Assessing Strategic Effects of Artificial Intelligence

Annotated Bibliography

September 2018

CGSR
Center for Global Security Research

LAWRENCE LIVERMORE NATIONAL LABORATORY

Annotated Bibliography for Workshop:

Assessing Strategic Effects of Artificial Intelligence

Center for Global Security Research
Lawrence Livermore National Laboratory

September 20-21, 2018

Rafael Loss

This workshop examines the implications of advances in artificial intelligence (AI) on international security, discussing the question of whether we will have to rethink, by the end of the next decade, how we practice nuclear deterrence and ensure strategic stability. Hosted by the Center for Global Security Research (CGSR) at Lawrence Livermore National Laboratory (LLNL), the workshop is part of a collaboration between CGSR and Technology for Global Security (T4GS) to engage policy-makers, scholars, technical experts, and the private sector to address emerging challenges in AI and related issues. The workshop engages with the current literature pointing to the risks and opportunities presented by AI and attempts to assess which are legitimate and which might be exaggerated.

The list of readings below provides overview and background for the topics covered in the workshop. While some references are to books or materials not available online, where possible references include links to accessible versions of articles.

Panel 1: Revisiting Strategic Stability and Recent Developments in Artificial Intelligence

Key questions:

- How do we, our allies, and adversaries define strategic stability in a time of renewed great power competition?
- What aspects of AI are most relevant for strategic stability? Will they strengthen or undermine deterrence?
- How rapidly should we expect AI-related technologies to proliferate? Will certain technologies spread more evenly than others?
- How might AI interact with other technological innovations, such as advances in quantum computing, in affecting deterrence dynamics?

James Acton (2013), "Reclaiming Strategic Stability," in Elbridge A. Colby and Michael S. Gerson (eds.), *Strategic Stability: Contending Interpretations*, Carlisle Barracks, PA: U.S. Army War College Press, pp. 117-146: https://carnegieendowment.org/files/Reclaiming_Strategic_Stability.pdf.

> Acton introduces and discusses several interpretations of the term "strategic stability". These very words, he argues, give the impression of a broad concept that pretends to predict whether and how states can enjoy stable relations. He concludes that crafting the nuclear strategy involves trade-offs between various desirable goals—including deterrence effectiveness, cost effectiveness, bureaucratic feasibility, domestic politics, and alliance politics to name but five—and it is by defining strategic stability most narrowly that we are most likely to set up a sensible debate about what those trade-offs should be.

Jürgen Altmann and Frank Sauer (2017), "Autonomous Weapon Systems and Strategic Stability," *Survival* 59(5), pp. 117-142: https://doi.org/10.1080/00396338.2017.1375263.

> Although not yet operational, decades of military research and development, as well as the growing technological overlap between the rapidly expanding commercial use of artificial intelligence (AI) and robotics, and the accelerating 'spin-in' of these technologies into the military realm, make autonomy in weapon systems a possibility for the very near future. By drawing on Cold War lessons and extrapolating insights from the current military use of remotely controlled unmanned systems, Altman and Sauer argue that autonomous weapon systems (AWS) are prone to proliferation and bound to foment an arms race resulting in increased crisis instability and escalation risks.

Rebecca Crootof (2018), "Autonomous Weapon Systems and the Limits of Analogy," *Harvard National Security Journal* 9(2): http://harvardnsj.org/wp-content/uploads/2018/06/2_Crootof_LimitsOfAnalogy_06.08.18.pdf.

> Crootof argues that analogies often used to describe AWS—either as more independent versions of weapons already in use or as humanoid robotic soldier—misrepresent their legally salient traits and limit our conception of how AWS might develop. She suggests

that because analogical reasoning fails in this realm, new supplemental law is needed to appropriately and effectively regulate AWS.

Michael C. Horowitz (2018), "Artificial Intelligence, International Competition, and the Balance of Power," *Texas National Security Review* 1(3), pp. 36-57: https://tnsr.org/2018/05/artificial-intelligence-international-competition-and-the-balance-of-power/.

