Artificial Intelligence and Deterrence: Science, Theory and Practice

Alex S. Wilner, PhD
Assistant Professor of International Affairs
Norman Paterson School of International Affairs (NPSIA), Carleton University
CANADA

alex.wilner@carleton.ca

https://alexwilner.com/

ABSTRACT

While a consensus is forming among military experts, policymakers, and academics that Artificial Intelligence (AI) will prove useful for national security, defence, and intelligence purposes, no academic study has explored how AI will influence the logic, conceptualization, and practice of deterrence. Debates on AI in warfare are largely centered on the tactical use and misuse of the technology within autonomous weapons systems, and the associated risks AI may pose to the ethical use of force. No concomitant debate exists, however, as to the strategic and deterrent utility of AI in times of crisis, conflict, and war or in matters of cybersecurity. Nor has any country openly published a strategic document on the nexus between AI and deterrence. The dearth of knowledge is surprising given the expectation that the future of warfare will be autonomous. This paper will provide a comprehensive conceptual map of how AI influences both deterrence in theory and in practice. It does so by exploring the science of AI and by providing a synthesis of how states are approaching AI in warfare and deterrence.

1.0 INTRODUCTION

While a consensus is forming among military experts, policy makers, and academics, that Artificial Intelligence (AI) will prove useful for national security, defence, and intelligence purposes, very few academic studies have yet fully explored how AI will influence the logic, conceptualization, and practice of deterrence.[1] AI is a field of science that seeks to provide machines with human-like qualities in problem-solving. *Narrow* AI uses algorithms to complete a specific task, like learning to play Chess, or to recognize faces. *General* AI seeks to empower machines to solve any number of problems. AI includes a number of techniques, including, most notably, *machine learning*, which trains algorithms to identify regularities in reams of data, and *reinforcement learning*, in which a program, built with feedback mechanisms, is rewarded on the actions it carries out.

AI is expected to have far-reaching consequences in governance, human rights, politics, power, and warfare.[2] Russian President Vladimir Putin's recent assertion that AI "is the future," and that "whoever becomes the leader in this sphere will become the ruler of the world," suggests that an AI arms race may be upon us.[3] Missing from much of the hype and discussion on AI, however, is a theoretical and conceptual exploration of how the technology influences our understanding of strategic studies. Debates on AI in warfare largely revolve around the tactical use (and misuse) of the technology within autonomous weapons

[1] I would like to acknowledge and thank my Research Assistant, Jennifer O'Rourke, for contributing to earlier drafts of this paper.

[2] Alex Wilner, "Cybersecurity and its Discontents: Artificial Intelligence, the Internet of Things, and Digital Misinformation," *International Journal* 73:2, 2018.

[3] James Vincent, "Putin Says the Nation That Leads in AI 'will Be the Ruler of the World'," *The Verge*, September 4, 2017.

systems (sometimes referred to as lethal autonomous weapon systems (LAWS)), and the associated risks AI may pose to the ethical use of force. No concomitant debate exists, however, as to the deterrent utility of AI in times of crisis, conflict, and war or in matters of cybersecurity. Nor has any NATO country openly published a strategic document on the nexus between AI and deterrence. The dearth of knowledge is surprising given the expectation that the future of warfare will likely be autonomous.[4]

This paper will explore how AI influences deterrence across the domains of warfare. AI risks undermining deterrence in unique ways. For illustration, it may alter cost-benefit calculations by removing the fog of war, by superficially imposing rationality on political decisions, and by diminishing the human cost of military engagement. It may recalibrate the balance between offensive and defensive measures. It may shorten the distance between intelligence analysis, political decisions, and coercive action. It may augment the certainty and severity of punishment strategies, both in theatre and online. And it may altogether remove human emotions from the issuance and use of coercive threats. This article provides a conceptual map of how AI might influence deterrence theory and practice. It builds off the author's previous work on updating deterrence theory for non-traditional threats, like terrorism, violent radicalization, and cybersecurity.[5] The article is also based on the author's ongoing research project – *AI Deterrence* – which is funded by Canada's Department of National Defence (DND) and Defence Research and Development Canada (DRDC) through the Innovation for Defence Excellence and Security (IDEaS) program (2018-2019). As an exercise in speculation, this paper, and the larger IDEaS project from which it stems, will try to accomplish two overarching tasks: to synthesize the scientific literature on AI as it relates to crisis, conflict, and war, and to theorize how AI will influence the nature and utility of deterrence.

The paper is structed in four sections. It begins with a scientific overview of AI, providing a layman's review of the technology. Section two turns to a practical description of how AI is used in surveillance, data analytics, intelligence, and military planning. Next, the nascent theoretical literature on AI and defence is explored. Subsequent lessons for AI and deterrence are discussed in the conclusion.

2.0 ARTIFICIAL INTELLIGENCE: A SCIENTIFIC PRIMER

The amount of scientific literature on AI that explores its potential to industry, society, economics, and governance is staggering. And yet, actually defining AI remains problematic: The field is marked by different technological approaches that attempt to accomplish different and at times divergent scientific tasks.[6] In their seminal textbook on the subject, Stuart Russell – an AI pioneer who has gone on to publicly (and loudly) warn that the development of AI is as dangerous as the development and proliferation of nuclear weapons – and Peter Norvig, describe a number of AI taxonomies, including systems that think or act like humans (including those that can mimic human intelligence by passing the Turing test), and those that think or act rationally (in solving problems and behaving accordingly, via robotic or digital platforms).[7] Developing a singular definition that can be universally accepted among all of these disparate branches of inquiry is challenging. But doing so is also unnecessary for the purposes of this article. A broader definition of AI that lends itself to security studies should suffice.

