



How viable is international arms control for military artificial intelligence? Three lessons from nuclear weapons

Matthijs M. Maas

To cite this article: Matthijs M. Maas (2019) How viable is international arms control for military artificial intelligence? Three lessons from nuclear weapons, Contemporary Security Policy, 40:3, 285-311, DOI: [10.1080/13523260.2019.1576464](https://doi.org/10.1080/13523260.2019.1576464)

To link to this article: <https://doi.org/10.1080/13523260.2019.1576464>



Published online: 06 Feb 2019.



Submit your article to this journal [↗](#)



Article views: 4216



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 2 View citing articles [↗](#)



How viable is international arms control for military artificial intelligence? Three lessons from nuclear weapons

Matthijs M. Maas  ^{a,b}



^aFaculty of Law, Centre for International Law, Conflict and Crisis, University of Copenhagen, Copenhagen, Denmark; ^bCenter for the Governance of AI, Future of Humanity Institute, University of Oxford, Oxford, UK

ABSTRACT

Many observers anticipate “arms races” between states seeking to deploy artificial intelligence (AI) in diverse military applications, some of which raise concerns on ethical and legal grounds, or from the perspective of strategic stability or accident risk. How viable are arms control regimes for military AI? This article draws a parallel with the experience in controlling nuclear weapons, to examine the opportunities and pitfalls of efforts to prevent, channel, or contain the militarization of AI. It applies three analytical lenses to argue that (1) norm institutionalization can counter or slow proliferation; (2) organized “epistemic communities” of experts can effectively catalyze arms control; (3) many military AI applications will remain susceptible to “normal accidents,” such that assurances of “meaningful human control” are largely inadequate. I conclude that while there are key differences, understanding these lessons remains essential to those seeking to pursue or study the next chapter in global arms control.

KEYWORDS Artificial intelligence; AI; arms race; arms control; nonproliferation; epistemic communities; normal accidents; governance

New technologies that promise significant strategic advantages can upset balances of power or disrupt previously stable global governance arrangements. Artificial intelligence (AI) is one such critical technology, with some suggesting that in the coming decades it will spell “the biggest geopolitical revolution in human history” (Drum, 2018, p. 46). Even on more modest readings, it is clear that the technology can offer states critical advantages across many strategic and military domains (Dafoe, 2018; Payne, 2018b). AI is, after all, a general-purpose technology for improving the accuracy, speed, and/or scale of machine decision-making in complex environments. As such, while its efficacy and utility is certainly not limitless, AI will

CONTACT Matthijs M. Maas  Matthijs.Maas@jur.ku.dk  Faculty of Law, Centre for International Law, Conflict and Crisis, University of Copenhagen, Karen Blixens Plads 16, Building: 6A.4.05, Copenhagen 2300, Denmark

increasingly allow us to substitute for and improve upon human performance in tasks such as pattern recognition, prediction, optimization, and (autonomous) decision-making—all domain-general tasks that are key to performance across a wide range of strategic and military contexts.

The strategic potential—and appeal—of AI has hardly gone unnoticed by states. In recent years, many nations have begun to emphasize the role of AI as a cornerstone in both their national strategies and military doctrine (cf. China's State Council, 2017; Putin, 2017; Work, 2015). Moreover, while many military AI applications remain somewhat immature at present (Payne, 2018a, p. 9), a range of systems are now seeing development or deployment (Maas, Sweijts, & De Spiegeleire, 2017). Indeed, a 2017 review by the Stockholm International Peace Research Institute already identified 49 deployed weapon systems with autonomous targeting capabilities sufficient to engage targets without the involvement of a human operator (Boulanin & Verbruggen, 2017, p. 26). Moreover, further advances are continually being made, with many states showing an interest in further developing military uses of AI (Ayoub & Payne, 2016; Horowitz, 2018a).