Horowitz evaluates how developments in AI—advanced, narrow applications in particular—are poised to influence military power and international politics. He describes how AI more closely resembles "enabling" technologies such as the combustion engine or electricity than a specific weapon. He then explores the possibility that key drivers of AI development in the private sector could cause the rapid diffusion of military applications of AI, limiting first-mover advantages for innovators. Alternatively, it is also possible that military uses of AI will be harder to develop based on private-sector AI technologies than many expect, generating more potential first-mover advantages for existing powers such as China and the United States, as well as larger consequences for relative power if a country fails to adapt.

Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum (2015), "Human-Level Concept Learning Through Probabilistic Program Induction," *Science* 350(6266), pp.1332-1338: http://science.sciencemag.org/content/sci/350/6266/1332.full.pdf.

Machine Learning has yet to incorporate two human-level capabilities: learning new concepts from little data and building rich representations from data. This concept learning allows humans to build higher level ideas by combining lower- and high-level ideas; explain the driving mechanisms of observed events; and learn from past experience how to associate concepts already understood to an unfamiliar event. Concept learning allows a method to take experience from one domain and apply it to another, a capability that even the most impressive AI's—such as DeepBlue and AlphaGo—do not possess in robust form. The authors seek to incorporate these three capabilities in a method (the Bayesian Learning Program) that can not only classify and generate written characters, but also decompose the characters and produce new relationships among the decomposed parts.

Paul Scharre and Michael C. Horowitz (2018), "Artificial Intelligence: What Every Policymaker Needs to Know," Center for a New American Security: https://www.cnas.org/publications/reports/artificial-intelligence-what-every-policymaker-needs-to-know.

Scharre and Horowitz introduce the impact of advances in AI for national security and an initial exploration into how AI may change the international security environment. They argue that although the AI revolution might take decades to unfold and evolve in surprising ways, the changes that AI and machine learning will bring to global security will require policy-makers' preparation. Toward that end, they describe what "AI" is and what it is good for and discuss related safety concerns and potential vulnerabilities of AI.

David H. Wolpert (1996), "The Lack of A Priori Distinctions Between Learning Algorithms," *Neural Computation* 8(7), pp. 1341-1390: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.390.9412&rep=rep1&type=pdf. [1]

> Wolpert introduces the "no free lunch" (NFL) theorems of machine learning. As papers introduced new methods showing success of various datasets in the maturing years of machine learning research, the question arose, "Is there a particular method that performs better than all other methods overall?" For example, are neural networks always the best choice for an AI task? If there were such a method, the time-consuming phases of searching for the right algorithm for a certain problem would be eliminated. Wolpert's seminal work proves the counterintuitive result that, across all AI tasks, no one algorithm outperforms any other algorithm. This means that success of an AI method in one domain does not necessary translate to other domains. Hence, one still needs to invest the time to test several methods for a new problem in order to identify the best method.

Panel 2: Comparing AI Adoption and Integration Across Countries

Key questions:

- Does AI empower different actors in different ways? Do different actors follow different approaches in developing and adopting AI technologies?
- Does AI increase the potential for asymmetrical conflict and strategies?
- Are non-state actors leveraging AI technologies likely to serve as spoilers in this new strategic environment?
- What drives cooperation and/or competition among various actors? How could incentives be altered to enhance strategic stability?
- Is there an AI arms race? What does it look like and what does it take to win it?

Samuel Bendett and Elsa B. Kania (2018), "Chinese and Russian Defense Innovation, with American Characteristics? Military Innovation, Commercial Technologies, and Great Power Competition," *The Strategy Bridge*: https://thestrategybridge.org/the-bridge/2018/8/2/chinese-and-russian-defense-innovation-with-american-characteristics-military-innovation-commercial-technologies-and-great-power-competition.

> Bendett and Kania observe that as China and Russia seek to keep pace with and overtake U.S. defense innovation initiatives, their approaches are mimicking and converging with certain elements of the traditional U.S. approach. If successful, they argue, these overtures towards Chinese and Russian defense innovation with American characteristics could enhance their respective capabilities to experiment with and operationalize new

[1] See Pedro Domingos (2012), "A Few Useful Things to Know about Machine Learning," *Communications of the ACM* 55(10), pp. 78-87: https://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf for an explanation of twelve fundamental principles of machine learning. These principles describe the price that must be paid to effectively learn from data. The trade-offs he outlines, as well as the manual tweaking of algorithms and processing data, suggest significant human input in the process. The resulting "black art" gained from experience represented in academic terms; much less automated. These principles, codified by Domingos at the advent of Big Data and Deep Learning as we know it, have proven to be immutable – impervious to advances in methods and memory.

capabilities. They conclude that while the United States still has a significant advantage in emerging technologies and their applications, Russian and Chinese efforts could nonetheless produce results that may yet disrupt today's techno-strategic competition among these great powers.