[4] See, for instance, Chris Coker, *Future War*, (Polity: 2015); Paul Scharre, *Army of None: Autonomous Weapons and the Future of War*, (W.W. Norton: 2018); Ben Wittes and Gabriella Blum, *The Future of Violence: Robots and Germs, Hackers and Drones* (Basic Books: 2015).

[5] See, for instance, Alex Wilner, *Deterring Rational Fanatics*, (University of Pennsylvania Press: 2015); Andreas Wenger and Alex Wilner (eds.), *Deterring Terrorism: Theory and Practice*, (Stanford University Press 2012); Alex Wilner, "Cyber Deterrence and Critical-infrastructrue Protection," *Comparative Strategy* 36:4, 2017; Alex Wilner, US Cyber Deterrence: Practice Guiding Theory (review and resubmit *Journal of Strategic Studies*, Dec. 2018); Jerry Mark Long and Alex Wilner, "Delegitimizing al-Qaida: Defeating an 'Army whose Men Love Death'", *International Security* 39:1, 2014.

[6] Lee Spector, "Evolution of Artificial Intelligence," *Artificial Intelligence* 170:18, 2006.

[7] Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach* (3rd Edition) (Pearson, 2009). For Russell's warnings, see John Bohannon, "Fears of an AI Pioneer," *Science* 349, 2015.

To this end, AI is a field of science that attempts to provide machines with problem-solving, reasoning, and learning qualities akin to human qualities. Max Tegmark, the President of the Future of Life Institute – an organization dedicated to exploring the societal impacts of AI and other disruptive technologies – usefully divides AI into two broad categories: narrow (or weak) AI and general (or strong) AI (often referred to as Artificial General Intelligence, or AGI).[8]

Narrow AI is developed and used to accomplish a specific task, like playing poker or Go, an ancient Chinese board game, recognizing and categorizing objects, people, and speech from pictures and video content, driving vehicles autonomously, and accomplishing various other demands. In the past decade, narrow AI has proven exceptionally competent at accomplishing specific tasks, far outstripping human competitors along the way. These advancements are in large part a result of recent successes in machine learning, one of several approaches to achieving AI.

Machine learning relies on algorithms to sift through data in order to learn from and identify patterns within the data, and to then determine a prediction about new data as a result. For illustration, a deep neural network, a type of machine learning algorithm, will be trained on millions of pictures of human faces, in order to eventually predict whether a new picture includes a human face in it. "The magic of deep learning," writes Chris Meserole, "is that the algorithm learns to do all this on its own. The only thing a researcher does is feed the algorithm a bunch of images and specify a few key parameters … and the algorithm does the rest." Instead of painstakingly coding the software to accomplish the task, the machine is trained using the data and learns to accomplish the task autonomously.[9]

In practice, machine learning techniques have scored some recent resounding victories for narrow AI. In 2017, for example, Google's parent company Alphabet revealed that its AlphaGo Zero software (developed by DeepMind) taught itself, via reinforcement learning – which, in layman terms, trains a machine through trial and error – to become the world's most accomplished player of Go, a complex strategy game. While previous versions of the software were trained to play Go by watching thousands of recorded games played between human competitors, AlphaGo Zero learned from scratch, without being fed any human-derived data. Instead, equipped with the rules of the game, the AI played against itself, making random moves and learning from failure and success. In three days, the AI surpassed a 2016 version of itself that beat the sitting (human) Go world champion in four of five games. In 21 days, AlphaGo Zero reached a level of play matching its 2017 version that defeated 60 leading Go players. And in just 40 days, the AI surpassed all other versions of itself, becoming "the best Go player in the world."[10] The DeepMind team went on to develop a more generalized AlphaZero algorithm that uses self-play to learn other games. In 24 hours, AlphaZero achieved "a superhuman level of play in the games of chess and shogi (Japanese chess) as well as Go, and convincingly defeated a world-champion program in each case."[11] AI learns decisively, and quickly. The next step in using games to develop AI pairs machines against humans in StarCraft, a complex, real-time, computer game, which provides players with a far more nuanced and realistic military setting complete with logistics, infrastructrue, and various strategies.[12]

[8] Max Tegmark, "Benefits & Risks of Artificial Intelligence," Future of Life Institute, https://futureoflife.org/background/benefits-risks-of-artificial-intelligence/?cn-reloaded=1. See also, Paul Scharre and Michael Horowtiz, "Artificial Intelligence: What Every Policymaker needs to know," CNAS, June 2018.

[9] Chris Meserole, "What is Machine learning?" Brookings Institution, October 2018.

[10] DeepMind, "AlphaGo Zero: Learning from Scratch," October 2017; David Silver, et. al., "Mastering the Game of Go Without Human Knowledge," *Nature* 550 (2017).

[11] David Silver, et. al., "Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm," [https://arxiv.org/abs/1712.01815].

[12] Michal Certicky, "StarCraft AI Competitions, Bots, and Tournament Manager Software," IEEE Transactions on Games (early access), Nov. 2018.

Notwithstanding these recent achievements, an AI's success in one field or in one area – like playing games – does not necessarily translate into that AI's success in other fields. Narrow AI is limited by its task-specific programming. General AI, conversely, is meant to pick up where narrow AI ends, to empower a machine to learn, adapt, and solve any number of problems, much as a human would. For now, AGI is the stuff of science fiction, rather than of science. While several prominent laboratories and organizations are working on developing AGI, it does not currently exist. Even so, AGI and the related concept of "superintelligence" figure prominently in the discourse and literature on AI and security.[13] Russell's warnings against AI, in fact, are largely informed by a future landscape, perhaps still decades away, dominated by an AGI, a super-intelligent sentient and autonomous machine "that can learn from experience with humanlike breadth and surpass human performance in most cognitive tasks."[14] While there is a non-trivial likelihood that AGI will eventually be achieved, what is of greater immediate importance is exploring how advances in narrow AI – rather than general AI – will affect our strategic and military planning, including in deterrence. Skynet, of *Terminator* fame, can wait. This paper's primary objective is to explore the nexus between contemporary narrow AI and deterrence theory and practice.