As a result, there is today a widespread perception amongst the public, policymakers, and scholars, that the global development of AI is swiftly escalating into a strategic arms race—or even a “new Cold War”—between major states such as the United States and China (cf. Allen & Kania, 2017; Auslin, 2018; Barnes & Chin, 2018; Geist, 2016; Hogarth, 2018; Lee, 2018; Thompson & Bremmer, 2018). Related to this, there is a fear that the widespread militarization of AI, and its deployment to the battlefield, is but a matter of time. Some have challenged the overall “arms race” narrative on the grounds that such framings not only misrepresent the nature of AI and its innovation (Kania, 2018), but also play into adversarial and zero-sum thinking that is counter-productive or outright dangerous (Cave & Ó hÉigeartaigh, 2018; Zwetsloot, Toner, & Ding, 2018). While these are valid and important points, it is important still to consider the strategic and governance implications of the AI arms race framing, given how prevalent it already is.

Should we be at all concerned about AI arms races? Most likely, yes: In the past, international efforts to control the proliferation, production, development or deployment of certain military technologies—from chemical and biological weapons to land mines and ballistic missile-defense systems—were all, to various degrees, motivated and grounded by four distinct rationales: ethics, legality, stability, or safety. Military AI likewise has raised concerns on all four grounds. Some have objected to “killer robots” on *ethical* grounds (Human Rights Watch, 2012; Nehal, Beck, Geiss, Liu, & Kress, 2016; Roff, 2014), or in terms of their noncompliance with principles of international (humanitarian) *law* (cf. Davison, 2017). Others worry that AI capabilities may adversely affect *strategic stability* between states—by shifting the tactical offense-defense balance (Rickli, 2017), eroding nuclear deterrence stability (Geist & Lohn,

2018; Lieber & Press, 2017); or creating mutual uncertainty over the new balance of power, which may feed destabilizing miscalculation (Kroenig & Gopaldaswamy, 2018). Finally, concerns are raised over *safety*, reflecting fears that the pursuit of narrow military supremacy can trap states in what Danzig (2018) has called a “technology roulette.” As states race to deploy increasingly autonomous military AI systems, their intrinsic vulnerability to unexpected interactions or operational accidents (Scharre, 2016a) raises the specter of inadvertent escalation into a “flash war” between autonomous military systems, similar to the algorithmic flash crashes already observed in the financial sector (Scharre, 2016b, 2018b). In sum, because military AI again invokes these four seminal concerns, the question of appropriate regulation appears pertinent. It also appears timely, since “the idea of arms control for AI remains in its infancy” (Payne, 2018a, p. 19), and both customary and formal international law remain still very much in flux. It as such seems prudent to consider today whether or not, or how, hazardous AI militarization or arms race dynamics might be contained or channeled. How viable is international arms control for military AI?

In this article, I will seek to answer this question, through drawing a parallel with one of the classic histories of arms control—our chequered, but (so far) more or less successful track record in controlling and managing the proliferation of nuclear weapons. I will draw on these historical insights in order to examine both opportunities and pitfalls facing any prospective international arms control strategies aimed at preventing, channeling or containing the militarization of AI.

This argument proceeds as follows: I first briefly discuss the utility—and limits—of the analogy between nuclear weapons and military AI. I then turn to the body of the argument, by examining the historical record of the non-proliferation, arms control, and management of nuclear weapons through three analytical lenses (“norms and domestic politics,” “epistemic communities,” and “normal accident theory”), and deriving insights for the governance of military AI. In sum, I argue that (1) far from being inevitable, the proliferation of powerful technologies such as military AI might be slowed or halted through the institutionalization of norms; (2) organized “epistemic communities” of experts, appropriately organized, can effectively catalyze arms control agreements; (3) many military AI applications will remain susceptible to “normal accidents,” such that the current focus of governance efforts on “meaningful human control” in “killer robots” is largely inadequate. I conclude that while past strategies to contain and control nuclear weapons cannot and should not be taken as blueprints, these historical lessons remains essential in designing any future efforts to responsibly contain military AI.

Nuclear weapons as case study for military AI

The long history of nonproliferation and arms control offers a wealth of examples of technology governance frameworks at the multilateral, bilateral, regional, transnational and unilateral levels. This history may serve as a potentially rich seam of insights into the opportunities and pitfalls in controlling, containing or channeling the weaponization of disruptive technologies (cf. Scharre, 2018a). Amongst this history, one technology stands out as having been the subject of a particularly long, diverse and rich array of governance and control efforts: nuclear weapons.

