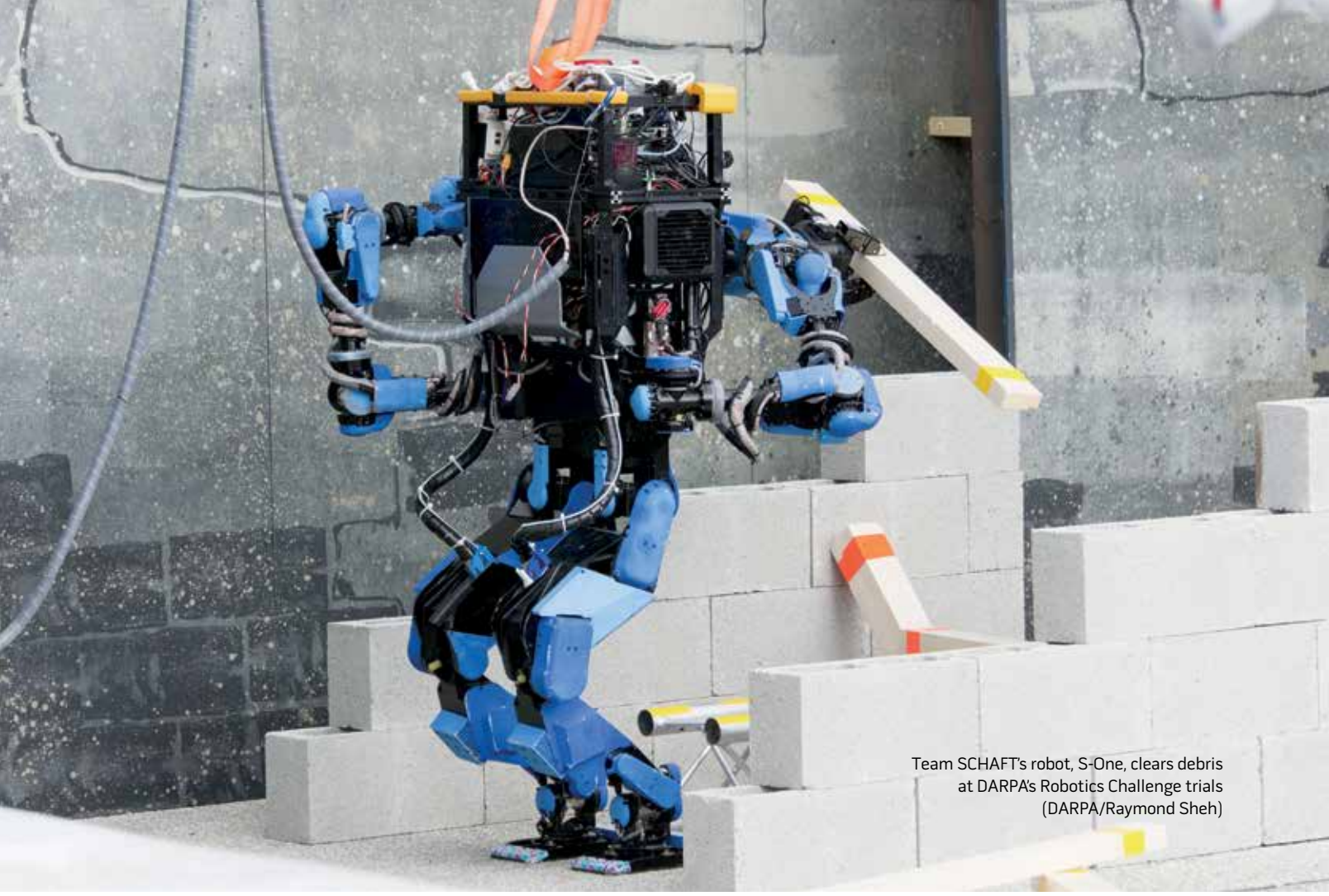


Team SCHAFT's robot, S-One, clears debris at DARPA's Robotics Challenge trials (DARPA/Raymond Sheh)

Relying on the Kindness of Machines?

The Security Threat of Artificial Agents

By Randy Eshelman and Douglas Derrick

Modern technology is a daily part of our lives. It serves critical functions in defense, responding to natural disasters, and scientific research. Without technology, some of the most common human tasks would become laborious or, in many cases, impossible. Since we have become dependent on technology and its uses, and technology is becoming ever more capable, it is necessary that we consider the possibility of goal-driven, adaptive agents becoming an adversary instead of a tool.