3.0 AI IN INTELLIGENCE, SECURITY, AND DEFENCE: PRACTICAL APPLICATIONS

To date, Artificial Intelligence has proven useful across the national security spectrum, including in surveillance, data analysis, intelligence assessment, and defence. The following section provides a summary overview and illustrative cases of how AI technology has been applied to each of these areas.

3.1 Surveillance

AI has been extensively used to collect, assess, analyse, and disseminate data and intelligence. Contemporary AI is capable of information aggregation and analysis, facial, speech, and handwriting recognition, opinion analysis, gait recognition, and predictive behavioural analysis.[15] A few illustrations highlight recent advancements in these areas. In China, for instance, facial recognition software and CCTV footage is being used to arrest fugitives (in public), deter jaywalking, and altogether control the movements of entire communities deemed a threat to national security.[16] Gait analysis and speech recognition have also become far more accurate at identifying individuals; one study using AI gait analysis achieved a 99.3 percent accuracy rate.[17] Predictive policing has likewise been bolstered by AI and surveillance data (often in the form of ubiquitous, live CCTV footage). For instance, over 90 cities, including New York, Chicago, and Cape Town, South Africa, use an AI-powered system that can monitor audio feeds for gun shots, autonomously triangulating their location, and immediately alerting police and first responders.[18] China takes predictive policing one step further, using AI to conduct behavioural analysis on citizens to determine

[13] See for instance, Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford, 2014); Amir Husain, *The Sentient Machine*, (Simon & Schuster, 2017); University of Cambridge, Centre for the Study of Existential Risk, https://www.cser.ac.uk/research/risks-from-artificial-intelligence/

[14] Stuart Russell, et. al., "Research Priorities for Robust and Beneficial Artificial Intelligence," AI Magazine, Winter 2015,109-112.

[15] Stephan De Spiegeleire, et. al., "Artificial Intelligence and the Future of Defense," The Hague Centre for Strategic Studies, 2017, 45-46.

[16] Agence France-Presse, "From Ale to Jail: Facial Recognition Catches Criminals at China Beer Festival," *The Guardian*, September 1, 2017; Bloomberg News, "China Uses Facial Recognition to Fence In Villagers in Far West," January 17, 2018; Daniel Oberhaus, "China is Using Facial recognition Technology to Send Jaywalkers Fines through Text Message," *Motherboard*, March 2018.

[17] Alex Perala, "Researchers Say Gait Recognition System Offers 99.3% Accuracy," *FindBiometrics*, May 28, 2018.

[18] Daniel Faggella, "AI for Crime Prevention and Detection," *TechEmergence*, August 3, 2018.

their likelihood of committing future crimes.[19] But the US Immigration and Customs Enforcement (ICE) is likewise exploring an "Extreme Vetting Initiative" that would use AI to screen immigrants in order to "determine and evaluate an applicant's probability of becoming a positively contributing member of society."[20] And Singapore launched its Smart Nation initiative in hopes of using AI to improve government services and public health and safety by leveraging and augmenting the digital connectivity that already exists within the city-state.[21] In each of these disparate cases, AI is paired with data derived from a variety of sources to better surveil people and places, and to shape and shift behaviour as a result.

3.2 Intelligence

Alongside surveillance, AI is likewise being adopted to make better sense and use of intelligence. As previously mentioned, AI is particularly good at identifying patterns in reams of data. In that vein, AI is being used to sift through massive troves of information to provide near real-time intelligence analysis. For example, Dawn Meyerriecks, the US Central Intelligence Agency's deputy director for science and technology, told a 2017 Washington, DC, conference, that the CIA was running over 135 pilot projects and experiments on AI that included things like automatically tagging objects in video in order to assist human analysts, and trolling social media for useful information, to better predicting future scenarios, including social unrest.[22] And at the GEOINT 2017 Symposium, Robert Cardillo, the director of the US National Geospatial-Intelligence Agency, warned that if the US were to attempt to manually interpret the commercial satellite imagery that it expects to collect over the next several years, it would need to employ some eight million imagery analysts, an impossible figure. The solution, Cardillo offers, is AI: his organization's goal is to "automate 75% of [human analyst] tasks … so they can look much harder at our toughest problems – the 25% that require the most attention."[23] And, finally, the US Department of Defense (DoD) has been particularly busy in exploring how AI might help it make better use of data and intelligence.

DoD launched the Algorithmic Warfare Cross-Functional Team (Project Maven) in 2017. Part of the project – which was facilitated by Google until its employees baulked at working with the Pentagon – involves using software to autonomously analyse surveillance material, including that derived from drones.[24] Project Maven's immediate, tactical objective is to "turn the enormous volume of data available to DoD into actionable intelligence and insights at speed." But DoD's much larger strategic goal is to "maintain advantages over increasingly capable adversaries and competitors" who are themselves exploring and integrating AI in defence planning.[25] To that end, one recommendation made by the Defense Innovation Board (DIB), an independent advisory committee launched in April 2016 and chaired Eric Schmidt (former Executive Chairman of Alphabet) that advises the Secretary of Defense, was to create a standalone DoD center for studying AI and machine learning. In response, in late 2018, DoD established the Joint Artificial Intelligence Center (JAIC), which ties the Pentagon's efforts in AI to those of the larger US Intelligence Community. The long-term ambition is to establish JAIC – or another organization like it – as a major US

[19] Yi Shu Ng, "China Is Using AI to Predict Who Will Commit Crime next," Mashable, July 24, 2017.