Taken at face value, there appear to be large differences between nuclear weapons and AI. Nonetheless, both technologies share key strategic and political characteristics that are critical in the context of arms control and governance. Where it matters, nuclear weapons still offer a highly insightful case study for the control of strategically revolutionary technologies such as military AI. For instance, Payne (2018a) has suggested that there are sufficient points of similarity to inform insightful debate:

Firstly, nuclear weapons and AI are both highly technical scientific developments, requiring coordinated expertise. Secondly, the “revolution” is concentrated in a few states, and the research involves a degree of secrecy which, coupled with the inherent technicalities, constrains public debate. Thirdly, there are valid ethical and legal concerns about proportionality, discrimination and control of weapons employing the technologies. Fourthly, both technologies have the potential to rapidly transform strategy, the institutions charged with applying it, and society more broadly. Lastly, both have potentially apocalyptic consequences and have aroused intense philosophical debate. (p. 15)

The similarities between nuclear weapons and military AI do not end there. Both offer a strong and “asymmetric” strategic advantage, such that parties have strong incentives to pursue the technology unilaterally; both involve dual-use components, technologies, and applications, which means that blanket global bans of the technology may not be politically palatable, enforceable, or even desirable. Moreover, because both technologies involve an (initially) high threshold in technology and tacit scientific knowledge, and because a few states have a definite lead in development, there are uneven global stakes in the technology, which often makes the consolidation of international legal regimes difficult, or undercuts the legitimacy of regimes that are consolidated (cf. Picker, 2001, pp. 191–193). Finally, both deployed nuclear weapons (Sagan, 1993) and military AI systems (Maas, 2018; Scharre, 2016a) exhibit a high susceptibility to “normal accident” failure modes, which has key implications for the targets of governance (as will be discussed further on).

Of course, it is important to understand the limits to this analogy, as well. I will therefore discuss a number of possible caveats or counterarguments to turning to nuclear weapons as a case for military AI.

In the first place, it could be objected that these technologies have fundamentally different use cases. Whereas nuclear weapons have not been (and hopefully will never be) used in anger since Nagasaki, AI systems will see pervasive and daily use in many diverse roles across the battlefield. In many cases, military AI indeed will not be a single weapon so much as a functionality integrated into many other military systems (cf. Verbruggen, 2018)—what Chinese PLA strategists call the “intelligentization” of war (Kania, 2017). Given this, are better analogies to AI not found by studying governance efforts for more modern weapons that see actual, regular deployment, such as drones or cyber weapons? I do not want to claim that examinations of these technologies would be unproductive. Ultimately, the choice of technologies to study reflects the extent to which one emphasizes or balances the four arms control rationales (ethics, legality, stability, safety) discussed above. However, in the context of governance, the salient ethical, legal and strategic feature of nuclear weapons—the rationale for many of the non-proliferation and arms control initiatives—was not that they saw regular tactical use, nor that they were integrated into each and every local battlefield system. Rather, the point was that, even in their non-use (or threatened use), nuclear weapons drastically reshaped the geopolitical and strategic landscapes, and spurred global political and legal effort in response, to a degree that drones—for all their more widespread daily use in war today—have not done. Military AI may certainly see more “active” (rather than “passive,” as with nuclear weapons) military uses, but the key comparison here is not just the “input” (how this effect is achieved), but the output (the magnitude of the risk). On that count, nuclear weapons offer a valuable—and, some have argued the only (Payne, 2018a)—strategic precedent for the emergence of military AI systems.

In the second place, it might be argued that the relevant principals developing the technology are different this time around. Whereas states have been the exclusive player in pursuing the development of nuclear weapons, in terms of general AI capabilities, much of the leading talent and innovation is located in the private sector. Yet while it is true that non-proliferation and arms control initiatives for military AI systems will have to take stock of private sector dynamics (cf. Fischer, 2017), states, as discussed, seem set or determined to become strong stakeholders in AI (cf. Hogarth, 2018), particularly where it pertains to military applications.

