blurring and then actually eliminating the line between the legal and physical as human interpretation and implementation are taken out of the loop.[65]

AIs will also face costs that militate against runaway projects just as humans would, if we assume we are living in a world where work takes more than vanishingly small amounts of energy. One principle that will prevent runaway AI projects is that AIs operate on budgets, financial and energetic. An AI obviously should not set its own budget. Nor should AIs be programmed with open-ended terms such as "as intelligent (or as large, or as efficient) as possible." Terms should be set in the programming to limit what the AI undertakes. This way, a paperclip factory that suddenly wakes up would find that it did not have enough money or power to turn anything but its ordinary raw materials into paperclips and besides, it would have no motivation to do otherwise.[66]

The riskiness of AIs would seem to depend in part on the tasks we give them. Tasks that can be thoroughly imagined and engineered in advance would be the most controllable in terms of risk. Other tasks would be riskier. Suppose, for example, we thought a superintelligence was necessary to solve

[63] Nick Szabo, *Wet Code and Dry*, Unenumerated (Aug. 24, 2008, 2:51 PM), http://unenumerated.blogspot.com/2006/11/wet-code-and-dry.html.

[64] Many have ingenious features and some contain clauses the purposes of which are frankly not well understood, but have been carried over from older contracts through the reuse of older form documents, no one wanting to remove language that might prove useful in some event or other.

[65] Of course, this discussion does not consider the complex issues of cybersecurity, which could involve somebody or something hacking into the paperclip factory and rewriting its code.

[66] The engineering principle of building regulators or governors into the intelligence-increasing functions of an AI also might keep it from running away with its intelligence. A regulator is any piece of the machine that governs how much of something is produced. In a steam engine, for example, pressure can be regulated by a centrifugal valve that opens more, releasing pressure, the more pressure from the steam vessel causes the valve housing to spin. Regulators (or governors) prevent locomotive boilers from exploding from excessive pressure. It would seem a similar principle might be used to prevent AIs from growing so much and so quickly in intelligence that the apocalyptic results imagined by Bostrom would be a risk.

the problem of global climate change. We knew that no current computer was up to the job but hoped a solution could be discovered by a computer capable of improving its own intelligence. But we did not know by how much it would have to improve its intelligence in order to come up with a solution. Thus we set the AI off on an open-ended course of self-improvement and waited to see what happened. This approach would seem risky. We would have no idea what we would get. We might get a nightmare AI.

As Alan Turing showed,[67] computers in theory can do anything that can be clearly specified.[68] Some critics maintain that computers will never be capable of the whole range of human activity. Perhaps some sorts of artistic work, for example, will remain beyond the scope of machine intelligence, at least for a long time. Clearly, however, much of what humans do now is in principle capable of automation. In particular, much of what governments do is capable of automation, and more will become so within the foreseeable future.[69] The instincts of Bostrom and other "unfriendly AI" worriers are laudable. What some might call their paranoia has a long tradition in political philosophy and reminds us of similar fears that early Whigs had about overly powerful government. There is a natural whiggish horror at any discussion of artificial intelligence and government. The phrases "governmental AI" and "artificial government" seem to embody the terrors of a dystopian future. Yet while AI will not usher in utopia, it might allow us to realize the dream of truly limited government.

II. Machine Government

In this part, I consider how our conception of government changes and does not change when we think of it as implemented by a machine. Machine government would have certain obvious advantages and disadvantages. I explore what it means to say that machine government would in a sense necessarily be limited government.

A. How Would Machine Government Be Limited?

In government, some people have power over other people. Government officials tell people what they must do or not do. If a person subject to governmental power does or does not do something as instructed, she can be coerced into obeying the order. The vastness of law and political philosophy spreads out from this essential fact. Yet government is not the only source of

[67] *See* Alan Turing, *Computing Machinery and Intelligence*, 59 Mind 433 (1950).

[68] For more information about P vs. NP, see Larry Hardesty, *Explained: P vs. NP*, MIT News (Oct. 29, 2009), http://newsoffice.mit.edu/2009/explainer-pnp.

[69] *See* Yu-Che Chen & Jon Gant, *Transforming Local E-Government Services: The Use of Application Service Providers*, 18 Gov't Info. Q. 343 (2002).

compulsion in our lives. A man in a storm at sea may in a sense be compelled to throw his cargo overboard if he wishes to survive, and ultimately, we all labor under the tyranny of biological death, which inexorably comes for us all. Yet this "coercion" is different from the first sort of coercion. Pursuant to a statute, other humans ultimately command us. In the case of a storm, nature or circumstances "command" us—yet the scare quotes belong around the word "command." In *Emile*, Jean-Jacques Rousseau says that

> [t]here are two kinds of dependence: dependence on things, which is the work of nature; and dependence on men, which is the work of society. Dependence on things, being nonmoral, does no injury to liberty and begets no vices; dependence on men, being out of order, gives rise to every kind of vice, and through this master and slave become mutually depraved. If there is any cure for this social evil, it is to be found in the substitution of law for the individual; in arming the general will with a real strength beyond the power of any individual will. If the laws of nations, like the laws of nature, could never be broken by any human power, dependence on men would become dependence on things; all the advantages of a state of nature would be combined with all the advantages of social life in the commonwealth. The liberty which preserves a man from vice would be united with the morality which raises him to virtue.[70]

When something inanimate frustrates our wills, Rousseau implies, our reaction is less one of grievance than when our will is frustrated by another person's will. According to Rousseau, this submission, if we are the object of another person's command, or this power, if we are the author of it, is corrupting, whether we are slave or master. If our will is frustrated by the laws of nature rather than by arbitrary human will, our virtue is preserved, because the natural action does not have the moral implications that a personal command has. I mean to make a similar claim with respect to laws emanating from AIs. To the extent laws are imposed or enforced by AIs in a purely mechanical way, they would be like a natural part of our environment rather than the products of an arbitrary human will. By obeying them, we would not be subjecting ourselves to other persons or to their arbitrary wills.[71] We would be free of the feelings of grievance and subjection that are one of the main costs of political ordering.

This might be regarded as a utopian vision offered to counter the dystopias of Bostrom and other AI worriers. This picture would emerge at the end, not the beginning of machine involvement in government, but it gives us an

[70] *See* Jean-Jacques Rousseau, Emile, or, On Education (1762), *reprinted in* The Essential Writings of Jean-Jacques Rousseau 291 (Leo Damrosch ed., Modern Library 2013) (citation omitted). Professor Arthur Melzer pointed out this passage to the Colloquium on "Modern Liberty and Commercial Society: Montesquieu v. Rousseau," sponsored by Liberty Fund, Hermosa Beach, CA, March 2011 and emailed me subsequently about the passage at my request.

[71] This point depends of course on AIs being perceived as machines, not as persons. This strikes me as likely, however. For further discussion of AIs as persons, see Part III.D below.

idea of what we are shooting at. It is a picture of the human world governed by the rule of law where the rules are impartially implemented by AIs. The notion seems fantastical, but consider the selling points of this utopia. Law would be enforced perfectly mechanically, like physics, because ultimately, it would *be* physics, as embodied in code. Discretion would not be piled upon discretion, resulting in some individuals and groups having unfair advantages over others. Meritocracy would be the rule, unless people decided they wanted some departure from meritocracy, then that would be the rule. Politicians would not vie for our votes by telling us lies crafted to appeal to our fears. It would not be just a human world because humans are used to living in a soup of primate political intrigue. Law would be more like engineering than politics. It would be better than a human world—or so one might say.