[20] Sidney Fussell, "AI Experts Say ICE's Predictive Extreme Vetting Plan Is Tailor-Made for Discrimination," *Gizmodo*, November 16, 2017.

[21] "Many Smart Ideas: One Smart Nation," Smart Nation | Strategic Projects and Enablers, Singapore, August 2018, https://www.smartnation.sg/.

[22] Frank Konkel, "The CIA Says it can Predict Social Unrest as Early as 3 to 5 days out," *Defense One*, October 2016; Patrick Tucker, "Russian Weapons Maker to Build AI-Directed Guns," *Defense One*, July 2017.

[23] Robert Cardillo (Prepared Remarks), GEOINT 2017 Symposium, June 2017, https://www.nga.mil/MediaRoom/SpeechesRemarks/Pages/GEOINT-2017-Symposium.aspx

[24] Cheryl Pellerin, "Project Maven to Deploy Computer Algorithms to War Zone by Year's End," US Department of Defense, News, July 21, 2017.

[25] US Deputy Secretary of Defense, *Memorandum: Establishment of an Algorithmic Warfare Cross-Functional Team*, April 26, 2017.

defense laboratory for all things AI, akin to the various scientific labs, like the Sandia National Lab run by the US National Nuclear Security Administration, that explore the development of other critical technologies. The opportunities for DoD of AI are "so ubiquitous and the adversary threat … so competitive," warns DIB, that "without exaggeration, the Board likens this situation to that which existed in the first (nuclear weapons) and second (precision munitions and stealth) offsets."[26] The DIB concludes that AI represents DoD's "Third Offset thinking." JAIC, and future iterations of it, are at the frontline of these efforts.

3.3 Autonomous Weapon Systems

Of all the debates surrounding AI and warfare, greatest popular and media concern is reserved for Lethal Autonomous Weapons Systems. [For the latest iteration of this movement, simply Google "Slaughterbots."] By broadest definition, critics present LAWS as any weapon platform that has the ability to select, target, and engage an adversary autonomously.[27] While important ethical, practical, and legal concerns have been levied against fully autonomous offensive weapons,[28] the purpose of this article is centered on exploring the coercive effect, rather than the moral consequence, of these systems. For clarity, weapon systems can be provided different levels of autonomy. As Paul Scharre describes in *Army of None*, if a human remains "in the [Observe, Orient, Decide, Act] loop deciding which target(s) to engage," the system in question should be considered a semiautonomous weapon. In this case, the search and detection of a target may be autonomous, but a human decides to engage and destroy a target. Contemporary drone warfare follows this pattern of behavior. Conversely, with autonomous weapon systems, the entire process of identifying, detecting, and engaging a target is done autonomously. Yet even here, autonomous weapons can be further sub-subdivided. On one hand, *supervised* autonomous weapons, like those widely used to defend ships, bases, and other potential targets from missile or rocket attack, engage autonomously with a target (usually an incoming projectile), though humans remain in the loop and supervise the weapon's use. A human can intervene if and where needed. *Fully* autonomous systems, on the other hand, perform the entire decision process autonomous and human intervention is not possible. Using the loop analogy, Daniel Hoadley and Nathan Lucas (and others) suggest that humans can be *in* the loop (semi-autonomous), *on* the loop (human supervised autonomous systems), and *out* of the loop (fully autonomous systems).[29]

While Scharre argues that very few contemporary weapon systems have crossed into the fully autonomous category, some have, and more are expected to. Examples include the Israeli Aerospace Industries' Harpy – a drone-like weapon that can loiter above a prescribed location for hours until it engages with a specific target. As Scharre explains, while a human decides to launch the Harpy in order to "destroy any enemy radars" within a proscribed geographic area and timeframe, the Harpy itself "chooses the specific radar it destroys."[30] In this case, the human does not know in advance, even when launching the weapon, which specific target the weapons will choose to destroy. There is a distinction, then, between a machine ordered by a human to target something or kill someone, and a machine deciding on its own to target something or kill someone. At issue, for both opponents and proponents of these systems, is that fully autonomous and

[26] Defense Innovation Board, 2018, https://innovation.defense.gov/Recommendations/

[27] International Committee of the Red Cross, "Autonomous Weapons: Decisions to kill and destroy are a human responsibility," https://www.icrc.org/en/document/statement-icrc-lethal-autonomous-weapons-systems

[28] iPRAW, "Focus on Computational Methods in the Context of LAWS," November 2017; Ariel Conn, "The Risks Posed by Lethal Autonomous Weapons," Future of Life Institute, Sept 2018.

[29] Dan Hoadley and Nathan Lucas, "Artificial Intelligence and National Security," Congressional Research Service, April 2018, 24-26.

[30] Scharre, *Army of None*, Chapter Three

offensive weapons systems are being developed and are likely to be more widely used in future conflicts and wars.[31]

4.0 AI AND DEFENCE: A REVIEW OF THE LITERATURE

The academic literature on AI and deterrence is exceptionally slim. With few exceptions, no study has yet to unpack the various ways in which the technology might intersect with deterrence logic, theory, and practice writ large. There is, however, a budding and related grey literature, mostly published by think tanks over the last 18 months, that explores AI and strategic studies, international security, global competition, and warfare. Lessons derived from this literature is useful for thinking about AI deterrence more specifically.[32] What follows is a summary of several key studies.