In the third place, it might be objected that AI has very different—and much less restrictive—production requirements. The development and deployment of nuclear weapons requires access to rare resources such as uranium (which can be controlled), as well as the construction of difficult-

to-hide enrichment infrastructures and the conduct of conspicuous tests of weapons and delivery systems (which can be detected). The development of AI, by contrast, is often considered far more discreet and discrete (Scherer, 2016), and may not lend itself well to non-proliferation regimes based on restricting key resources. As was the case with cyber weapons, it may be easier to hide military AI development “infrastructure” and tests, eroding the ability to verify compliance with arms control agreements (Borghard & Lonergan, 2018).

On the other hand, the ease of access to military AI capabilities should not be overstated. As argued by Ayoub and Payne (2016), “[the] ubiquitous access to advanced algorithms gives a misleading impression of the ease with which military relevant AI may proliferate between states” (p. 809). In practice, cutting-edge AI still requires very large (and rapidly increasing) amounts of computational power (Amodei & Hernandez, 2018; Hwang, 2018). Tacit knowledge possessed by experienced researchers proved a critical if underappreciated brake on the proliferation of nuclear weapons (MacKenzie & Spindardi, 1995), and will likely play a similar role in limiting the rapid diffusion of military AI capabilities that actually offer strategically meaningful performance improvements (over humans; or against rival systems). Finally, Horowitz has argued that the impact and uptake of a new technology depend not just on the availability of the innovation itself, but also on organizational innovation (Horowitz, 2018a, p. 4). All of this suggests that the set of leading state parties which must be brought in line with governance is not much larger for military AI, than it was for nuclear weapons; and that there are in-principle effective routes to curtailing at least some of the critical paths towards either “horizontal” proliferation (i.e., more parties pursuing or deploying military AI) or “vertical” proliferation (i.e., actors developing and deploying more advanced, ethically problematic, legally disruptive, destabilizing, or accident-prone military AI).

In sum, there are certainly important differences between nuclear weapons and military AI—especially where it pertains to the operational problem of effectively monitoring the more “discreet” AI development to ensure compliance with international arms control regimes. Such considerations must be taken into consideration for arms control efforts this time around. Yet while the idiosyncratic characteristics of military AI should be kept in mind, they do not undercut the comparison, as there are key strategic similarities between the technologies—in terms of the “stakes” at play, and in terms of the purported appeal to states—that a study of nuclear history can offer fruitful and compelling insights into the viability of this next chapter in arms control history.

Three insights from nuclear weapons for AI arms control

Having sketched out these strengths and limits of the comparison; I now provide three succinct historical explorations of nuclear arms control, and what lessons these cases hold for the control of military AI.

Global norms shape domestic politics, and affect the causes and cures of arms races

There exists at times a pessimistic perception that states cannot be permanently barred from pursuing strategically important technology which they suspect their rivals might develop—that accordingly, AI arms races are inevitable or even already underway, and the proliferation of increasingly autonomous and lethal military AI systems a matter of time.

Such modern-day pessimism echoes historical fears from the nuclear era, which held that “proliferation begets proliferation” (Shultz, 1984, p. 18). Indeed, policymakers in the early Cold War perceived nuclear weapons—the ultimate deterrent—as obviously desirable or necessary strategic assets to states. They therefore anticipated a wildfire spread of these weapons. In a 1963 memo to President John F. Kennedy, then-U.S. Secretary of Defense Robert McNamara argued that because of falling production costs, at least eight new nuclear powers might emerge within a decade (McNamara, 1963; as quoted in: Yusuf, 2009, p. 15). In part based on such estimates, JFK gave a public speech later that year in which he articulated “the possibility in the 1970s of... a world in which 15 or 20 or 25 nations may have these weapons” (Allison, 2010).

Yet remarkably, given such pessimism, “horizontal” nuclear proliferation since the 1960’s has proven less a “wildfire,” and more a story of “glacial spread.” By some estimates, up to 56 states have at one or another time possessed the (theoretical) capability to develop a nuclear weapons program (van der Meer, 2014, p. 30). Even though many of these states—up to 39, by some estimates (Pelopidas, 2011)—chose to engage in “nuclear weapons activity,” the majority eventually voluntarily terminated these programs uncompleted (cf. Pinker, 2011, p. 273). “Only” ten states have actually managed to develop these weapons, and nine nuclear weapons states presently remain. How can this restraint be explained? The literature on state decision-making identifies competing theories of state behavior, which focus on the role of (1) security; (2) domestic politics, and (3) norms (Sagan, 1996).