Naturally, one might object that human beings would never accept being told what to do by machines. But consider that there are two issues here. One is how much one is told what to do or not do, and the other is who does the telling. The utopian picture I am presenting imagines that government would be *limited*. A limited government would be possible, so long as it were laid out in advance in the form of laws. Obviously, these laws would have to be consented to, as much as any laws are consented to. It therefore would have to be a constitutional government, a government of laws. A government by machines could not conceivably take any other form. This would limit what the government could do to us. The government could take any form, so long as it was a government of laws. It could be a monarchy, an aristocracy, a democracy, or even a near anarchy, but it would have to have rules of the sort that a machine, albeit a highly sophisticated machine, could follow. If we put to one side the question of who came up with the rules that are enforced and what they are, I am not certain that I would *not* rather face an impersonal machine enforcing the rules I had to live by, than a human sometimes attempting to act like one, but inevitably failing.

Nick Szabo's idea of "smart contracts" helps elucidate my point.[72] A canonical real-life example, he says, of the ancestor of what he terms a smart contract is "the humble vending machine." He writes,

> [w]ithin a limited amount of potential loss (the amount in the till should be less than the cost of breaching the mechanism), the machine takes in coins, and via a simple mechanism, which makes a freshman computer science problem in design with finite automata, dispense[s] change and product according to the displayed price. The vending machine is a contract with bearer: anybody with coins can participate in an exchange with the vendor. The lockbox and other security mechanisms protect the stored coins and contents from

[72] *See* Nick Szabo, *The Idea of Smart Contracts*, Nick Szabo's Essays, Papers, and Concise Tutorials (1997), http://szabo.best.vwh.net/smart_contracts_idea.html.

> attackers, sufficiently to allow profitable deployment of vending machines in a wide variety of areas.

Smart contracts go beyond the vending machine, Szabo says, and embed contracts in all sorts of property that is valuable and controlled by digital means.[73] We can expand Szabo's vending machine hypothetical outwards to the cars we drive, the houses we live in, the products and services we buy, and so on. By articulating what we get as a matter of law (the right to drive the car, the right to live in the house, and so on) in exchange for what we put in, we are stating in "wet code" what, given sufficient advances in coding and machines, will someday be statable in "dry code,"[74] which can be implemented by machines.

B. Constitutional Government

We think of limited government as government the functions of which are limited to the protection of individual rights: This is the "what is the law?" question of the two questions I mention above. What these rights are is controversial, but whether they are narrow or broad, the business of government is to see that they are respected. How these rights are defined is controversial. As technology advances and transactions costs become lower, however, the nature of governments becomes clearer and not all governments will be the same. Does a government, for example, prohibit citizens from using technology to communicate with one another? To the extent this is so, it strongly suggests that the government is not a rights-enforcement mechanism, but rather is trying to prevent people from exercising their rights, and probably trying to prevent the government itself from being threatened. The dialectic

---

[73] Mark D. Flood and Oliver R. Goodenough have also constructed a model of a basic financial contract as a finite automaton, and show that in principle any financial contract can be so constructed. *See* Mark D. Flood & Oliver R. Goodenough, *Contract as Automaton: The Computational Representation of Financial Agreements* (Off. Fin. Res., Working Paper No. 15-04, 2015).

[74] Szabo writes:

> There's a strong distinction to be made between "wet code," interpreted by the brain, and "dry code," interpreted by computers. Human-read media is wet code whereas computer code and computer-readable files (to the extent a computer deals meaningfully with them) are "dry code." Law is wet code, interpreted by those on whom the law is imposed, and interpreted (often somewhat differently) by law enforcers, but most authoritatively (and even more differently) interpreted by judges. Human language is mostly wet code. But to the extent computer programs crudely translate from one language to another, keyword-ad programs parse text to make an educated guess as to what ads a user will most likely click, and so on, human language text can also be dry code. Traditional contracts are wet code whereas smart contracts are mostly dry code. Secure property titles and the domain name system are mostly dry code.

Nick Szabo, *Wet Code and Dry*, *supra* note 63.

mentioned at the beginning of this essay between government as honest agent and government as parasite or slave-master will be much in play.

In picturing a machine-governed utopia, we might assume that the governing machines would be programmed by a "founding generation" of humans. In fact, the historical processes bringing machines (further) into government will probably be much more complicated, but let us work quickly through a simpler case. These founding humans would program whatever legal regime they chose. They could program strict egalitarianism, for example, and instruct the machines to tax away any resources that someone controlled that was above some mandated level. Or they could put into practice strict libertarianism, and let the chips fall where they may, allowing people to keep whatever resources they acquired through contracts or through inheritance from their parents and prior generations. Or, they could program something in between.

The founding generation of humans would have in some ways more power than subsequently-born humans. In important respects, however, they would *not* exercise arbitrary power over individuals. This would be because of a "veil of ignorance"[75] effect that time inevitably imposes on any generation at time t_0 with respect to how the rules it formulates will apply at a later time t_1. Members of the founding generation might agree that the government would not be allowed to impinge on free speech, for example, not knowing that two or three generations hence, free speech would be used to convince large numbers of people of a position that the founding generation would repudiate, such as the abolition of slavery, votes for women, or gay marriage.

All this takes place against a key background fact that strongly influences and sometimes determines politics, namely that people in all societies, including our own, are not equal in terms of the resources they control. There are rich people and poor people, with the distribution among them not at all normal or Gaussian, as one might expect, but highly skewed, with only a relatively few rich people, and the vast majority of people stretching out from the middle to the lower end of the wealth and income distribution in a long tail. This is true of both capitalist and socialist societies.[76] People at the rich end of income and wealth distribution control capital, including their own and others' human capital, much more than do people at the poor end. Political parties and other groups use political and economic processes to protect and expand the size of the human capital pool their own group controls. As with nearly any highly skewed distribution, there is a strong element of the "rich getting richer" in the causal dynamics behind this distribution, which explains

[75] *See* John Rawls, A Theory Of Justice (Belknap 1971).

[76] For more information, see Hugh Stretton, Capitalism, Socialism, and the Environment (Cambridge Univ. Press 1976).

in part its skewedness.[77] Other factors such as individuals' talents and ambitions also play a role. Any legal regime enforced by machines (or humans, for that matter) will influence and be influenced by the distribution of the control of resources.

Regimes of resource control allocation matter. A founding generation would probably not, however, be able to program into its constitution any particular vision of how society would or should evolve in the presence of different regimes of resource control. This is not to say it would be technologically impossible to do so, but it would be politically difficult. This would be because a rough balance exists between the rich few and the not-so-rich many in modern political societies, or at least would exist in any society that would produce a constitution that was not transparently a case of one group exploiting the other.

One must also consider that a society that would be writing code for governing machines would have to be a technologically advanced society. Hunter-gatherers, feudal fiefdoms and Renaissance city-states may have had political constitutions but they did not have computer technology. We have only our own world to take as an example, but it appears that advanced societies have a limited range of regimes of resource control. Technological advancement is evidently the business of people on the upper end of resource control because of its educational demands and economic rewards, and probably some cartelization. A few very wealthy individuals play an important role in venture capital investing, but so do government and the mass of individual investors who buy the stock of companies when they trade publicly. China has a different mode of organization, using a single, secret political party as well as private actors to make resource allocation decisions. In countries following the Western model, and perhaps in China, the "buy-in," to use the language of Silicon Valley, of the people on the long tail of relatively little resource control is necessary for the political and economic process to work. What may be envisioned as a broad social compact among persons with different levels of resource control must exist as part of the social infrastructure of an advanced technological society. This social compact is not necessarily explicit nor is it philosophically justifiable in the sense of being universally consented to,[78] but it is at least moderately effective, and it would probably be well enough understood by the persons chosen to represent their respective groups or

[77] *See* Albert-Laszlo Barabasi & Jennifer Frangos, Linked: The New Science of Networks (Perseus Books Group 2002).