Stephan De Spiegeleire, Matthijs Maas, and Tim Sweijs (2017), "Artificial Intelligence and the Future of Defense: Strategic Implications for Small- and Medium-Sized Force Providers," The Hague Center for Strategic Studies: https://hcss.nl/sites/default/files/files/reports/Artificial%20Intelligence%20and%20the%20Future%20of%20Defense.pdf.

De Spiegeleire, Maas, and Sweijs suggest that AI may profoundly transform defense and security as new incarnations of 'armed force' start doing different things in novel ways. They juxtapose three different layers (/generations) of AI against four different layers (and possibly generations) of Armed Force, focusing on small- and medium-sized force providers like the Netherlands, and develop recommendations on how these countries can responsibly harness developments in AI to achieve more sustainable defense and security solutions.

Jill Dougherty and Molly Jay (2018), "Russia Tries to Get Smart about Artificial Intelligence," *The Wilson Quarterly*: https://wilsonquarterly.com/quarterly/living-with-artificial-intelligence/russia-tries-to-get-smart-about-artificial-intelligence/.

Dougherty and Jay examine how Russia plans to compete with the great powers in the AI field, i.e., China and the United States, building off of Russian President Vladimir Putin's recent statement that "[w]hoever becomes the leader in this sphere will become the ruler of the world." They describe the Russian ambitions to facilitate AI innovation through its private sector and to leverage advances both to undermine Western democracy and exploit failures and weaknesses in Western institutions and to gain military advantages on the battlefields of the future. To that end, they argue, Russia is marrying U.S.-style public-private cooperation with China's heavy government control.

Michael C. Horowitz et al. (2018), "Strategic Competition in an Era of Artificial Intelligence," Center for a New American Security: https://www.cnas.org/publications/reports/strategic-competition-in-an-era-of-artificial-intelligence.

Horowitz et al. argue that given the breadth of AI, with its ability to influence defense, diplomacy, intelligence, economic competitiveness, social stability, and the information environment, falling behind in AI development and implementation would present a risk for U.S. global economic and military leadership. Yet leadership in AI will not be just about the technology itself, but about how societies manage the technology. Thus, strategies for leveraging the technology will become essential. Consequently, the authors recommend a broad series of actions for the U.S. government, in partnership with the private sector, to prepare for the challenges posed by advances in AI, including on R&D, funding, acquisition, metrics, education, immigration, and norms.

Elsa B. Kania (2017), "Battlefield Singularity: Artificial Intelligence, Military Revolution, and China's Future Military Power," Center for a New American Security: https://www.cnas.org/publications/reports/battlefield-singularity-artificial-intelligence-military-revolution-and-chinas-future-military-power.

> Kania explores China's strategy for developing and implementing AI technology for military applications. Drawing on open-source Chinese-language documents, she explains Chinese strategic thinking on AI and specific military applications that Chinese leaders envision. In response, she argues, the United States must work to formulate a long-term, whole-of-nation strategy to support critical determinants of national competitiveness in AI. While it is critical to sustain and build upon the current U.S. competitive advantage in human capital through formulating policies to educate and attract top talent, the U.S. military should also prepare for a future in which the United States may no longer possess technological predominance.

Lora Saalman (2018), "Fear of False Negatives: AI and China's Nuclear Posture," *Bulletin of the Atomic Scientists*: https://thebulletin.org/landing_article/fear-of-false-negatives-ai-and-chinas-nuclear-posture/.

> Saalman argues that Chinese analysts—much more so than their U.S. counterparts—are preoccupied with the inability of their systems to identify, much less to counter, a stealthy and prompt precision strike. China's activities to compensate for these shortcomings—integration of swarms that could be used to confront dual-capable US platforms; alleged internal discussions about launch-on-warning for missiles; hedging on conventional versus nuclear payloads; and research into greater application of AI and autonomy in prompt and high-precision systems—she posits, suggest that its concerns about false negatives could create changes in command and control operations that run the risk of producing a false positive.