In September 2018, the Washington-based Brookings Institution began publishing a series, *A Blueprint for the Future of AI*, that explores the effects AI might have on society, including in health care provision, education, governance, business, and security. The series builds off an earlier, April 2018 piece co-published by Brookings' President, John Allen, in which he and Darrel West explore the many ways in which AI and society will intersect. In security, West and Allen argue that AI will shorten the distance from intelligence gathering to decision-making. Under conditions of "hyperwar", they argue, massive amounts of data will be "sifted in near real time – if not eventually in real time," providing decision-makers with greater intelligence awareness and more options far more quickly. Under some conditions, they continue, perhaps especially in cybersecurity, AI might be able to make better sense of malicious code more quickly, alerting human handlers of incoming threats, and if empowered to, responding autonomously. And on the question of keeping the kill chain within human hands, West and Allen warn that US adversaries, like Russia, China, and North Korea, are "not nearly so mired in this debate."[33] A moral divide exists.

As for the Brookings' series itself, three papers focus on AI and security.[34] Mara Karlin's contribution explores how AI might alter the "strategic level of national security," and influence "the management, employment, and development of military force." She argues that AI will augment the influence the American private sector will have on national security decision-making, given their current control over AI research, development, and use. Like West and Allan, Karlin hypothesizes that by providing decision-makers with new and alternative options based on a wide-ranging assessment of an unimaginably large trove of data, that AI may eventually convince decision-makers to delegate some tasks (including targeting) to machines under specific time-sensitive conditions, and to re-evaluate existing military narratives, plans, and paradigms. In another piece, Alina Polyakova focuses on AI as a tool of asymmetric war; she calls it "AI-driven asymmetric warfare," or ADAW. With Russia in mind, Polyakova illustrates how weaker adversaries might "co-opt existing commercially available" AI technology to challenge stronger states with AI-enhanced cyberattacks and AI-generated disinformation or political influence campaigns. For instance, she suggests that AI-driven "deep fake" technology – which allows a user to swap one person's face for another in video content – can produce fake but highly realistic and customized content that can be used to strategically shift

[31] See, Paul Scharre, "A Million Mistakes a Second," *Foreign Policy*, Sept. 2018.

[32] At times, however, some authors (e.g. Horowitz, Scharre, Cummings, and Hoadley and Lucas) are less sanguine about the utility AI will have in national security, suggesting that the technology is easily duped, spoofed, or exploited, does not easily lend itself to very simple cross-domain tasks (like interpreting between formal writing and informal social media writing), and cannot often explain to human operators how decisions and outputs were produced (e.g. AI as a black box). See, for instance, Michael Horowitz, "The Promise and Peril of Military Applications of Artificial Intelligence," *Bulletin of the Atomic Scientists*, April 2018.

[33] Darrell West and John Allen, "How Artificial Intelligence is Transforming the World," Brookings Institution, April 2018; Brookings, A Blueprint for the Future of AI, https://www.brookings.edu/series/a-blueprint-for-the-future-of-ai/

[34] Mara Karlin, "The Implications of Artificial Intelligence for National Security Strategy," Brookings, Nov. 2018; Alina Polyakova, "Weapons of the Weak: Russia and AI-driven Asymmetric Warfare," Brookings, Nov. 2018; Michael O'Hanlon, "The Role of AI in Future Warfare," Brookings, Nov. 2018.

narratives and perceptions, and ultimately, behavior. Finally, Michael O'Hanlon provides the third Brookings' article, in which he links advancements in AI and robotics together to illustrate how new tactics on the battlefield might require a rethink at the strategic level. For illustration, he argues that in a future battlefield, thousands of miniature autonomous drones deployed both at sea and in the air might lead to new swarm or "saturation" tactics, a topic Scharre first broached in 2014.[35] O'Hanlon concludes that these tactics might put an end to "the kind of impunity that US forces have enjoyed for decades" in Europe, Asia, and elsewhere.

Much like Brookings, the Center for a New American Security (CNAS), a Washington-based, non-partisan organization founded in 2007, developed its *AI and Global Security Initiative* to explore the different ways in which AI and security intersect. The Initiative has produced a series of publications, most of them linked to Paul Scharre and Michael Horowitz, who help direct CNAS's Task Force on Artificial Intelligence and National Security.[36] None of the publications explore deterrence, but they do provide a detailed assessment of the various ways in which AI might affect security and military affairs. Several highlights are worth unpacking. In a CNAS reprint from *Foreign Policy*, Horowitz suggests that as a general-use technology largely based on software developments, competition for AI will be broad, uniquely combining the efforts of countries and corporations alike. In another reprint from *Foreign Policy*, Scharre explores the ramifications of letting machines dictate the speed of warfare, including the associated risk of inadvertently augmenting the result of minor algorithmic accidents. In other pieces, Scharre, Horowitz, and CNAS colleagues outline the state of the art, and explore narrow AI's applicability across the various security, intelligence, and defence disciplines. They argue that AI will help classify data; detect anomalies in patterns of behavior; predict future behavior; improve low-skilled human performance; facilitate labor-intensive activities; automate tasks in both physical and cyberspace; uncover new cyber vulnerabilities, threats, and solutions; develop, distribute, and counter "targeted propaganda"; and provide super-human qualities in speed, precision, reliability, patience, and vigilance.