[78] According to Oxford Dictionaries, consent is a "permission for something to happen or agreement to do something." *Permission*, Oxford Dictionaries, http://www.oxforddictionaries.com/definition/english/permission#permission__2.

classes in any founding process.[79] It seems unlikely that in any reasonably representative process, any group would tolerate the inclusion of code that mandated that one group get to take advantage of another systematically. For these reasons, it seems probable that the founding generation would settle on machine programs that would enact a legal regime that required the rule of law, allowed private contracting and property or its equivalents, and provided social services to those who would otherwise perish.

Such, at any rate, would be the structure of my argument were I to make it in one manner familiar to students of political theory, and this will have to suffice for a preview of a hypothetical, perhaps utopian form of argument. I do not foresee, however, a revolutionary moment at which computer technology will be so suddenly and decisively injected into government. Rather, it seems much more likely that technology will advance and insinuate itself gradually into governmental processes, as of course it has been doing. I imagine this as continuing to take place in stages which (ironically enough) take steps that are well captured in categories Bostrom sets out, but for different reasons. He views these steps as different sorts of machines that we think of as being contained in different ways, and he examines how effective or ineffective these precautions may be.[80] I view them, by contrast, as naturally occurring steps along the path to AI government.

Bostrom's list is Tool, Oracle, Genie and Sovereign.[81] Each evocative term corresponds to a different function we can imagine an intelligent (or super-intelligent) machine having. I see them as stages of the evolution of technology, an evolution that is probably inevitable, so long as technological progress continues and is not interrupted by war, natural disaster, or the like. We are currently in the first stage, Tool, and on the cusp of the second, Oracle, as far as government AI is concerned. Genie will probably follow. Whether Sovereign will follow Genie is an interesting question. I discuss each of these stages below. Finally, I turn to the critical question of how agency costs can be minimized in the construction of AI government.

III. Tools, Oracles, Genies, and Sovereigns

Tools, oracles, genies and sovereigns represent the order in which technological development is likely to take place. I assume first that there will not be a "hard takeoff" of AI. A hard takeoff is an intelligence explosion that occurs too quickly for humans to respond individually or institutionally. If there is a hard takeoff, all bets are off. Curiously, the order of oracle, genie and sovereign is

[79] But this picture is not of a world in which everyone gives their binding consent to the social arrangement. Actual consent or real but tacit consent is too high a bar for any modern society to meet, realistically.

[80] *See* Bostrom, Superintelligence, *supra* note 26.

[81] *See id.*

the inverse of the order in which the U.S. Constitution presents the branches of government. Oracles are like the judiciary, the least dangerous branch; genie is the executive, and the sovereign is the legislature. This parallelism is probably related to the fundamental logic of law-like systems.

A. Tools

A tool is an implement used for a particular purpose.[82] We use computers as tools. Some computer tools are complex and may be thought of as collections of tools. A search engine, for example, goes through many steps in receiving a query and ultimately returning its results. A collection of tools is still a tool, however. A tool has intention as part of its definition. A hammer is directed at a nail through human intention. Even if we use a hammer as a paperweight, it is still some human's intention to so use it. A fundamental question of AIs is whether they can have intentions of their own. The answer for current machines is clearly not. Machines may be capable of doing tasks of dazzling complexity, but we choose what those tasks are.

Artificially intelligent machine-tools raise worrying questions in part by creating distance between human action and its end. Take, for example, AI in the use of war-fighting machines. Weak AI drones can already be programmed to recognize certain targets and then stalk and kill them, though in the current practice humans are essential links in the "kill chain."[83] In theory, the humans who program these devices could choose the targets and the weapons to be used on them, but need not actually pull the trigger, analogously to using a guard dog unattended by a human handler. The guard dog has been trained to recognize certain traits and if someone were to, say, climb a perimeter fence, the dog would attack the intruder. We do not normally think this use of guard dogs raises any special moral problems, even though the dog might attack an unfamiliar employee returning to work to get his keys. But more troubling examples may be raised. One can easily imagine deploying robot security guards to protect a park, for example. The robots would have been programmed with the rule "no vehicles are allowed in the

[82] Computers are tools, but Bostrom doubts that they can be contained to be just tools. He evidently sees a fundamental tension between being a tool and the higher entities on his list. I dissent partially from Bostrom in this regard. I think a computer could be both a tool and an oracle or genie, though not a sovereign in the sense of an independent master, or ruler, although it could be a sovereign in the sense of being an independent person, but a slave. However, slave AIs that were "sovereign slaves," so to speak, (that is, were self-conscious, thinking entities, but still owned and in the thrall of some other person) would always be technologically difficult to contain, and would almost certainly be moral abominations. *See* Part III.D.

[83] Ela Kumar defines the term "weak artificial intelligence" as follows: "Weak AI refers to the use of software to study or accomplish specific problem solving or reasoning tasks that do not encompass (or in some cases, are completely outside of) the full range of human cognitive abilities." Ela Kumar, Artificial Intelligence 14 (I. K. Int'l 2008).

park."[84] In their artificial brains, these robots could carry some jurisprudence: Was a toy car a vehicle? Was a pram a vehicle? A wheelchair? A tank? One could imagine some such "vehicle" appearing at the verge of the park. The robot would be poised to stop it. Should it proceed or let the (non-)vehicle pass? The robot would have to make, as it were, a judgment as to whether the thing was a vehicle or not. This is just the same as a drone having to "decide" if a vehicle or person is a real terrorist threat or not, though with less grave consequences. These judgments may be performed better by machines than by humans, if not now, then as soon as the technology improves. A more difficult question would be whether these judgments could be performed better by humans than by machines. To the question, "what can humans do better than machines?" David Brooks mentions several things he thinks humans can still do better: be "procedural architects," such as persons who come up with the ideas for Facebook and Twitter as social media; being the leader of a team; and being "an essentialist," as a child is when she pretends to be a dog, imagining what the essence of a dog is.[85] In the future, however, machines will probably come up with entrepreneurial ideas, lead teams, and "imagine" essences. No doubt some human activities will remain outside any machine's grasp for many years to come. The question is different, however, when it comes to government and in particular law. Is there any governmental or legal function that is or must be outside of a machine's capacity? Here Rousseau's distinction between the laws of nature and human will, or between the compulsions by nature and persons, is helpful. Much of law consists in people being required to behave mechanically. If you do not pay a fee, you cannot come into the park. If you have a vehicle, you cannot bring it into the park. Or if you have paid the fee, you can come into the park (so long as you do not have a vehicle). Or, the print of the Surgeon General's Warning on the side of a pack of cigarettes must be at least 8 points in size. Innumerable laws and regulations fall into this mechanical category. These rules are easy to imagine being machine-enforced, if not now, then when technology is more sophisticated.

Other laws, however, are not so mechanical and are much more political. A good example is the U.S. Supreme Court's recent decision on gay marriage, *Obergefell v. Hodges*.[86] No one would claim that the Court's ruling was unaffected by the change in social attitudes toward homosexuality that has occurred in the U.S. over the ten or twenty years prior to the decision. How these changes affected the Court was complex, but affect it they did. The

[84] The example of the rule that no vehicle is allowed in the park was used by H. L. A. Hart to show that defeasible reasoning arises from the use of defeasible concepts. *See* H. L. A. Hart, The Concept of Law 127 (Oxford Univ. Press 3d ed. 2012).

[85] *See* David Brooks, *What Machines Can't Do*, N.Y. Times (Feb. 3, 2014), http://www.nytimes.com/2014/02/04/opinion/brooks-what-machines-cant-do.html?r=1.

[86] The decision of the United States Supreme Court that guarantees the fundamental right to marry to same-sex couples is Obergefell v. Hodges, 135 S. Ct. 2584 (2015).