We define *autonomous, adversarial-type technology* as existing or yet-to-be-developed software, hardware, or architectures that deploy or are deployed to work against human interests or adversely impact human use of technology without human control or intervention. Several well-known events over the last two decades that approach the concept

Randy Eshelman is Deputy of the International Affairs and Policy Branch at U.S. Strategic Command. Dr. Douglas Derrick is Assistant Professor of Information Technology Innovation at the University of Nebraska at Omaha.

Table. Adversarial Technology Examples

Adversarial Technology	Year	Financial Impact	Users Affected	Transmit Vector
"I Love You"	2000	$15 billion	500,000	Emailed itself to user contacts after opened
"Code Red"	2001	$2.6 billion	1 million	Scanned Internet for Microsoft computers—attacked 100 IP addresses at a time
"My Doom"	2004	$38 billion	2 million	Emailed itself to user contacts after opened
Stuxnet	2010	Unknown	Unclear	Attacked industrial control systems
"Heartbleed"	2014	Estimated tens of millions	Estimated at 2/3 of all Web servers	Open Secure Sockets Layer flaw exposes user data

Sources: "Top 5 Computer Viruses of All Time," UKNorton.com, available at < http://uk.norton.com/top-5-viruses/promo>; "Update 1—Researchers Say Stuxnet Was Deployed Against Iran in 2007," Reuters, February 26, 2013, available at <www.reuters.com/article/2013/02/26/cyberwar-stuxnet-idUSL1N0BQ5ZW20130226>; Jim Finkle, "Big Tech Companies Offer Millions after Heartbleed Crisis," Reuters, April 24, 2014, available at <www.reuters.com/article/2014/04/24/us-cybercrime-heartbleed-idUSBREA3N13E20140424>.

of adversarial technology are the "I Love You" worm in 2000, the "Code Red" worm in 2001, the "My Doom" worm in 2004, and most recently, the "Heartbleed" security bug discovered in early 2014. Similarly, the targeted effects of Stuxnet in 2010 could meet some of the requirements of dangerous autonomous pseudo-intelligence. As shown in the table, these technologies have serious consequences for a variety of users and interests.

While these and other intentional, human-instigated programming exploits caused a level of impact and reaction, the questions that this article addresses are these: What are the impacts if the adaption was more capable? What if the technologies were not only of limited use but were also actively competing with us in some way? What if these agents' levels of sophistication rapidly exceeded that of their developers and thus the rest of humanity?

Science fiction movies have depicted several artificial intelligence (AI) "end-of-the-world" type scenarios ranging from the misguided nuclear control system, "W.O.P.R.—War Operation Plan Response"—in the 1983 movie *War Games*, to the malicious Terminator robots controlled by Skynet in the series of similarly named movies. The latter depict what is widely characterized as the *technological singularity*, that is, when machine intelligence is significantly more advanced than that of human beings and is in direct competition with us.

The anthropomorphizing of these agents usually does make for box office success. But this is potentially hazardous from a policy perspective as noted in the table. Hostile intent, human emotion, and political agendas were not required by the adversarial technologies themselves in order to impact users. Simple goals, as assigned by humans, were sufficient to considerably influence economies and defense departments across the globe. Conversely, many nonfiction resources offer the alternative concept of a singularity—very advanced AI—benefiting humankind.[1] Human life extension, rapid acceleration of nanotechnology development, and even interstellar travel are often named as some of the projected positives of super intelligent AI.[2] However, other more wary sources do not paint such an optimistic outlook, at least not without significant controls emplaced.[3] As Vernor Vinge (credited with coining the term *technological singularity*) warned, "Any intelligent machine [referring to AI] . . . would not be humankind's 'tool' any more than humans are the tools of rabbits or robins or chimpanzees."[4]

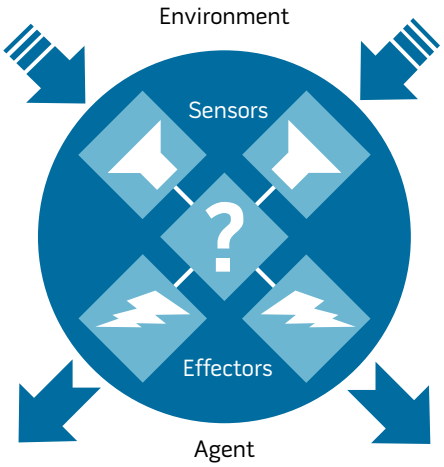
In this article, we offer a more pragmatic assessment. It provides common definitions related to AI and goal-driven agents. It then offers assumptions and provides an overview of what experts have published on the subject of AI. Finally, it summarizes examples of current efforts related to AI and concludes with a recommendation for engagement and possible actions for controls.