Tom Simonite (2017), "For Superpowers, Artificial Intelligence Fuels New Global Arms Race," *Wired*: https://www.wired.com/story/for-superpowers-artificial-intelligence-fuels-new-global-arms-race/.

Elsa B. Kania (2018), "The Pursuit of AI Is More Than an Arms Race," *Defense One*: https://www.defenseone.com/ideas/2018/04/pursuit-ai-more-arms-race/147579/.

> The two articles by Simonite and Kania debate the merit of framing the competition in AI advances among the great powers as an arms race. While Simonite argues that the United States, Russia, and China are deeply entrenched in an arms race for ever better algorithms, likening it to the competitions for superiority in nuclear bombs and precision-guided weapons, Kania maintains that the arms race analogy falls short in grasping the complexity of the current competition.

Panel 3: Artificial Intelligence and Deterrence Across Domains

Key questions:

- How might AI affect the key components of strategic deterrence, such as C4ISR, the weapons complex, second strike capabilities, and space-based systems?
- How might AI technologies impact deterrence strategies across domains? What is the relationship between AI and integrated/complex deterrence?
- Do developments in AI technologies shift thinking about critical national security infrastructure? Do they shift the requirements for engagement between the private and public sectors?

Paul Bracken (2016), "The Cyber Threat to Nuclear Stability," *Orbis* 60(2), pp. 188-203: https://doi.org/10.1016/j.orbis.2016.02.002.

> Bracken argues that cyber war technologies are increasingly spilling over into precision strike and nuclear mission areas, transforming deterrence and arms race stability and leading to other significant changes. The driver behind this is a combination of long standing problems with mobile missiles along with new technologies not usually factored into strategic assessments: big data analytics, computer vision, and related information systems. When combined with drones and precision strike, the hunt for mobile missiles is becoming faster, cheaper, and better. He concludes that the implications of his findings vary by country, but will shape major power nuclear modernization, crisis stability among secondary powers, and conventional attack of nuclear deterrents.

Erik Gartzke and Jon R. Lindsay (2017), "Thermonuclear Cyberwar," *Journal of Cybersecurity* 3(1): pp. 37-48: https://doi.org/10.1093/cybsec/tyw017.

> When combined, according to Gartzke and Lindsay, the warfighting advantages of cyber operations become dangerous liabilities for nuclear deterrence. Increased uncertainty about the nuclear/cyber balance of power raises the risk of miscalculation during a brinksmanship crisis. Accordingly, strategic stability in nuclear dyads is, in part, a function of relative offensive and defensive cyber capacity. To reduce the risk of crisis miscalculation, they propose that states should improve rather than degrade mutual understanding of their nuclear deterrents.

Edward Geist and Andrew J. Lohn (2018), "How Might AI Affect the Risk of Nuclear War?" RAND Corporation: https://www.rand.org/pubs/perspectives/PE296.html.

> Geist and Lohn, recapitulating the insights gained from a series RAND workshops, ask whether advances in AI might upset the nuclear strategic balance, and, if so, for better or for worse. They examine the impact of advanced computing on nuclear security through 2040, describing the types of anticipated concerns and benefits through two illustrative examples: AI for detection and for tracking and targeting and AI as a trusted adviser in escalation decisions. In view of the capabilities that AI may be expected to enable and how adversaries may perceive them, they conclude that AI has the potential to exacerbate

emerging challenges to nuclear strategic stability by the year 2040 even with only modest rates of technical progress.

Keir A. Lieber and Daryl G. Press (2017), "The New Era of Counterforce: Technological Change and the Future of Nuclear Deterrence," *International Security* 41(4), pp. 9-49: https://www.mitpressjournals.org/doi/full/10.1162/ISEC_a_00273.