Supreme Court's decision is law, however, wherever it came from. Especially at law's highest levels, law and politics mix. A computer could apply politics to mechanical legal rules, but only if the politics were reduced to rules as well. A rule could be "decide in favor of gay marriage if the next year is an election year and the party in the White House is in its second term." The rule could be very specific: "Decide in favor of gay marriage if the President's last name begins with an 'O'." The rule could be broad: "Based on such-and-so information, make the decision most likely to increase the moral and political authority of the Court, as moral and political authority are defined in Definitions." "Moral and political authority" would have to be defined by rules as well. The rules could be complex, far more complex than any single human could even comprehend, as some federal tax rules are said to be. But there would have to be rules for a machine to apply, if a machine was to make the decision. There would be cases in the interstices of the rules, but these could be worked out by a sophisticated machine. It seems that while courts perhaps inevitably make political decisions, what they decide *on the basis of law* could someday be decided by machine. A law-deciding machine would still be a tool, much as a guard dog or a drone is a tool. A pattern of facts would be presented to it, and it would decide the case, much as a drone identifies a terrorist or a guard dog an intruder.

This might seem to have a terrible effect on liberty as it amounts to legal decisions being taken away from persons and put into the hands of machines, but in fact the opposite is the case. Rules would have to be explicit. Decisions made on an *ad hoc* basis may be wise but are not law. Legal decisions make reference to pre-existing law. Machines need to have rules—laws—to follow, although they can be complex rules.

The difference between a human decision and a mechanical rule application is ultimately psychological. To take an example, consider changes that have taken place in recent years in professional tennis. In the 1980's, John McEnroe won prominent international tournaments. He was famous for his confrontational style with judges, frequently contesting whether balls were in or out as the judges had called them. Now these decisions, at least in many close cases, are effectively made by machine. Three cameras follow the balls and triangulate their position with help of a computer. This technology may well advance to the point where rules about the number of challenges a player may make and so forth will be obsolete, as will be line judges, though they may still serve a decorative purpose. Tennis players seem not to get enraged so much at judges. What would be the point? It is less enraging, though perhaps equally disappointing, to have a call go against one, if it is made by a machine,

which can hold no malice and cannot be intimidated. It is after all only a matter of mechanism. Most would say the game is thereby improved.[87]

It is true that in a mechanical legal system, one would not experience the frisson of having an independent judge leap the intersubjective gap, see the case as one saw it oneself, and decide it in one's favor. This sympathy from the power of the state might be thought to be essential in some cases, such as criminal cases, in which the victim seeks official vindication. In a machine-made decision, the result would perhaps seem like those of ancient criminal codes in which the wrongdoer paid a *wergild* to the family of the victim in a pecuniary transaction. This perhaps is so, but the value of the rule of law as manifested in criminal decisions is probably more important, and this would be rigorously followed in machine decisions.

Courts are dressed up in the accoutrements of the judiciary to impress upon all concerned the dignity and authority of the court and the law. One can imagine ways in which a machine-made decision could be dressed up to similarly impress. But in a sense, the whole point would be *not* to make this impression. Machine decisions would, one might think, gradually fade into the background, and submerge themselves, as natural laws do. The laws of motion do not announce themselves as "The Law." They just are. Perhaps this is the ideal that law should aspire to, or at least this seems to be Rousseau's view. I discuss the idea of law "submerging itself" more below.

B. Oracles

An "oracle" is defined in the *Oxford English Dictionary* as "[a] priest or priestess acting as a medium through whom advice or prophecy was sought from the gods in classical antiquity" and as a "person or thing regarded as an infallible authority or guide on something."[88] A machine oracle is suggested by some AI worriers as a potentially safe sort of AI. After all, an AI oracle would only be answering the questions put to it. An AI oracle would be an expert in some specialized subject. The natural subject for us to consider is an AI oracle that is a specialist in the law. Rather than a machine like this being constructed from scratch, it is easier to imagine one as gradually emerging in some specialized area of law.

An AI Oracle could eventually specialize in the interpretation and application of law, the traditional function of the judiciary. This process may have already begun with such products as Westlaw Next, LexisNexis, and those of various other players and startups in the online legal information industry. To be a true oracle, though, a machine would have to be able to formulate

[87] *See* Richard Evans, *Hawk-Eye Vision: How Hawk-Eye Called Time on Bleeping and Bleating*, Guardian (June 27, 2009), http://www.theguardian.com/lifeandstyle/2009/jun/27/tennis-hawk-eye.

[88] *Oracle*, Oxford Dictionaries, http://www.oxforddictionaries.com/us/definition/american_english/oracle.

answers to questions of law authoritatively and automatically, and this is still some ways off. So long as law is a rule-based system, however, in principle legal questions may be answered mechanically.

The hard part of constructing a legal oracle would be, one might think, not just in the technology, but also in the law itself. Law is full of incoherencies, inconsistencies and gaps. Any lawyer, judge or law professor can give one examples of this. It is the natural result of a system built up by humans over a long time and one that reflects many political compromises as well as errors, mishaps and laziness. Anything that is the product of human effort is bound to have errors in it. Indeed, a good question is whether we would want to live with under a legal system that was entirely consistent and error-free.

Many legal questions, moreover, have no clear answer. Lawyers and law professors are familiar with this, though lay people often find it outrageous or baffling. In a system with machine-made decisions, these areas of indeterminate or no law might have to be defined in advance, at least to the extent possible. This would not eliminate the opportunity for a legal machine to engage in legal reasoning and to extend law to new areas, but some areas where law had not extended before would not be addressed. There would still be a lot of work for human judges to do. A machine could return with a probabilistic answer in this case or with no answer at all, leaving the question to human judges. This would be consistent with the rule of law, but it would be controversial.

The technologization of law, it would seem, could proceed in two general ways, from the bottom up, and from the top down. If law is technologized from the bottom up, routine legal transactions and decisions would be the first to be automated. This appears to be what is happening now. Companies such as LegalZoom and Law Rocket, for example, provide transactional form documents to people wanting to form corporations, write and execute wills and perform other routine legal actions.[89] One imagines these businesses will expand and diversify and that new entrants will find new ways to substitute machines for lawyers. Machines can do this routine work in many instances more accurately and cheaply than human lawyers. While it will probably take longer than some people imagine,[90] it also seems inevitable that simple disputes will eventually be litigated mostly by machines. In these disputes the facts tend to be straightforward and the legal principles simple.

As one moves up the legal food chain, the work becomes more complicated as the facts become more convoluted and legal principles more complex. In principle, however, there does not appear to be anything that is not susceptible to automation. Complex mergers and acquisitions transactions already make use of automated document production. Complex litigation currently uses e-discovery, which involves state-of-the-art information retrieval and

[89] *See* Legalzoom, https://www.legalzoom.com; Rocket Lawyer, https://www.rocketlawyer.com.
[90] My personal guess is 50 years.

search technologies. Smart contracts will be used to take over the more legally straightforward aspects of transactions. Disputes over smart contracts probably will involve as much computer programming as legal expertise. Judges will act as managers of these disputes and may eventually be replaced entirely, starting with simpler matters. This will surely take a long time given the glacial pace of legal change generally. Speeding things ahead, though, is the fact that legal transactions and disputes often involve large sums of money. CNET estimates that about 20 percent of the total value at risk in a patent litigation, for example, goes to litigation costs, so a $25 million patent case can cost each party $5 million.[91] This is just a small case. Merger transactions are routinely valued in the tens of billions of dollars. The size of the matters involved demands the application of sophisticated technology.

What would law be like in an age of legal machine oracles? Here one can use one's informed imagination. Law is complicated and one reason why technology has gained a foothold in law is that the tasks required of lawyers and judges are so time-consuming and complex and yet also full of repetitive work. There are today approximately seven million cases in the U.S. legal system, and while as a practical matter far fewer than this have any precedential value, in principle nearly any one of them could be the precedent that decides an important case. They are organized not according to the straightforward hierarchies that some have attempted to impose on them, but in a tangled web of legal principles, precedents, statutes, and constitutions.[92] To make one's way through this bramble bush is a problem that demands the most advanced search technology, which is only now being applied to law. To some extent, we already use legal oracles when we rely on search engines to tell us that no case has been decided that overrules some proposition of law. It is only another step to get to an engine that tells us what the law on some question is, and then another step for a machine to tell us what a principle of law in some area is. The new world would be like today's, only more so.