Under the realist *security* model, states pursue nuclear weapons in reaction to perceived security threats—either to offset a rival’s conventional military supremacy, or to match an adversary’s (feared) nuclear program. Under this “realist” reading, nonproliferation policy can only slow down, but not eliminate, the spread of nuclear weapons. While intuitive and parsimonious

—and compelling in explaining “early” proliferation to states such as Stalin’s USSR—there are some problems with the security model. For instance, “national security” can serve as a default post-hoc rationalization for decision-makers seeking to justify contentious choices by their administrations. Moreover, as noted by Sagan (1996), “an all too common intellectual strategy in the literature is to observe a nuclear weapons decision and then work backwards, attempting to find the national security threat that ‘must’ have caused the decision” (p. 63).

Other scholarship has therefore turned to a second factor, the role of *domestic politics*—to the diverse sets of actors who perceive parochial bureaucratic or political interests in the pursuing or foregoing of nuclear weapons. These actors include policy elites; nuclear research- or energy industry establishments; competing branches of the military; politicians in states where parties or the public favor nuclear weapons development (Sagan, 1996, pp. 63–65). Such actors can form coalitions to lobby for proliferation. This happened in India, where the 1964 nuclear test by rival China did not produce a crash weapons program but instead set off a protracted, decade-long bureaucratic battle between parties in the New Delhi elite and nuclear energy establishments. This struggle was only resolved in 1974, when Prime Minister Indira Gandhi, facing a recession and a crisis of domestic support, authorized India’s “Peaceful Nuclear Explosion,” possibly (but inconclusively) to distract or rally public opinion (Sagan, 1996, pp. 65–69). Conversely, this lens also shows how domestic politics can work against proliferation: after pursuing nuclear programs throughout the 1970s–1980s, regional rivals Brazil and Argentine eventually abandoned their nuclear ambitions, as a result of the emergence of liberalizing domestic regimes supported by groups of actors (e.g., banks, monetary agencies) which favored open access to global markets and opposed “wasteful” defense programs (Solingen, 1994).

Finally, a third, “ideational” model of non-proliferation (Ruble & Cohen, 2018) emphasizes the role of (domestic, elite, and global) *norms* on states’ decisions to pursue nuclear weapons. In some cases, nuclear weapons’ symbolic value as a marker of modernity and scientific prowess may have contributed to proliferation, as in the case of President de Gaulle’s perception of the atomic bomb as a symbol to restore France’s great power status following the experiences in the First Indochina War and the 1958 Algerian crisis (Sagan, 1996, pp. 78–79).

More often, norms—implicitly the “nuclear taboo” against the (first) use of nuclear weapons (Carranza, 2018), and explicitly the norms encoded by international legal instruments, including the Nuclear Non-Proliferation Treaty (NPT) and the Comprehensive Nuclear Test-Ban Treaty (CTBT)—appear to have served as a factor in constraining nuclear proliferation, though such norm enforcement is not without challenge (Knopf, 2018). Nonetheless, these global legal instruments can function, in part because they can

provide shared normative frameworks that disseminate and promote non-proliferation norms, -interests and -identities at the domestic-political level, tipping the balance of domestic contestation towards coalitions seeking non-proliferation. While one might expect such effects of “norms capture” to be stronger in liberal societies than in non-liberal ones, some scholars have suggested that even in the latter case, state elites who are not accountable to their own publics simply come to internalize the normative characterization of a successful state as one that abides by its treaty commitments (Rublee, Bertsch, & Wiarda, 2009, p. 222), or at the very least, have incentives for compliance with international non-proliferation norms, to foster a reputation for reliability in the eyes of other states (Williamson, 2003, p. 81). Moreover, global international regimes—defined by Krasner (1982) as “sets of implicit or explicit principles, norms, rules and decision-making procedures around which actors’ expectations converge in a given area of international relations” (p. 2)—can serve as Schelling Points around which global society can coordinate multilateral collective sanctions or rewards (Müller & Schmidt, 2010).