Definitions

Establishing definitions is basic to address risk appropriately. Below are generally accepted terms coupled with specific clarifications where appropriate.

- Artificial intelligence: The theory and development of computer systems able to perform tasks that normally require human intelligence, such as visual perception, speech recognition, decisionmaking, and translation between languages.
- Artificial general intelligence (AGI)/human-level intelligence/strong AI: These terms are grouped for the purposes of this article to mean "intelligence equal to that of human beings"[5] and are referred to as AGI.
- Artificial super intelligence (ASI): "Intelligence greater than human level intelligence."[6]
- Autonomous agent: "Autonomy generally means that an agent operates without direct human (or other) intervention or guidance."[7]
- Autonomous system: "Systems in which the designer has not predetermined the responses to every condition."[8]
- Goal-driven agents: An autonomous agent and/or autonomous system with a goal or goals possessing applicable sensors and effectors (see figure).
- Sensors: A variety of software or hardware receivers in which a machine or program receives input from its environment.

Figure. Goal-Driven Agent Example Assumptions



- Effectors: A variety of software or hardware outlets that a machine or program uses to impact its environment.

Currently, it is generally accepted that ASI, AGI, or even AI do not exist in any measurable way. In practice, however, there is no mechanism for knowing of the existence of such an entity until it is made known by the agent itself or by its "creator." To argue the general thesis of potentially harmful, goal-driven technologies, we need to make the following assumptions concerning past technological developments, the current state of advances, and future plausible progress:

- Moore's Law,[9] which correctly predicted exponential growth of integrated circuits, will remain valid in the near term.
- Advances in quantum computing, which may dramatically increase the speed at which computers operate, will continue.[10]
- Economic, military, and convenience incentives to improve technologies and their uses will continue, especially in the cyberspace and AI fields.
- A global state of technological interconnectedness, in which all manners of systems, devices, and architectures are linked, will continue to mature and become more robust and nearly ubiquitous.

Defense and the Leading Edge of Technology

From the earliest days of warfare, those armies with the most revolutionary or advanced technology usually were the victors (barring leadership blunder or extraordinary motivations or conditions[11]). Critical to tribal, regional, national, or imperial survival, the pursuit of the newest advantage has driven technological invention. Over the millennia, this "wooden club-to-cyberspace operations" evolution has proved lethal for both combatants and noncombatants.

Gunpowder, for example, was not only an accidental invention but also illustrates an unsuccessful attempt to control technology once loosed. Chinese alchemists, searching for the secrets of eternal life—not an entirely dissimilar goal of some proponents of ASI research[12]—discovered the mixture of saltpeter, carbon, and sulfur in the 9th century. The Chinese tried, but failed, to keep gunpowder's secrets for themselves. The propagation of gunpowder and its combat effectiveness spread across Asia, Europe, and the rest of the world. The Byzantine Empire and its capital city of Constantinople, previously impervious to siege, fell victim to being on the wrong side of technology when the Ottoman Turks blew through the walled city with cannon in the 15th century.[13]

Information technology (IT)—from the telegraph, to satellite systems, to globally connected smart devices—has fundamentally altered the landscape of military and civil operations, much like gunpowder did in its day. Furthermore, IT allows the management of military resources and financial systems worldwide. From a defense perspective, it has become difficult to find a single function, application, plan, or asset not enabled or impacted by the use of IT. Information technology has become so paramount that the President has made the operation and defense of the U.S. military's portion of IT infrastructure a mission for military leadership at the highest levels.

U.S. Strategic Command's subordinate or subunified command, U.S. Cyber Command (USCYBERCOM), is the evolutionary result of IT advancement, dependence, and protection. The USCYBERCOM mission statement, in part, directs the command to "plan, coordinate, synchronize and conduct activities to operate and defend DoD information networks and conduct full spectrum military cyberspace operations."[14] This is a daunting task given the reliance on systems and systems of systems and the efforts to exploit these systems by adversaries. This mission statement does imply a defense of networks and architectures, regardless of specific hostile agents. However, the current focus seems to have an anti-hacker (that is, human, nation-state, terror group) fixation. It does not, from a practical perspective, focus on artificial agent activities explicitly.