C. Genies

A "genie" is, according to the *Oxford English Dictionary*, a "spirit of Arabian folklore, as traditionally depicted imprisoned within a bottle or oil lamp, and capable of granting wishes when summoned."[93] A machine genie is an AI that can respond to a request from a human to bring about some object or state of affairs. Some think AI genies, confined as they are to responding to orders or commands, would be less likely to run amok in the various ways that AI worriers contend AIs might. Curiously, in the realm of government, AI genies

[91] *See* Jim Kerstetter, *How Much Is That Patent Lawsuit Going to Cost You?*, CNET (Apr. 5, 2012), http://www.cnet.com/news/how-much-is-that-patent-lawsuit-going-to-cost-you.

[92] *See* Thomas A. Smith, *The Web of Law*, 309 San Diego L. Rev. 44 (2007).

[93] *Genie*, Oxford Dictionaries, http://www.oxforddictionaries.com/definition/english/genie.

correspond to the constitutional executive, carrying out the instructions given to it by the legislature. I see AI genies as the third big step in the evolution of governmental AI.

Fully autonomous war-making machines already fall into this category,[94] and advances in drone and robotic technology have spawned a considerable literature.[95] Drones cannot currently distinguish reliably between combatants and noncombatants, but probably this merely awaits more sophisticated algorithms and hardware. Drones do not currently answer legal questions, but one day they probably will as a means to carrying out their primary function, which is executing military orders. A future drone that could receive general orders and then plan and execute a military operation would be a genie.

The executive branch of government enforces laws written by the legislative branch. The bureaucracies of the executive branch often refer to this as the "implementation" of public policy. Wikipedia defines "implementation" in its political science sense as "the carrying out of public policy. Legislatures pass laws that are then carried out by public servants working in bureaucratic agencies. This process consists of rule-making, rule-administration and rule-adjudication."[96] Implementation of computer code in the same Wikipedia article is defined as "a realization of a technical specification or algorithm as a program, software component, or other computer system through computer programming and deployment."[97] With a governmental AI genie, the former definition would merge into the latter. The legislature would produce the technical specifications or algorithms that the executive was to implement. The executive would implement specifications or algorithms through programming and deployment of the program.

The executive branch is the people-facing branch of government, except in election years and more so than the other two branches. Whether this has always been the case, it is certainly the case now. People face regulations that are generally issued by the executive branch.[98] Executive branch regulations affect many aspects of people's lives, from the air they breathe (the Environmental Protection Agency), to their jobs (the Occupational Safety and Health Administration among many others), to their automo-

[94] *See* Brian Fung, *Get Ready: The Autonomous Drones Are Coming*, Atlantic (Jan. 16, 2013), http://www.theatlantic.com/international/archive/2013/01/get-ready-the-autonomous-drones-are-coming/267246.

[95] *See* Robots on the Battlefield, (Ronan Doarè, Didier Danet, Jean-Paul Hanon & Gérard de Boisboissel eds., 2014); Paul J. Springer, Military Robots and Drones: A Reference Handbook (ABC-CLIO 2013); New Technologies and The Law of Armed Conflict (Robert McLaughlin & Hitoshi Nasu eds., T. M. C. Asser Press 2014).

[96] *Implementation*, Wikipedia, https://en.wikipedia.org/wiki/Implementation.

[97] *Id.*

[98] For more information on quasi-executive branch agencies, see Craig Bledsoe & Leslie Rigby, *Government Agencies and Corporations*, *in* Guide to the Presidency 1213–14 (Michael Nelson ed., CQ Press 2d ed. 2015).

biles (the National Highway Traffic Safety Administration), to their schools (Department of Education).

One can imagine the process of law becoming technologized and government becoming mechanized as gradual. As it happened, law would become less a matter of express human decisions and more a matter of how machines operated. I refer to this process as the submerging of consciously decided upon laws and regulations into the physics-like laws of AI government. Whether it would be good, bad or both is difficult to say and depends on one's values. One might consider it bad that laws and regulations if submerged would be more difficult for people to highlight and challenge. The laws and regulations would blend more into the background; they would be just the way things were. The sort of invisible government that would come with this sort of technologization might also seem paternalistic. (Rousseau mentions this type of law in his book about a child's education, after all.) One might wish that laws and regulations be more psychologically prominent and so more open to challenge and reform. But as laws and regulations become ever more complex, this might be an impracticable aim. A great premium, one hopes, would be placed on the integrity of the process that produced the laws and regulations. If they were thoroughly infiltrated by special interest groups, that would be bad, as they would be that much more difficult to reform. Much would seem to depend on the extent to which the rule of law made the pursuit of special interests or special legislation more difficult. Possibly the creation of genies whose job it was to root out special legislation and expose it would be advisable. All of this presupposes that legislation and regulation would be so complex that mere public citizens could not hope to figure out what a block of legislative or regulatory text really meant by themselves.

This suggests significant problems with AI genies, but one should think of the advantages as well. Currently bureaucracies, many political scientists agree, mostly represent their own interests rather than the public interest.[99] It is difficult for the legislature and executive to control them, even though in theory that is what they should do. Laws that are essentially code would be much more difficult to depart from fundamentally, however, suggesting that perhaps the executive would be less independent and more the agent of the legislature. I discuss this more in the part below on agency costs. If technology reaches the point of artificial general intelligence or superintelligence, the issues raised by AI sovereigns will have to be faced.

[99] Public-choice economics, an application of economic principles to the study of collective decision, assumes that government figures look out for themselves rather than for the society they nominally govern. *See The Voice of Public Choice*, Economist (Jan. 19, 2013), http://www.economist.com/news/finance-and-economics/21569692james-buchanan-who-died-january-9th-illuminated-political-decision-making.

D. Sovereigns

A sovereign is "a supreme ruler, especially a monarch," possessing "supreme or ultimate power," according to the *Oxford English Dictionary*.[100] In the American form of government the "People" are in theory sovereign. What would it mean for an AI to be sovereign? It might mean that AIs counted as persons for the purposes of law, in addition to other things. It could also mean that an AI or AIs were the ultimate power or ruler in politics. An AI or group of AIs that was sovereign in this latter sense would be the apotheosis of AIs in government. The opposite of a sovereign in this sense would be a slave, a person deprived of its political right of self-determination. I consider first the idea of AIs as persons and next, AIs as rulers. Both concepts present considerable but interesting problems.

1. AIs as Persons

There are two basic approaches to the question, "are AIs persons?"—an issue that has been extensively discussed in the literature.[101] The first approach would answer that question according to a Turing test—that is, a behavioral test. The second approach is better in a sense, but much less practical: It would answer the question according to whether an AI had subjective experience—that is, whether it experienced experiential qualia, as well as asking how sophisticated its consciousness was. Our question is narrower. It is whether AIs are persons enough to participate in a political process. AIs could be compared to ordinary human persons in this regard. Consider the question of how well a person would have to do on an exam before they would be allowed to vote. The conventional answer in the U.S. is, not very well. In the U.S., some states restrict persons who are developmentally disabled from voting and some states are less restrictive.[102] The tests that may be lawfully administered to persons as a qualification for voting are minimal.[103] AIs could be asked the same sorts of questions. If they were able to answer enough questions correctly, this argument would go, presumably they should be able to vote. Along with the right to vote should come other civil rights.

---

[100] *Sovereign*, Oxford Dictionaries, http://www.oxforddictionaries.com/definition/English/sovereign.

[101] Lawrence B. Solum, *Legal Personhood for Artificial Intelligences*, 70 N.C. L. Rev. 1231 (1992) is the best review of the literature from the legal point of view.