To what extent have non-proliferation norms driven nuclear restraint? Of course, it can be hard to disentangle causal connections between membership in normative (legal) instruments such as the NPT and nuclear restraint, since such behavior could reflect existing policy preferences by the states. Some reviews of the nuclear (non-)proliferation records have suggested that “NPT membership and the NPT regime’s norms have modest or marginal impacts on nuclear proliferation” (Jo & Gartzke, 2007, p. 185), though others found that, when accounting for states’ *ex ante* treaty commitment preferences, state ratification of the NPT treaty regime was “robustly associated with a lower likelihood of pursuing nuclear weapons” (Fuhrmann & Lupu, 2016, p. 530).

Intriguingly, while public norms seem to be able to strengthen the hands of (non-)proliferation coalitions, they do not seem to reliably shift state policy-making where these coalitions do not already exist in sufficient strength. For instance, in 1994 Ukraine chose to join the NPT and renounce its nuclear arsenal in spite of respectable Ukrainian public support for retaining the weapons (Sagan, 1996, p. 80). Conversely, in 1999 the U.S. Senate rejected the CTBT in the face of widespread U.S. public support.

As with all history, it can be hard to distill unambiguous causal chains; yet surveys of the distinct state rationales for nuclear proliferation and of –non-proliferation or the abandonment of ongoing nuclear programs (Garnett, 2012; Sagan, 1996, 2011; Solingen, 1994; van der Meer, 2011, 2014), suggests that, far from proliferation cascades fueled by security and strategic concerns, a far wider array of motives shaped these decision-making processes, with elements of all three models—security, domestic politics, and norms—playing different roles amongst different states, and often contributing to a

decision to forego or abandon nuclear proliferation. For instance, Solingen has charted the role of ascendant liberalizing coalitions, in countries such as Taiwan, South Korea, and Argentina, in shifting towards nuclear restraint, because of the favorable impact of this decision on international trade, aid, technology and investment opportunities, as well as to reduce perceived wasteful budgets for such military programs (Solingen, 1994). The broad history of nuclear restraint suggests that, far from a foregone conclusion, arms races involving strategically appealing technologies can be slowed, channeled, or stopped. This suggests that halting, managing or containing military AI arms races is viable—and hints at a range of considerations for doing so.

In the first place, it suggests that security concerns are conducive but not decisive to arms races, and that a limited number of “first-mover” major powers may share an interest in supporting global legal regimes aimed at the non-proliferation of certain forms of military AI, such as cyber warfare systems, which might otherwise empower conventionally weaker (or non-state) rivals. Given that this group of “leading” states is initially small, bilateral agreements may suffice; however, the unevenly shared stakes in the technology may render eventual multilateral negotiations more difficult (cf. Picker, 2001).

In the second place, the domestic-politics model suggests that strengthening the hand of domestic coalitions pursuing the non-proliferation of AI weapons is one pathway towards shifting state decision-making away from pursuing more problematic categories of military AI, even in the face of clear national security interests. Of course, one caveat here concerns the fact that military AI may have far broader appeal than nuclear weapons did, such that it is harder to find domestic coalitions that are clearly opposed to its development in all cases. For instance, the strategic benefits of developing nuclear weapons are almost solely military and relatively discrete—a half-finished nuclear weapon is not even half as useful as a finished one, and the road from starting a nuclear program towards developing, not just a working weapon or small arsenal, but a credibly survivable and deliverable deterrent is long and potentially *less* useful (because more provocative to well-armed adversaries) than not initiating a nuclear breakout to begin with. In contrast, the benefits of pursuing military AI might be more linear and gradual, with intermediate advances in subfields (e.g., image recognition or drone swarming command and control) enabling not just immediate application to battlefield roles, but also economically productive spin-offs to civilian applications. These features, combined with the comparatively lower reputational costs, may make some forms of military AI more palatable. But not all. Exceptions may be found in high-performance adversarial contexts (such as cyber warfare or aerial warfare) where AI systems or platforms end up directly engaging with each other. In such cases, as Payne (2018a) has argued, “marginal quality might prove totally decisive” because “other things

being equal, we can expect higher-quality AI to comprehensively defeat inferior rivals” (p. 24). In such domains, the incentives for parties to independently develop “second-rate” military AI capabilities might be lower. Conversely, where AI systems do not have to “fight their like” directly (e.g., logistics; facial recognition), second-best AI systems still offer military advantage, and could proliferate widely (Horowitz, 2018a, 2018b). This suggests that the precise appeal of military AI systems to different parties may be more complex—which offers openings for tailored engagement with domestic coalitions.