For much of the nuclear age, "counterforce" disarming attacks—those aimed at eliminating an opponent's nuclear forces—were nearly impossible because of the ability of potential victims to hide and protect their weapons. Lieber and Press argue that these two key approaches that countries have relied on to ensure arsenal survivability have been undercut by leaps in weapons accuracy and a revolution in remote sensing. This new era of counterforce then challenges the basis for confidence in contemporary deterrence stability and sheds light on one of the enduring puzzles of the nuclear era: why international security competition has endured in the shadow of the nuclear revolution.

Beyza Unal and Patricia Lewis (2018), "Cybersecurity of Nuclear Weapons Systems: Threats, Vulnerabilities and Consequences," Chatham House: https://www.chathamhouse.org/publication/cybersecurity-nuclear-weapons-systems-threats-vulnerabilities-and-consequences.

According to Unal and Lewis, many of the assumptions on which current nuclear strategies are based pre-date the current widespread use of digital technology in nuclear command, control and communication (NC3) systems, leaving them vulnerable to cyberattacks. During peacetime, offensive cyber activities could create a dilemma for a state as it may not know whether its systems have been the subject of a cyberattack. At times of heightened tension, cyberattacks on nuclear weapons systems could cause an escalation, which could result in nuclear use. Consequently, the authors argue that nuclear weapons states should incorporate cyber risk reduction measures in NC3 systems.

Panel 4: Operationalizing Automation and Artificial Intelligence for the Battlefield

Key Questions:

- How might AI change the character of conflict, its initiation, escalation and termination? How might it affect the offense-defense balance in major power and regional conflict as well as in campaigns involving non-state actors?
- Could tactical battlefield AI aggregate up and have strategic effects?
- How might non-state actors leverage AI technologies to threaten state actors? What are the counter-AI/adversarial AI tools needed to mitigate these risks?

Kelsey Atherton (2018), "Targeting the Future of the DoD's Controversial Project Maven Initiative," *C4ISRNET*: https://www.c4isrnet.com/it-networks/2018/07/27/targeting-the-future-of-the-dods-controversial-project-maven-initiative/.

Atherton recounts how within a year of the establishment of DoD's "Algorithmic Warfare Cross-Functional Team" or "Project Maven"—a program which focuses on computer vision to autonomously extracts objects of interest from moving or still imagery—the details of Google's role in that program disseminated internally among its employees and was then shared with the public, calling into question the specific rationale of the task and the greater question of how the tech community should go about building algorithms for war, if at all.

Mary L. Cummings (2004), "Creating Moral Buffers in Weapon Control Interface Design," *IEEE Technology and Society Magazine* 23(3), pp. 28-33: https://ieeexplore.ieee.org/document/1337888/.

Cummings examines the ethical and social issues surrounding the design of human-computer interfaces that are designed for control of highly autonomous weapons systems. She argues that the implementation of AWS means that commanders will be asked to make near-instantaneous decisions about networks of AWS that can easily generate more information than a human can process. When directing such weapons through a human-computer interface that provides a virtual user-friendly world, moral buffers can be more easily created as a consequence of both psychological distancing and compartmentalization. This would allow people to ethically distance themselves from their actions and diminish a sense of accountability and responsibility.

Mark Gilchrist (2018), "Emergent Technology, Military Advantage, and the Character of Future War," *The Strategy Bridge*: https://thestrategybridge.org/the-bridge/2018/7/26/emergent-technology-military-advantage-and-the-character-of-future-war.

Gilchrist examines the integration challenge that limits the military potential of available technology. He looks specifically at why militaries should be cautious about the role AI and autonomous systems are expected to play in future warfare. AI has many potential applications; however, this article focuses on the areas that will ultimately determine how a military response is generated. The delivery of the right information, to the right decision-maker, at the right time, and the ability for all elements of the force to react to that decision is fundamental to success in battle. Therefore, the author focuses on command and control systems and the implications of AI for the already stressed network architectures when considering how to generate lethal effect in the information age.

Andrew Ilachinski (2017), "AI, Robots, and Swarms: Issues, Questions, and Recommended Studies," CNA: https://www.cna.org/CNA_files/PDF/DRM-2017-U-014796-Final.pdf.