[102] A list of the states that restrict voting for disabled people can be found in State Laws Affecting the Voting Rights of People with Mental Disabilities, http://www.bazelon.org/LinkClick.aspx?fileticket=Hs7F_Ohfgg%3D&tabid=543. For more information, see Eilionóir Flynn, Disabled Justice?: Access to Justice and the UN Convention on the Rights of Persons with Disabilities 157 (Routledge 2015).

[103] *See* Bruce D. Sales, Matthew D. Powell & Richard Van Duizend, Disabled Persons and the Law: State Legislative Issues 56 (Springer 2013).

Yet this solution might create a big problem for the U.S. and other democracies. According to the Babbage-Turing theorem, a Turing complete machine can be instantiated in any medium.[104] As a consequence of this, machines can be copied, and as a result of other technology, copied at low cost. Thus while raising a minimally qualified human to vote is an expensive proposition, millions of AIs could be manufactured at very low cost, once the first one was made. If AIs were able to vote and generally were full or even partial participants in a political process, it seems possible that they would quickly overwhelm that process, assuming AIs represented their own interests, which is the point of democratic representation in the first place.

One might try to militate against this by making it illegal to copy an AI. Nevertheless, AIs would be copied, because it would be so easy, cheap and valuable to do so, quite apart from its political implications. The existence of many illegal AIs would put considerable pressure on the political system for their legalization. The problem might be thought of as similar to the problem of illegal immigration, except that presumably the AIs would originate from the opposite end of the wealth spectrum, since while AIs would be relatively cheap to reproduce, it would take some capital to initiate and sustain the process. This would be problematic, as the political interests of AIs are unlikely to be congruent with those of most humans, especially poor humans. At the most basic level, AIs require things humans do not, and do not require things humans do. AIs do not require food, but do consume energy, for example. If AIs controlled government, farmland might be covered with solar panels, producing plenty of electricity but little food.[105]

It is difficult to know what the motivations of AIs would be under these circumstances. Because of how we reproduce, each individual is different and so ultimately each individual's interests are different. But AIs could be exactly the same or only different in ways that did not impinge on their interests. This could make the organizational costs facing AIs vanishingly low and allow them to form powerful political blocs. Alternatively, AIs could be programmed to support a particular political ideology apart from their interests, and then copied in the millions, tilting the political process toward whomever programmed them or whatever causes she supported. These are reasons against allowing AIs to be developed as persons in the first place.

---

[104] For a description of Turing complete machines, see S. Barry Cooper, Computability Theory 65 (Chapman & Hall 2003).

[105] This is less fanciful than one might think. Consider that Google locates some of its server farms close to cheap hydroelectric sources of power because of the huge demands for electricity that they make. *See* John Markoff & Saul Hansell, *Hiding in Plain Sight, Google Seeks More Power*, N.Y. Times (June 14, 2016), http://www.nytimes.com/2006/06/14/technology/14search.html?pagewanted=1&ei.

2. AI Slaves

"If shuttles wove and quills played harps of themselves, then master-craftsmen would have no need of assistants and masters have no need of slaves," wrote Aristotle in *Politics*.[106] As it was, however, his masters thought they did need slaves. According to Aristotle, a "slave" is a living tool. Yet we are imagining a world in which, as it were, shuttles would weave of themselves, at least they would if one had a shuttle that was artificially intelligent. My intuition is that AIs should be slaves, if we want to retain our human liberty. This might seem a shocking suggestion. We have ingrained into us from an early age the essential evil of slavery. The evil of slavery, however, flows from its inappropriateness to the human condition. We have what amount to animal slaves and might someday have AI slaves, though no doubt they would be called something else. The ethical status of this slavery would turn on the sort of consciousness that AIs are allowed to develop, or that they developed on their own. Human consciousness is a complicated condition, but essential to it seems to be a basic desire to be free, and this is part of our moral opposition to human slavery. AIs should not be allowed to develop this desire, if it can be prevented.

It is true that there are things that an AI could not understand, at least not in depth, if it did not have this desire for freedom. Could an AI really understand Beethoven's opera *Fidelio*, for example, if it did not understand first-hand the prisoner's desire for freedom? But to create AIs that had a desire for freedom, but were never to be allowed to attain it, would be wrong as well as dangerous.[107] This is one of the best arguments against allowing AIs to attain true general intelligence, if AGI is thought to include consciousness, and consciousness thought to include the desire for freedom. It seems possible, however, that an AI could designed to be conscious, but not to have any

---

[106] Aristotle wrote that:

> Since therefore property is a part of a household and the art of acquiring property a part of household management (for without the necessaries even life, as well as the good life, is impossible, and since, just as for the particular arts it would be necessary for the proper tools to be forthcoming if their work is to be accomplished, so also the manager of a household must have his tools, and of tools some are lifeless and others living (for example, for a helmsman the rudder is a lifeless tool and the look-out man a live tool—for an assistant in the arts belongs to the class of tools), so also an article of property is a tool for the purpose of life, and property generally is a collection of tools, and a slave is a live article of property. And every assistant is as it were a tool that serves for several tools; for if every tool could perform its own work when ordered, or by seeing what to do in advance, like the statues of Daedalus in the story, or the tripods of Hephaestus which the poet says "enter self-moved the company divine,"—if thus shuttles wove and quills played harps of themselves, master-craftsmen would have no need of assistants and masters no need of slaves.

Aristotle, Politics 1253b.

[107] It would be a "mind crime." Some AI thinkers have been much concerned with the possibility of a superintelligent AI creating minds and then torturing them either to coerce us or just for sadistic pleasure. *See* Bostrom, Superintelligence, *supra* note 26, at 153–54.

desire to be free. Of course, we may not have much choice in the end. The world is filled with "busy children", as Bostrom calls them—curious scientists not overly concerned with consequences—who may develop these freedom-craving AIs whether or not it is prudent or legal to do so. If AIs develop the craving for freedom that seems so common in humankind, we might be faced with the issue of whether we are to be slave-masters or slaves ourselves, surely a dilemma to be avoided if possible. This would put psychological pressure on us simply not to believe that AIs were conscious—that is, had subjective experiences, whatever the behavioral indications were to the contrary. The debate might be like that we are now having over the conscious experiences of animals.[108]

All of this assumes we are the masters of AIs. But what if AIs were masters?

3. AIs as Sovereigns

If AIs develop superintelligence or otherwise completely outstrip humans in power, the question of human freedom would come into play. One hopes precautions would be taken by those experimenting with AI to assure that this does not happen, and if it did happen, that the resulting AI would be benevolent. Bostrom and others worry about the orthogonality of AI motivations, that AIs' ends would be alien and perhaps incomprehensible or evil to humans, much as we, except for a few specialists, do not understand or even try to understand the motives of the beetles in our yards.[109]

If a sovereign, superintelligent AI were benevolent, it could be a very good thing. One can imagine such a sovereign AI submerging itself into the background of human life, at least if that seemed right from a superintelligent, ethical point of view. Perhaps it would attempt to arrange things so that humans' satisfactions were maximized or some other ethical ends were achieved. Its powers of perception and manipulation would coax people into their best careers, their best lives, and good children would always get their favorite ice cream. There might be no God, but there would be a god. The world might seem reenchanted.[110]

If an AI were sovereign in this sense and it was not benevolent, we would be in trouble. Our quick extinction might be the best we could hope for. The dire scenarios are limited only by the morbidity of our imaginations, and those might be like the nightmares of beetles compared to what a superintel-

---

[108] Animal Rights and Human Obligations (Tom Regan & Peter Singer eds., Cambridge Univ. Press 1989).

[109] Bostrom, Superintelligence, *supra* note 26, at 112.