Elsewhere, Horowitz, writing in the *Texas National Security Review* – a new, policy-oriented publication linked to *War on the Rocks* – categorizes AI as the "ultimate enabler", an all-purpose "technology with a multitude of applications" rather than as a weapon in any traditional sense.[37] It allows developers, innovators, and adopters to apply it across the security spectrum in unique ways, unlike other, less flexible technological developments, like the ballistic missile or machine gun, which proved useful and revolutionary but in a much more limited context. Like Scharre and O'Hanlan, Horowitz suggests that low-cost, autonomous drones, coordinating their actions at machine speed, might undermine high-cost, high-quality legacy weapon systems. Horowitz argues further that the way AI develops in the future will help dictate the utility and advantage it might lend to its developers and early adopters. If advancements are led by the private sector, for instance, AI might more quickly "diffuse" to militaries around the world, who purchase and adopt it for their own use. That would reduce the original developer's "first-mover advantage," and could narrow the balance of power between innovators and adopters. But if AI – or certain types of AI particularly useful to defence and security – is developed primarily by states, government labs, and their militaries, the technology will be slower to spread between countries, and innovators may retain a technological edge that might translate into a longer-lasted strategic advantage. These assertions are explored further by M. L. Cummings in a 2017 brief published with Chatham House, a British think tank. Cummings suggests that private sector innovation in AI currently has the advantage because top engineering talent find more lucrative careers in the commercial applications of AI than they do in the more narrowly-focused aerospace and defence industry. This is especially true in the US, Canada, and Europe. "The global defence

[35] Paul Scharre, "Robotics on the Battlefield Part II: The Coming Swarm," CNAS, Oct 2014.

[36] CNAS, Artificial Intelligence and Global Security, https://www.cnas.org/artificial-intelligence-and-global-security

[37] Michael Horowitz, "Artificial Intelligence, International Competition, and the Balance of Power," *Texas National Security Review*, May 2018.

industry," she warns, "is falling behind its commercial counterparts in terms of technology innovation."[38] Bridging the gab may be difficult.

The RAND Corporation, another leading US think tank with a long track record of publishing work on strategic studies, held several workshops in 2017 that explored future US security challenges, circa 2040, and published a number of subsequent reports in 2018. One of these short reports explores the way AI might interfere with nuclear strategy and strategic deterrence in particular.[39] Several useful findings stand out. First, if AI creates the perception among nuclear states that one country has the ability to detect, locate, and target all of another state's nuclear weapon launchers – an infeasibility today – then vulnerable states may be especially inclined to use these weapons more quickly at the risk of losing them altogether. Second, hacking an AI's system by, for instance, "poisoning its training data" or degrading hardware, might render AI-generated strategic advice suspect, fallible, and maliciously manipulated by third parties. Third, autonomous nuclear weapons akin to aerial or underwater drones – a "doomsday drone" – could be used to further dissuade a challenger from launching a debilitating nuclear first strike by augmenting the credibility and survivability of new-age second strike weapons. And fourth, AI might be used in arms control, augmenting the trust, control, and transparency that underpins the counter-proliferation verification process.

Beyond the limited academic and think tank literature, AI and national security has been extensively explored by the US government. Leading the charge, in October 2016, the Executive Office of the President of the United States published a broad strategy document on AI, *Preparing for the Future of Artificial Intelligence*.[40] Of interest to us are its brief discussions dedicated to exploring how AI might be used in cybersecurity and military affairs. On the former, the document suggests AI might create more robust, agile, and cheaper forms of cybersecurity, able to more quickly detect and respond to malicious events. Citing DARPA's 2016 Cyber Grand Challenge, which sought to develop and test autonomous systems for exploiting and patching digital flaws within an all-machine tournament, these systems could eventually "perform predictive analytics to anticipate cyberattacks." On the latter, the document suggests that offensive AI weapons systems may have "greater precision", and could introduce "safer, more humane military operations" that rest on limiting the risk to military personnel, diminishing the number and type of munitions used, and producing less collateral damage. And defensive AI weapons could be used for "protecting people and high-value fixed assets," and possibly, "deterring attacks through non-lethal means." Finally, beyond the battlefield, the document argues that AI could "provide significant benefits," in logistics, maintenance, base operations, veterans and personnel affairs, emergency and medical response, navigation, communications, and intelligence, all of which would make "American forces safer and more effective."

Some of these themes are picked up by Daniel Hoadley and Nathan Lucas in their 2018 brief to the US Congressional Research Service, in which they provide a detailed overview of the challenges and opportunities AI introduces to combat environments.[41] While their work does not mention deterrence at all, they do provide a useful classification or clustering of the military effects of AI that lends itself to an exploration of coercion. First, AI will provide military systems with autonomy, replacing humans in dangerous, risky, complex, and labor-intensive jobs. Second, AI greatly augments the speed with which actions can be taken: it will have the "ability to react at gigahertz speed," and accomplish "long-duration tasks that exceed human endurance." Relatedly, AI will make sense of a huge quantity of data from disparate sources. The result is that AI might provide a decision-maker with advice, suggested actions, or solutions that allow it to outpace an adversary's own assessment of and ability to strategically react to a situation if left to human analysis alone. Third, AI might boost the productivity and capability of human soldiers and of less

[38] M. L. Cummings et. al. *Artificial Intelligence and International Affairs: Disruption Anticipated*, Chatham House Report, (2018), 7-18.

[39] Edward Geist and Andrew Lohn, "How Might Artificial Intelligence Affect the Risk of Nuclear War?" RAND (2018).

[40] Executive Office of the President of the United States, "Preparing for the Future of Artificial Intelligence", October 2016.