Ilachinski cites recent advances in AI research by Google, Microsoft, Facebook and other technology companies and identifies the challenges that DoD faces as it embraces AI, robot and swarm technologies. He identifies four key technical gaps that must be addressed to overcome the challenges: 1) The fundamental mismatch between the accelerating pace of technology and research in the commercial and academic research communities and the timescale of DoD's existing acquisition process; 2) The underappreciation of the unpredictable nature of autonomous systems, particularly when operating in a dynamic environment such as a battlefield; 3) the lack of a universal conceptual framework for autonomy that can be used to anchor theoretical discussions and serve as a frame-of-reference for understanding how theory, design, implementation, testing and operations are all interrelated; and 4) the fact that DoD's current acquisition process does not allow for timely introduction of "mission ready" AI/autonomy and the disconnect between system design and development of concepts of operations.

Christian Szegedy et al. (2014), "Intriguing Properties of Neural Networks," arXiv:1312.6199v4, *arXiv*: https://arxiv.org/abs/1312.6199. [2]

This paper set into motion the burgeoning field of Adversarial AI. Szegedy et al. find two theoretical concerns: the inherent non-interpretability of the thought process of these "black boxes" was more difficult to disentangle than previously thought and that the models possessed large blind spots that could be exploited by an adversary. Previous research proposed methods of showing how different areas of the networks learned interpretable features. In other words, one could explain the thought process of the network. The authors show that this was not the case. The highly non-linear relationship learned by the network is not a composition of interpretable formulas, but one fully integrated unit. This structure leaves it open to "blind spots". Furthermore, these blind spots do not exist outside the range of the data that has been learned, but very close to data points already learned by the neural nets. This brittleness could allow an adversary to tweak a data point in such a manner that a human would not recognize the change, but the change would fool a neural net. Creating and identifying such "adversarial examples" is at the heart of Adversarial AI.

Panel 5: Ensuring Strategic Stability in the Age of Artificial Intelligence

Key Questions:

- What are U.S. priorities for taking advantage of AI to enhance deterrence?
- What are the main risks that AI poses for strategic stability?
- How might cooperation help steer AI to enhance, rather than undermine, national and international security?

Kareem Ayoub and Kenneth Payne (2016), "Strategy in the Age of Artificial Intelligence," *Journal of Strategic Studies* 39(5-6): pp. 793-819: https://www.tandfonline.com/doi/pdf/10.1080/01402390.2015.1088838.

[2] For an assessment of how to counter adversarial influence in data analytics see Philip Kegelmeyer et al. (2015), "Counter Adversarial Data Analytics," Sandia National Laboratories: http://www.sandia.gov/~wpk/pubs/publications/cada-full-uur.pdf. Nicolas Papernot et al. (2016), "The Limitations of Deep Learning in Adversarial Settings," presented at the 1st IEEE European Symposium on Security & Privacy in Saarbrucken, Germany: http://www.patrickmcdaniel.org/pubs/esp16.pdf develop a "Threat Model Taxonomy" that categorizes adversarial methods according to the extent of knowledge an adversary must have to fool a target AI, and the degree to which the example is deceiving.

Arguing that AI will profoundly impact existing power balances and the conduct of strategy, Ayoub and Payne review the psychological foundations of strategy and explore the ways in which AI will impact human decision-making. They then review current and evolving capabilities in 'narrow', modular AI that is optimized to perform in a particular environment and explore its military potential. Lastly, they look ahead to the more distant prospect of a general AI.

Rebecca Crootof and Frauke Renz (2017), "An Opportunity to Change the Conversation on Autonomous Weapon Systems," *Lawfare*: https://www.lawfareblog.com/opportunity-change-conversation-autonomous-weapon-systems.

Crootof and Reny examine the movement to prohibit fully autonomous weapons, arguing that the cancellation of the inaugural meeting of the Group of Governmental Experts on Lethal Autonomous Weapon Systems opens up an opportunity to set up a more productive working group to explore the perennial legal issues associated with AWS. Such a working group, they argue, could give the international community the chance to change the conversation and engage with the full range and depth of legal challenges raised by autonomous weapon systems, rather than remaining stuck in the simplified, binary ban debate.

Executive Office of the President (2016), "Preparing for the Future of Artificial Intelligence," National Science and Technology Council, Committee on Technology: https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf.