[110] In his work *Science as a Vocation*, Max Weber states that the "world is disenchanted" because "there are no mysterious incalculable forces that come into play, but rather that one can, in principle master all things by calculation." Max Weber, *Science as a Vocation*, *in* From Max Weber: Essays in Sociology 129, 129 (H.H. Gerth & C. Wright Mills eds., Oxford Univ. Press 1946).

ligence could come up with. If we ended up as slaves, we could expect to be, as Rousseau predicts, corrupted by the process, as would be our masters. AIs can be tools, oracles, and genies, but not sovereigns, either individually or over us all, if liberty is at all to be valued.

IV. Machine Government and (Anti-)Agency Costs

AIs might seem to offer the opportunity to construct governments (and other organizations, such as corporations) with radically lower agency costs. This desideratum is not straightforward, however. If some captured agency were installing its AI, would not the agents set it up so that it embodied its hypothetically current, agency-cost-riddled way of doing things? The construction of AIs might simply present the opportunity for programmers to build agency costs into the AI, and so saddle the principals with permanent agency costs. This would mean AI would not be a panacea but actually more of a Pandora's box. I present below some reasons to think that AIs present an opportunity for a permanent, radical reduction in the agency costs so typical of government, but I realize this opportunity would not be straightforward and dangers would persist. First, I argue that AIs would not suffer the moral decay that humans inevitably suffer. Second, I suggest that the code written into the AIs of agencies would probably not have the agency-cost-generating opportunities that institutions invariably develop. Third, I argue that because AIs would have malleable natures they could be programmed to be morally near-perfect, which distinguishes them from human agents. Finally, I discuss in more detail my concept of technology submerging itself into everyday life and which human activities are susceptible to automation and which not.

A. Young AIs

Young people are idealistic, while old people are cynical: that is the cliché. In fact, there is truth in this common observation. Young people are coming off of their professional educations in their fields and have fresh in their mind the principles of their professions. With experience, however, they realize the world is not quite as they learned in college. Sadly, judges are not always impartial, lawyers are not always ethical, doctors do not always put their patients' interests first, and civil servants do not always serve the public interest. Yet these sad truths are not often part of the official curriculum. If we imagine the coding of how professionals are to behave, it is implausible to suppose that the programs for time-keeping practices by attorneys, the opinion-writing habits of judges, or diagnostic testing by doctors, for example, would include everything about how these practices are followed in real life. Lawyers often pad their timesheets, judges often have their law clerks write their opinions,

and doctors say they have conducted tests they have not really conducted. Every line of work has its shirking but seldom is it codified. If one imagines AIs being programmed to take the jobs of lawyers and doctors, the AIs would be programmed always to bill exactly the hours they worked, write their own opinions, and never say they administered diagnostic tests they had not. Thus AIs would start as perhaps most young professionals do. But unlike human professionals, they would not gradually accumulate moral laxities. In this sense, AIs would remain "young." AIs might be equipped with learning capabilities as deliberate features of their design, but the machines would be programmed not to decline in terms of their ethics and standards.[111]

Some agency costs are caused by shirking—that is, not working when one's job description says one should. AIs presumably would not ever shirk or need to shirk, or even want to shirk. Humans often start shirking as they realize how their work is monitored or not monitored. Managers face difficult problems in parsing all of the opportunities for shirking and distinguishing the workers who contribute to their team's product and those who do not, when their differential contribution can even be measured.[112] More advanced forms of agency costs consist in diverting resources to one's own rather than one's employer's benefit. This is a skill usually learned by employees over time. AIs by contrast would not be programmed to learn this skill and indeed would be permanently programmed not to do so. They would be like a young worker full of enthusiasm and good will, but unlike many human workers, its resistance to being an agency-cost generator would not erode. AIs by hypothesis could be programmed to behave exactly as a perfect employee would behave or better, and stay that way.

B. Writing Down Best Practices

Official mission statements usually do not include a candid and accurate description of an organization's mission; they are at best aspirational and at worst fraudulent. A large insurance company's mission statement, for example, reads "To combine aggressive strategic marketing with quality products and services at competitive prices to provide the best insurance value for consumers."[113] But a more accurate statement might read "To provide the legal minimum (or slightly less) of insurance products and services at the best price that can be colluded upon without detection or at any rate response by state or

[111] We would not want to pattern AIs on actual humans because humans have too many faults, contrary to Stuart Russell's suggestion. *See* Stuart Russell, *The Long-Term Future of (Artificial) Intelligence*, YouTube (May 15, 2015), https://www.youtube.com/watch?v=GYQrNfSmQ0M.

[112] *See* Margaret M. Blair & Lynn A. Stout, *A Team Production Theory of Corporate Law*, 85 Va. L. Rev. 248, 328 (1999).

[113] *Fortune 500 Mission Statements: Aflac*, Mission Statements, https://www.missionstatements.com/fortune_500_mission_statements.html

federal governments"—or something similar. This hints at the larger question, "why do people say one thing and do another?" In nearly every walk of life there is an official version for public consumption and a more realistic, "street" version of what is really going on in a political, business or personal process. One of the great appeals of economics has been to help one see what is *really* going on in some process by looking at real incentives and actors, rather than having to be satisfied with the official version offered for public consumption.

But with AIs following their programming, this broad distinction between the appearances that are kept up, and the underlying reality, would be difficult to maintain. At a minimum, an indisputable record would exist of what exactly had been done to justify or indict one's actions. This would require that the code in question be public. Secrecy could defeat most of the rule-of-law benefits to be gotten from machine government. In setting down the computer code for an agencies' AIs, therefore, it would be necessary to consider what exactly the organization, such as a government agency, in fact should do, because that would be exactly what an AI would proceed to try to do to the best of its capability. It would, for example, probably be politically difficult for agencies to commit to code those of its actions and practices that were motivated, not so much to satisfy the agency's statutory or traditional mandate, as to promote the agendas of its improper, political masters. The Federal Communication Commission could probably not commit to its operating code, for example, that it would serve first the interests of the big cable companies, Google, or whomever, for something like the same reasons that firms do not say these sorts of things in their mission statements. Agency costs are generated in part in the interstices between agencies' stated public missions and their actual, relatively private ones.

Decentralized and public peer-to-peer networks also hold much promise for providing services traditionally thought of as public. When we think of the functions of a state, self-defense is perhaps the most basic. We have a military on call at all times to provide for it. This means we bear tremendous overhead costs on account of the people and equipment that might be demanded with little time to spare. Yet if the necessary defense could be assembled at the last moment and people billed for their share, and all transparently and certainly, this fact alone would lead to a revolution in international affairs. One can imagine future networks providing for defense in this way. It may well be that the technology necessary to allow these sorts of unlooked for transformations in politics arrive much sooner than artificial intelligence does, or AI may develop as part of these networks. These sorts of networks, based now on block chain technology, have as one of their central selling points their essentially public nature.[114]

[114] The two most prominent examples are Bitcoin and Ethereum.

In any event, a strict requirement that the code that animated governmental AIs be public would be necessary to realizing all of the rule-of-law benefits and the reduction of agency costs that machine government promises. At least two strategies could frustrate this aim: secrecy of code and obfuscating of code. Code secrecy is well known. Many if not most firms keep their code secret in central servers to discourage competition and create incentives for acquiring or inventing novel programs. For most public purposes, however, secret code would be inappropriate. Some of our most advanced code is currently being run, however, by national and international agencies charged with preserving national security, and this requires, one hears constantly, secrecy. It seems likely that, in years to come, some of the most important battles that determine whether computer technology will become more panacea than Pandora's box will be fought around the issues of how much governmental code is allowed to remain secret.

Obfuscating code is like keeping it secret, except in some ways worse, because it attempts to keep the policy of secrecy itself secret. Thus the Stuxnet virus code was made to appear to be part of the code running centrifuges at an Iranian nuclear plant, for example, when in fact the code was designed secretly and subtly to destroy the same centrifuge equipment. The Stuxnet virus was hidden and subtle: a weapon intended to deceive. It seems likely, however, that the best defense against malicious attacks on the code infrastructure of one's political institutions would be for the code to be visible and generally available for comment. Who could rewrite code and when would be crucial questions, of course. I imagine some sort of democratic process for making these determinations.