[41] Hoadley and Lucas, "AI and National Security".

sophisticated military systems. Like various other authors, Hoadley and Lucas focus on drones. Alone, a single drone is no match for a fighter jet, but algorithmically lassoed together, a fleet of thousands might well overwhelm it. And fourth, AI might provide out-of-the-box and unpredictable strategic advice that stretches the human imagination.

5.0 AI DETERRENCE: A SPECULATIVE ACCOUNTING

Deterrence has been around a long time.[42] It has, often and repeatedly, proven its theoretical flexibility in responding to shifting international dynamics and to emerging (and disruptive) technologies. The literature on deterrence is, as a result, robust, cumulative, and expansive. This evolution has occurred within the context of "four waves" of scholarship.[43] Briefly, the first wave began following the Second World War, and focused on the development of nuclear weapons within a US-dominated strategic backdrop. The second wave corresponds to the great power dynamics of the Cold War period itself, and to the proliferation of nuclear weapons. Game theory and behavioral models were used to better understand alliances and escalation, with a focus on strategic and conventional deterrence by punishment, and on mutually rational and unitary state actors. The third wave followed in the 1970s, with a focus on empirically testing the frameworks and theories previously proposed. Concepts like communication, commitment, resolve, and expected utility, along with central versus extended deterrence, were refined. And the psychology of coercion – cognition, bias, fatigue, misperception – along with notions of strategic culture, bureaucracy, and leadership also emerged.

The fourth wave began at the end of the Cold War, with a shift in focus from bipolarity, strategic weapons, and great power rivalry to an eclectic (and even chaotic) diffusion of interests on "rogue" states, violent non-state actors – from pirates to terrorists – and to processes like radicalization, rebellion, and insurgency. While a resurgent Russia and a rising China have garnered renewed interest in traditional coercive dynamics, the scope of deterrence nonetheless broadened during the fourth wave, with reinvigorated theoretical research on the logic of denial, delegitimization, and conventional deterrence. And altogether novel concepts for deterrence in both space and cyberspace, like cross-domain deterrence, were also developed. This article, and the larger AI deterrence project upon which it is based, reflects perfectly the speculative nature of fourth wave scholarship.

At its logical core, deterrence entails using threats to manipulate an adversary's behavior. More concretely, it rests on convincing an adversary to forgo an unwanted action. Compellence – a related concept – flips the process around, prompting an adversary to pursue an action it might otherwise have avoided. Deterrence restrains behavior while compellence encourages it. Both are subsumed within the logic of coercion more broadly. In all cases of deterrence, compellence, and coercion, a minimum of two actors are involved. Defenders issue threats to alter an opponent's behavior; challengers consider these threats and decide whether or not to acquiesce. In some cases, a third actor is involved. With extended deterrence, for instance, a proxy, partner, or ally is also included in the process, either on the side of the defender or the challenger. Threats of punishment – usually in the form of retaliation – and promises of denial – usually in the form of depriving an adversary the expected benefits of a particular behavior – form the basis for most deterrent and compellent engagements, though ideological considerations that inform behavior have also been recently developed within the context of deterrence by delegitimization.

[42] The following sub-section on defining deterrence borrows from my previous work on the subject. For an expansive list of sources, see: Wilner, *Deterring Rational Fanatics*, Chapters 2 and 3.

[43] Jeffrey Knopf, "Terrorism and the Fourth Wave in Deterrence Research," in Wenger and Wilner, (eds.) *Deterring Terrorism: Theory and Practice* (Stanford UP: 2012); Amir Lupovici, "The Emerging Fourth Wave of Deterrence Theory," *International Studies Quarterly* 54 (2010); Alex Wilner, "Deterring the Undeterrable: Coercion, Denial, and Delegitimization in Counterterrorism," *Journal of Strategic Studies* 34:1 (2011); Thomas Rid, "Deterrence Beyond the State: The Israeli Experience", *Contemporary Security Policy* 33:1 (2012).

Deterrence weighs on a challenger's cost-benefit calculus, on the strategic choice it has available to it. Militarily crushing an adversary, Thomas Schelling writes, such that it has no alternative but to accept demands made of it is not deterrence, but rather military defeat.[44] The same logic holds for denial: Developing a foolproof defence, however rare, does not deter by denial but rather deters by defeat. In this case, adversaries do not penetrate a target, for instance, because they are prevented from doing so, not because they decide against doing so. In deterrence, changes in behavior are an option, not a necessity. Finally, though deterrence is usually anchored to International Relations theory, it also functions below the level of the state among non-state actors, individuals, groups, and other organizations. Criminological deterrence, for instance, and scholarship on deterring terrorism both provide insights on deterring violence within communities and emanating from violent non-state actors.

In practice, deterrence rests on a number of pre-requisites. First, actors must retain a degree of rationality, such that a combination of threats will suffice to shift their behavior. Second, challengers and defenders must agree, to some point, that non-violence is preferable to violence: if conflict is the only preference both sides share, then deterrence will not work. Third, defenders must not only communicate (or otherwise signal) their threats and expected changes in behavior, but must also reassure challengers that if they acquiesce to a threat, punishments will not be meted out. Fourth, defenders must have the perceived capability, and the resolve, to punish or deny as threatened. And finally, deterrence works best again a known or suspected adversary; anonymity robs deterrence of a coercive target.

Taken together, how might AI and deterrence intersect? A speculative accounting – a laundry list of thoughts – follows.

First, better defence equals better denial. AI, by improving the speed and accuracy of some defensive weapons, and by subsequently improving the reliability of defending infrastructure, weapons platforms, and territory against certain kinetic attacks (e.g. missile or rocket attack), might deter some types of behavior by altogether denying their utility. The same logic holds in defending cyber platforms: by denying aggressors access to information more persistently, a defender's AI might compel a challenger not to bother attacking in the first place.