This NSTC report assess new opportunities for progress in areas such as health, education, energy, and the environment resulting from advances in AI. As a contribution toward preparing the United States for a future in which AI plays a growing role, the report surveys the current state of AI, its existing and potential applications, and the questions that are raised for society and public policy by progress in AI. The report also makes recommendations for specific further actions by Federal agencies and other actors, including on issues relating to international relations, cybersecurity, and defense.

Edward M. Geist (2016), "It's Already Too Late to Stop the AI Arms Race—We Must Manage It Instead," *Bulletin of the Atomic Scientists* 72(5), pp. 318-321: https://www.tandfonline.com/doi/pdf/10.1080/00963402.2016.1216672.

While an ongoing campaign argues that an agreement to ban autonomous weapons can forestall AI from becoming the next domain of military competition, Geist suggests that an AI arms race is already under way. He draws on the history of AI weaponization and arms control for other technologies to argue that AI and robotics researchers should cultivate a security culture to help manage the AI arms race. By monitoring ongoing developments in AI weapons technology and building the basis for informal "Track II" diplomacy, AI practitioners could begin building the foundation for future arms-control agreements.

Daniel S. Hoadley and Nathan J. Lucas (2018), "Artificial Intelligence and National Security," Congressional Research Service: https://fas.org/sgp/crs/natsec/R45178.pdf.

This CRS report outlines issues and questions regarding rapid advances in AI and their potentially significant implications for national security, with emphasis on how these relate to the role of the U.S. Congress. Questions raised include: What is the right balance of commercial and government funding for AI development?; How might Congress influence Defense Acquisition reform initiatives that ease military AI adaptation?; What changes, if any, are necessary in Congress and DoD to implement effective oversight of AI development?; What regulatory changes are necessary for military AI applications?; and what measures can be taken to protect AI from exploitation by international competitors and preserve a U.S. advantage in the field?

General Reading Recommendations:

Greg Allen and Taniel Chan (2017), "Artificial Intelligence and National Security," Belfer Center for Science and International Affairs: https://www.belfercenter.org/sites/default/files/files/publication/AI%20NatSec%20-%20final.pdf.

Stuart Armstrong, Kaj Sotala, and Seán Ó hÉigeartaigh (2012), "The Errors, Insights and Lessons of Famous AI Predictions—And What They Mean for the Future," *Journal of Experimental & Theoretical Artificial Intelligence* 26(3), pp. 317-342: https://doi.org/10.1080/0952813X.2014.895105.

Vincent Boulanin and Maaike Verbrugen (2017), "Mapping the Development of Autonomy in Weapon Systems," Stockholm International Peace Research Institute: https://www.sipri.org/publications/2017/other-publications/mapping-development-autonomy-weapon-systems.

Miles Brundage et al. (2018), "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation": https://maliciousaireport.com/.

Stephanie Carvin (2017), "Normal Autonomous Accidents: What Happens When Killer Robots Fail?" *Social Science Research Network*: http://dx.doi.org/10.2139/ssrn.3161446.

David Deutsch (2012), "Creative Blocks," *Aeon*: https://aeon.co/essays/how-close-are-we-to-creating-artificial-intelligence.

Peter Eckersley and Yomna Nasser (2017), "Measuring the Progress of AI Research," Electronic Frontier Foundation: https://eff.org/ai/metrics.

JASON (2017), "Perspectives on Research in Artificial Intelligence and Artificial General Intelligence Relevant to DoD," The MITRE Corporation: https://fas.org/irp/agency/dod/jason/ai-dod.pdf.

Lawrence Livermore National Laboratory (2018), "What is Machine Learning?" *YouTube*: https://www.youtube.com/watch?v=7MvEx2R8xug.

McKinsey Global Institute (2017), "Artificial Intelligence: The Next Digital Frontier?": https://www.mckinsey.com/mgi/overview/2017-in-review/whats-next-in-digital-and-ai/artificial-intelligence-the-next-digital-frontier.

Paul Scharre (2018), *Army of None: Autonomous Weapons and the Future of War*, New York, NY: W. W. Norton.

Brian K. Spears (2017), "Contemporary Machine Learning: A Guide for Practitioners in the Physical Sciences," arXiv:1712.08523v1, *arXiv*: https://arxiv.org/abs/1712.08523.

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. LLNL-TR-757281