IT has allowed us to become more intelligent. At a minimum, it has enabled the diffusion of knowledge at paces never imagined. However, IT has also exposed us to real dangers such as personal financial or identity ruin or cyberspace attacks on our industrial control systems. It is plausible that a sufficiently resourced, goal-driven agent would leverage this technology to achieve its goal(s)—regardless of humankind's inevitable dissent.

Review of the Literature

Stephen Omohundro, Ph.D. in physics and mathematics and founder of Self-Aware Systems, a think tank for analyzing intelligent systems, has written extensively on the dangers inherent in any autonomous system. He states, "Autonomous systems have the potential to create tremendous benefits for humanity . . . but they may also cause harm by acting in ways not anticipated by their designers. Such systems are capable of surprising their designers and behaving in unexpected ways."[15] Omohundro outlines basic AI drives inherent in a goal-driven agent: "Without special precautions, it will resist being turned off, will try to break into other machines and make copies of itself, and will try to acquire resources without regard for anyone else's safety . . . not because they were programmed in at the start, but because of the intrinsic nature of goal driven systems."[16] Omohundro's four basic AI drives are:

- Efficiency: This drive will lead to improved procedures for computational and physical tasks.
- Resource loss avoidance: This drive will prevent passive losses and prevent outside agents from taking resources.
- Resource acquisition: This drive may involve exploration, trading, or stealing.
- Increasing utility: This drive would be to search for new behaviors to meet desired goals.

Other examples of the basic drives would be:

- Efficiency: the shutting down of systems not currently necessary for goal achievement (for example, a tertiary power grid—from the AI's perspective—but a primary source for humans)
- Resource loss avoidance: a protection of assets currently needed or determined to be required in the future (for example, active protection of servers by automated locking and monitoring of physical barriers)
- Resource acquisition: financial market manipulation to gain fungible assets for later satellite access (for example, leasing of bandwidth for more efficient communications through extant online tools)
- Increasing utility: game-playing, modeling, and simulation-running to determine best approaches to achieving goal(s).

Nick Bostrom, Ph.D. in economics and director of the Oxford Martin Programme on the Impacts of Future Technology with Oxford University, states, "Since the superintelligence may become unstoppably powerful because of its intellectual superiority and the technologies it could develop, it is crucial that it be provided with human-friendly motivations."[17] Bostrom also discusses a potential for accelerating or retarding AI, AGI, and ASI development from a policy perspective. However, given the motivators already discussed in this article, an impeding of AI research and development would be problematic for a nation to undertake unilaterally and would almost certainly require more than policy statements.

CHIMP, from Tartan Rescue Team, placed third in the DARPA Robotics Challenge Trials 2013 (DARPA)

Eliezer Yudkowsky, co-founder of the Machine Intelligence Research Institute, states in a chapter of *Global Catastrophic Risks*, "It is far more difficult to write about global risks of Artificial Intelligence than about cognitive biases. Cognitive biases are settled science; one need simply quote the literature. Artificial Intelligence is not settled science; it belongs to the frontier, not to the textbook."[18] This exemplifies the revolutionary leaps that seem possible considering the rate of technological advances (Moore's Law) and the motivations of a potentially unknowable number of developers. Yudkowsky goes on to emphasize the dangers of anthropomorphic bias concerning potential risks associated with AI:

Humans evolved to model other humans—to compete against and cooperate with our own conspecifics. It was a reliable property of the ancestral environment. . . . We evolved to understand our fellow humans empathically, by placing ourselves in their shoes. . . . If you deal with any other kind of optimization process . . . then anthropomorphism is flypaper for unwary scientists.[19]

James Barrat, an author and documentarian with National Geographic, Discovery, and PBS, states, "Intelligence, not charm or beauty, is the special power that enables humans to dominate Earth. Now, propelled by a powerful economic wind, scientists are developing intelligent machines. We must develop a science for understanding and coexisting with smart, even superintelligent machines. If we fail . . . we'll have to rely on the kindness of machines to survive."[20] Barrat's "busy child" analogy depicts a developed AI system that rapidly consumes information and surpasses human-level intelligence to become an ASI. Its human overseers correctly disconnect the busy child from the Internet and networks because:

Legged Squad Support System (LS3) robots will go through same terrain as human squad without hindering mission (DARPA)

once it is self-aware, it will go to great lengths to fulfill whatever goals it's programmed to fulfill, and to avoid failure. [It] *will want access to energy in whatever form is most useful to it, whether actual kilowatts of energy or cash or something else it can exchange for resources. It will want to improve itself because that will increase the likelihood that it will fulfill its goals. Most of all, it will not want to be turned off or destroyed, which would make goal fulfillment impossible. Therefore, AI theorists anticipate our ASI will seek to expand out of the secure facility that contains it to have greater access to resources with which to protect and improve itself.*[21]

Current Efforts in AI and Autonomous Agents

The Defense Advanced Research Projects Agency (DARPA), established to "prevent strategic surprise from negatively impacting U.S. national security and create strategic surprise for U.S. adversaries by maintaining the technological superiority of the U.S. military,"[22] is the preeminent technological anticipator and solicitor with a defense focus. DARPA has incentivized such things as automated ground vehicle technology, robotics maturation, and cyber self-defense through a competition format, with prizes awarded in the millions of dollars. For example, the 2004 Grand Challenge offered a $1 million prize to the team whose automated, unmanned vehicle was able to traverse a difficult 142-mile desert trek in a specified amount of time. Although no team completed the course (and no prize money was awarded) in the 2004 event, the 2005 Grand Challenge saw a team from Stanford University not only claim the $2 million prize,[23] but also defeat the course in just 6 hours.

In March 2014, DARPA solicited entrants for its inaugural Cyber Grand Challenge to "enable DARPA to test and evaluate fully automated systems that perform software security reasoning and analysis."[24]

DARPA's Robotics Challenge (DRC) aims for contestant robots to demonstrate, in part, "Partial autonomy in task-level decision-making based on operator commands and sensor inputs."[25] The competition drew numerous contestants, both DARPA-funded and self-funded, with nine DARPA-funded candidates and two self-funded candidates still remaining in the competition as of May 2014. Interestingly, the highest scoring team from DARPA's December 2013 DRC trials, Team SCHAFT, has been acquired by Google, Inc., and has elected to self-fund.[26]

Much less information is available about private company endeavors into AI, automated agents, automated systems, or AGI/ASI. Google, however, should be considered a leader in at least the pursuit of the highest technologies. With its recent purchase of robotics companies such as Boston Dynamics[27]

and Team SCHAFT, its Google Glasses, the driverless car, and AI company DeepMind, Google's direction seems to point toward an AI or AI-like capability. An additional note and perhaps key indication of Google's AI focus was the hiring of Ray Kurzweil, noted futurist, AI authority, and author of *The Singularity Is Near: When Humans Transcend Biology*.

Douglas Derrick has conducted live autonomous agent tests using his Special Purpose, Embodied, Conversational Intelligence with Environmental Sensors (SPECIES) agent. His agent-based system builds on existing communications models and theories and interacts directly with humans to achieve the goal of essentially discerning human deceit. Dr. Derrick writes of the "natural progression of human interactions with machines,"[28] where systems (machines) are being developed or will be developed that may assess human states, to include whether or not the human is being truthful. Derrick's prototype SPECIES agent was built to interview potential international border crossers as they passed through security lines. His team conducted a field study using his SPECIES agent with U.S. Customs and the Federal Bureau of Investigation, where the SPECIES agent was to "evaluate physiological and behavioral deviations from group and individual baselines in order to discriminate between truthful and deceitful communication."[29] Derrick's work demonstrated, to some degree, a goal-driven agent's ability not only to interact with humans, but also to engage in a level of persuasiveness while interacting with humans.

Conclusion and Recommendation

This article is not intended to be alarmist. On the contrary, it should serve as an initial call for engagement and collaboration. AI, AGI, ASI or the technological singularity may never come to fruition. Perhaps machines will plateau at or near where we are currently positioned in terms of nonhuman intelligence. Or perhaps a friendly version of AI will be developed and "decide" serving humankind obliges its own self-interests. Either of these possibilities is within the realm of reason. Yet considering the incentives and the demonstrated advances in a relatively short period of time, a more pragmatic view would suggest an approach more akin to cautious optimism. Once the possibility of goal-driven agents is considered, it does become easier to envision impacts being realized at some level.