Second, and on the flip side, under other conditions AI may augment the feasibility of certain types of attack, favouring punishment strategies over denial strategies. As noted, autonomous swarming drones have garnered the greatest attention. If and when these capabilities are developed and refined, swarming bots may provide challengers with a unique coercive tool not easily deflected or defeated. Saturation tactics that rely on thousands of disposable and easily replaced robotic platforms may tip the balance towards offensive measures and the promise of punishment strategies. Conversely but relatedly, swarms might likewise be used by a defender to fend off an attack against it employing conventional weapons – think of a defensive robotic swarm launched to counter an incoming fighter jet – providing it with a tactical-level threat of denial. But then again, and historically speaking, offensive developments usually spur defensive developments in kind: Just as AI feeds offensive swarming tactics, so, too, might it eventually result in defensive swarming responses. The resulting robotic dog fight might recalibrate coercion towards the middle.

Third, and moving beyond kinetics alone, AI might improve a state's ability to plan and carryout both offensive and defensive coercive threats, by improving logistics, navigation, communications, recruitment, training, deployment, and so on. The back-office AI that coordinates the machinery of warfare may make some coercive threats more robust and persuasive as a result.

Fourth, by rapidly providing unique and novel advice to decision-makers that supersedes human innovation and capability, AI may provide situational awareness that dips into predictive analytics. By melding an improved analysis of what adversaries have done in the past and are currently doing today – indeed this very minute – AI will provide users with the ability to anticipate an adversary's next action. Defenders can

[44] Thomas Schelling, *Arms and Influence* (1966), 1-34.

preemptively respond accordingly.[45] If, over time, a challenger believes that a defender can anticipate its next move, it may be deterred from trying, or might alternatively calculate that only brash, unexpected, and novel actions will help it accomplish what it wants (at least, until the machine learns from these episodes, too).

Fifth, by manipulating public information through deep fakes and other related processes, AI might provide users with a new form of deterrence by delegitimization. The threat, in this case, is the ability to create, release, and disseminate fake video or audio material threatening or embarrassing to a target. Think of Russia – or of any other state or non-state actor, if only because the technology will be cheaply available – surreptitiously threatening a US Democratic Party presidential nominee with engineered content that could influence the candidate's standing among the electorate in the leadup to a future US presidential election. Because determining the veracity of AI-manipulated content and attributing its source is difficult to do, countering these types of coercive misinformation campaigns may be challenging.[46]

Sixth, fighting at "machine speed" may change the calculus of taking action. If AI-based decision-making provides one side of a conflict an advantage in responding quickly and decisively, then others will eventually mimic and come to rely on these processes, too. But as both sides of a contest come to rely on machines for insights, the very rationale of these AI-generated insights may degrade more quickly over time, as one side's AI responds and reacts to another's. Put another way, in this scenario an AI-generated insight may have a short shelf life, and windows of opportunity may prove fleeting. If so, the logic and value of striking first, and fast, may prevail, upending long-standing coercive and escalatory calculations along the way.

Seventh, AI might provide traditionally weak actors with novel means to coerce traditionally strong actors. The dual-use nature of AI along with private-sector developments in the technology, suggests that many states – including relatively weak ones – and non-state actors, organizations, and individuals as well, may eventually be able to purchase the technology for their own use. While weak actors may face other limitations, like acquiring access to useful training data, AI might nonetheless help level the playing field with more powerful actors. If so, diffusion of the technology may diminish how the strong deter or compel the weak, and might otherwise allow the weak with new avenues for coercing the strong.

Eighth, ubiquitous real-time surveillance could deter criminal behavior. If a state where to establish AI-powered surveillance of urban centers, border crossings, and other sensitive locations to generate biometric identification and behavioural analytics, and if it were to publicly announce its use of these tools, it might convince criminals, terrorists, spies, and other nefarious actors that their plans are unlikely to succeed. China's experiment in countering jaywalking at busy urban intersections is informative.[47] Facial recognition cameras monitor roads, snapping pictures of jaywalkers, and matching the offender to photo IDs stored in a database. The photos, along with some of the individual's personal information, can then be displayed on roadside screens, and fines can be issued automatically. In the city of Ji'Nan, police report that this technology has reduced jaywalking by 90 percent. Used at scale, the technology could curtail other unwanted behavior and activity.

And finally, ethical and legal limitations on how AI is used in battle may dictate how some countries behave and others respond. While some states, notably the United States and several European allies, are openly against providing AI with the right or the means to kill individuals without human intervention – French President Emanuel Macron explained, for instance, while promoting his country's new AI innovation strategy in 2018, that he was "dead against" the idea[48] – other countries appear far less concerned. China,

[45] Yaakov Lappin, "Artificial Intelligence Shapes the IDF in Ways Never Imagined," *The Algemeiner*, October 2017.

[46] Will Knight, "The Defense Department has Produced the First Tools for Catching Deep Fakes," *MIT Technology Review*, August 2018.

[47] Meghan Han, "AI Photographs Chinese Jaywalkers; Shames them on Public Screens," *Medium*, April 9, 2018.

[48] Nicholas Thompson, "Emanuel Macron Talks to Wired about France's AI Strategy," *Wired*, March 2018.

Russia, Israel, and others, for example, may be more willing to delegate decisions – including those that result in human death – to Artificial Intelligence. Under certain conditions, doing so may provide these countries with a tactical, strategic, or coercive advantage over those inclined to keep humans in or on the loop. It may likewise provide these countries with a means to counter-coerce, influence, or otherwise manipulate countries that are more constrained and refrained in the way they use their AI in battle.