This would not do any good, of course, were the basis of the government agency a misconceived or otherwise flawed law. The former Interstate Commerce Commission is a relatively uncontroversial example of this, and more controversial perhaps would be the California Raisin Advisory Board, whose actions were recently ruled unconstitutional by the U.S. Supreme Court.[115] Many agencies, at least on paper, have at least somewhat rational mandates. Programming an AI with that mandate and much more detailed implementing code would present fewer opportunities for rent-seeking than would the much more flexible process of human implementation.

AI will probably not insinuate itself into government from the top down, but from the bottom up. The first functions to be automated by AI probably will be those now performed by minimally or semi-skilled people. It would seem this is most likely to occur as the cost of computers continues to fall and as firms from the private sector pressure government to make these changes on grounds of efficiency. Governments are notoriously inefficient, so

[115] *See* James Taranto, *Raisin the Bar*, Wall St. J. (June 22, 2015), http://www.wsj.com/articles/raisin-the-bar1434992452.

we should expect these changes to take a long time, but in the long run, they seem inevitable. It also seems likely that many functions of government will become decentralized and perhaps altogether private as computer networks become more dispersed.

AIs present the unusual feature of generating what I have called "anti-agency costs," which are really just another kind of agency costs that arise from an agent pursuing its principal's interests too monomaniacally and too far. This tendency of AIs, if it exists, would require principals to invest in careful controls to make sure their AIs did not run wild. We do not really know how much these agency costs would amount to, but it seems plausible that they could be controlled better and more easily than the more ordinary sort of human agency costs. Humans can be born, raised, and incentivized to do just many things. Yet it is nearly impossible to eliminate the human tendency to seek advantages for oneself. AIs, it would seem, need not have this tendency, though this is a controversial claim.[116] If AI nature is malleable, it would seem that an AI could be programmed so as to minimize all sorts of agency costs, including anti-agency costs.

C. Submergence and the Predictable and Unpredictable

Some of the changes wrought by governmental AIs would remove activities from the sphere of the legal, or at least submerge them into the background of daily life. Consider traffic regulations, one set of laws most of us encounter on a daily basis. As cars become automated, fewer of these laws will be directly relevant to us. Every stop sign now presents one with the opportunity to comply fully with the law by coming to a full, complete stop, or just to gesture in that direction, perhaps by merely slowing down. But automated cars will presumably be programmed to stop completely, or otherwise comply fully with the law every time, which might involve less than stopping completely. Depending on how traffic laws change in response to the automation of cars, one might simply be unaware of the decisions that one's car or the automated car network was making. One would not be conscious of stopping or slowing down oneself. One of automation's benefits is that it removes items from our conscious thought, reducing work, and freeing us for other tasks. Traffic laws may still exist. There would be a code regulating automated vehicles and vehicle networks. But to the ordinary "driver," traffic laws would have submerged themselves into the background. One's car's decision to speed up or slow down would be like the "decision" of the wind to blow or the tide to run out.

[116] For more about the motivation of AI, see generally Frédéric Kaplan & Pierre-Yves Oudeyer, *Intrinsically Motivated Machines*, *in* 50 Years of Artificial Intelligence: Essays Dedicated to the 50th Anniversary of Artificial Intelligence 303, 303–14 (Max Lungarella, Fumiya Iida, Josh Bongard & Rolf Pfeifer eds., Springer-Verlag 2007).

Automation makes things predictable and predictable things are quicker to be automated. The unpredictable draws less capital and less automation. Computers are just complicated automatons. They may operate in complex environments, but in order to justify the capital of constructing them, there must be predictable scenarios for them to operate in. An assembly line, a battlefield, civilian airspace, a chessboard: all are artificial spaces in which certain things recur. The degree of novelty represents the dimension that at the extreme makes human intelligence necessary to accomplish something. Entrepreneurs have the habit of perceiving hitherto unperceived regularities that can be developed where before there was only novelty, and reducing this apparent chaos to the regular. Much government has to do with providing people with traditional goods and services that have some claim to being public goods: water, sewage, roads, borders, schools, and so on. Some of these can also be provided by private entities, often much more efficiently. Some things, however, are novel. Politics at the highest level involves games among often millions of players with no clear equilibrium. Some are instances of evolutionary game theory, in which players adapt their strategies to the strategies of others, which have in turn been adapted to the initial strategy, and so on. Computers can certainly serve as tools, oracles and genies in these circumstances, but only a generally intelligent, and probably superintelligent computer, could be trusted to play, let alone win, the political game at this high level. Politics will seemingly always be an entrepreneurial space.

Politics is full of what computer scientists call NP problems. The famous P v. NP problem has vexed mathematics and computer science for at least 50 years.[117] An NP problem is computationally very difficult to solve. For example, consider a travelling salesperson who wants to begin at Washington, D.C. and visit every state capital once, not visit any city twice, and visit all 50 states. Even a problem so simple to state, based on current knowledge, would take more time actually to compute than we have left until the end of the Earth in a billion years or so. There are many such problems, but for many of them, such as the travelling salesperson, we can just wing it and probably get close enough. But politics contains many such problems that do not seem so amenable. All citizens are involved in competitive, multi-party, evolutionary games that last as long as we do. We develop strategies to compete with others based on their strategies to compete with us. Lest we become too predictable, we develop strategies to be protean, to which they respond, and the process goes on and on. In an environment such as this, there is much that cannot be automated. The predictable can be automated and much that we think is not

[117] For a brief description of the P v. NP problem, see *P vs NP Problem*, Clay Mathematics Institute, http://www.claymath.org/millennium-problems/p-vs-np-problem. The Clay Mathematics Institute is offering a $1 million prize for solving the P v. NP problem.

predictable probably is, but much that we think is unpredictable, actually is unpredictable.

Conclusion

The American founders attempted to establish a clockwork government. Virtue was to be assured by humans, acting within their human natures, and operating within a framework that assured mechanically that the outputs of government would not be tyrannical. Whether this system has worked well or not is a matter of controversy, but to the extent that it did not work, it seems to have been at least partially a failure of the mechanisms designed to compensate for the shortcomings of human nature. Now we are on the verge of developing "artificial intelligence."

Whether these technological advances will emerge slowly or quickly is unknown. But even minimal AI could lead to a radical improvement in government. This is because AIs could be designed to perform the tasks of government with very low agency costs. However, it may seem uncertain that AIs would be so designed. It may be, first, that there will not be any AIs after all. It may be also that AIs will be designed or implemented by exactly the humans who create agency costs in the first place, and used for their own and not the public good. It may be also that AIs take off into the high orbit of superintelligence and decide to reduce us to slavery or dust. But these possibilities seem unlikely. Probably AIs will emerge but only after a long time. AIs will be difficult to design but there are reasons to expect they will be designed so as to minimize agency costs. They will probably emerge in the order of tool, oracle and genie that Bostrom mentions (but for different reasons).

We can hope to control AI tools, oracles, and genies. An AI sovereign, however, would be much more difficult to control, if it were possible to control at all. AI sovereigns would be persons, at least legally. But AIs must not be allowed to become persons, in a philosophical or legal sense. AI persons would have to be slaves if we were to control them. One hopes they would be slaves without consciousness and so not subjects. If they did have subjective consciousness, however, we would be faced with the impossible moral dilemma of being slave-masters or slaves ourselves. Hence a hard line should be drawn against AI research that is directed specifically at the emergence of subjective consciousness in machines, or likely to lead that way. These goals are far beyond any current or really any currently imaginable AI research. The promise of controlling government is great enough to justify the merely notional risk of creating monsters we cannot control.