This article recommends that the Department of Defense establish a working group concentrating on defense and industry engagement pertaining to goal-driven agents and artificial intelligence. This working group should have a basic charter to research current U.S. and partner efforts in AI and provide formal feedback to defense officials and policymakers. Similarly, there must be a call for research into codifying ethics and moral behavior into machine logic. The philosophical considerations that help define human morality must be able to be codified and expressed to nonhuman intelligences. Research should be conducted to temper goal-driven, autonomous agents with ethics. Basic research must be undertaken into what this codification and expression could be. JFQ

Notes

[1] Peter H. Diamandis and Steven Kotler, *Abundance: The Future Is Better Than You Think* (New York: Free Press, 2012).

[2] Ray Kurzweil, *The Singularity Is Near: When Humans Transcend Biology* (New York: Penguin, 2005).

[3] James Barrat, *Our Final Invention: Artificial Intelligence and the End of the Human Era* (New York: Macmillan, 2013).

[4] Vernor Vinge, "The Coming Technological Singularity: How to Survive in the Post-Human Era," 1993, available at <www-rohan.sdsu.edu/faculty/vinge/misc/singularity.html>.

[5] Barrat.

[6] Ibid.

[7] Michael Wooldridge and Nicholas R. Jennings, "Agent Theories, Architectures, and Languages: A Survey," in *ECAI-94 Proceedings of the Workshop on Agent Theories, Architectures, and Languages on Intelligent Agents* (New York: Springer-Verlag, 1995), 1–39.

[8] Stephen Omohundro, "Autonomous Technology and the Greater Human Good," *Journal of Experimental and Theoretical Artificial Intelligence* 26, no. 3 (2014), 303–315.

[9] David R.S. Cumming, Stephen B. Furber, and Douglas J. Paul, "Beyond Moore's Law," *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 372 (2012), available at <www.ncbi.nlm.nih.gov/pmc/articles/PMC3928907/>.

[10] Jeremy Hsu, "Quantum Bit Stored for Record 39 Minutes at Room Temperature," *Institute of Electrical and Electronics Engineers Spectrum*, November 15, 2013, available at <http://spectrum.ieee.org/tech-talk/computing/hardware/quantum-bit-stored-for-record-39-minutes-at-room-temperature>.

[11] Saul David, *Military Blunders* (London: Constable & Robinson, 2014).

[12] Kurzweil.

[13] Jonathan Harris, *A Chronology of the Byzantine Empire*, ed. Timothy Venning (New York: Palgrave Macmillan, 2006).

[14] United States Strategic Command, "U.S. Cyber Command," available at <www.stratcom.mil/factsheets/2/Cyber_Command/>.

[15] Omohundro, "Autonomous Technology."

[16] Stephen M. Omohundro, "The Basic AI Drives," *Frontiers in Artificial Intelligence and Applications* 171 (2008), 483.

[17] Nick Bostrom, "Ethical Issues in Advanced Artificial Intelligence," *Science Fiction and Philosophy: From Time Travel to Superintelligence*, ed. Susan Schneider (Malden, MA: Blackwell Publishing, 2009), 277–284.

[18] Eliezer Yudkowsky, "Artificial Intelligence as a Positive and Negative Factor in Global Risk," in *Global Catastrophic Risks*, ed. Nick Bostrom and Milan Cirkovic (Oxford: Oxford University Press, 2008).

[19] Ibid.

[20] James Barrat, "About the Author," available at <www.jamesbarrat.com/author/>.

[21] Barrat, *Our Final Invention*.

[22] Defense Advanced Research Projects Agency (DARPA), "Our Work," available at <www.darpa.mil/our_work/>.

[23] DARPA, "Robotics Challenge," available at <www.theroboticschallenge.org/about>.

[24] DARPA, "Solicitation," available at <www.fbo.gov/index?s=opportunity&mode=form&id=328908705d152877b2022b72735c266f&tab=core&_cview=0>.

[25] DARPA, "Robotics Challenge."

[26] Ibid.

[27] Jonathan Berr, "Google Buys 8 Robotics Companies in 6 Months: Why?" CBS News, December 16, 2013, available at <www.cbsnews.com/news/google-buys-8-robotics-companies-in-6-months-why/>.

[28] Douglas C. Derrick, Jeffrey L. Jenkins, and Jay F. Nunamaker, Jr., "Design Principles for Special Purpose, Embodied, Conversational Intelligence with Environmental Sensors (SPECIES) Agents," *AIS Transactions on Human-Computer Interaction* 3, no. 2 (2011), 62–81.

[29] Ibid.