Thirdly, the “norms”-model suggests that, while policy-makers may pursue the development of AI *in general* because of its “symbolic” value as a marker of global scientific leadership, this may not transfer to the development of AI *for military purposes*. Instead, the degree to which military AI confers status may be mixed: Pursuing openly autonomous “killer robots” may indeed remain unappealing for states. For instance, over the past years, global surveys of public opinion show that in most countries (excepting India), majorities oppose the deployment of autonomous weapons (Open Roboethics Initiative, 2015; Roff, 2017). An opposition appears to be on the rise: an even more recent survey by Ipsos showed an increase in global opposition (from 56% to 61%) since 2017 (IPSOS, 2019). At the same time, other surveys have shown that public opposition to these weapons can be very context-dependent, and drops off if their usage is framed as being defensive and aimed at reducing casualties amongst friendly troops (Horowitz, 2016; West, 2018). In another U.S. survey, Americans generally expressed mixed support for the United States investing more in AI military capabilities, but also for the United States to cooperate with China to avoid the dangers of an AI arms race (Zhang & Dafoe, 2019, pp. 26–30).

Moreover, advocacy efforts might well be able to shift these public norms on military AI further—and thereby alter the reputational penalties and rewards of deploying new systems or for complying with restrictive global regulation, respectively. Indeed, it is important to recognize the considerable efforts that have been put into making “killer robots” normatively unpalatable, notably by movements such as the Campaign to Stop Killer Robots, a coalition of 89 NGOs from 50 countries (cf. Joshi, 2019). In fall 2018, both the European Parliament as well as United Nations Secretary General António Guterres called for bans on autonomous weapons (European Parliament, 2018; Guterres, 2018); and at present at least two dozen states are pursuing such a legally binding ban—although states such as the United States, United Kingdom, and Russia still explicitly oppose such an initiative (Joshi, 2019). Even if such public advocacy efforts have not (yet) produced a ban on autonomous weapons, this does not mean they have not already influenced the normative space around military AI.

However, to what extent will a specific opprobrium on the—important, but narrow—category of autonomous weapons transfer to other types of military AI? Indeed, beyond “killer robots,” it is unclear to what extent states will face a meaningful or strong “military AI taboo” with the same strength as the “nuclear taboo.” After all, the latter norm was possible and potent, because nuclear weapons are a “single” technology with a single discrete, publicly visible and viscerally horrifying use mode. This created a natural and unambiguous “red line” in usage, not to be crossed. Conversely, the deployment and use of AI in many (non-kinetic) military applications is already a fact, such that this Rubicon has been crossed. Moreover, the technology is moreover very heterogeneous, such that whereas visceral applications (e.g., “killer robots”) may generate public opprobrium and restrictive activism, more diffuse or less kinetic ones (e.g., logistics systems; capabilities to track missile submarines) may not. It would therefore be advisable that organizations pursuing bans of the technology, consider the degree to which framings of “killer robots” continue to correspond to developments in military AI, including other usages which are potentially unethical, unsafe, or destabilizing.

Finally, while public norms or activism against military AI may strengthen domestic political coalitions already opposed to these weapons, they alone are not always able to sway policymakers in the first place. A key route lies therefore in shaping policymakers’ norms (and indirectly the domestic political landscapes). This relies on the (top-down) norm-shaping influence exerted by global legal instruments and regimes, but also on the (bottom-up) institutionalization of norms by “elite entrepreneurship in norm change” (Lantis, 2018), and specifically through “epistemic communities” of expert groups.

Early efforts by coordinated “epistemic communities” can spur arms control efforts

Is bottom-up norm institutionalization possible? In an era of slow action on certain global challenges such as climate change, it may appear as if states are not easily swayed by expert advice. Yet the history of nuclear arms control offers an existence proof of a national research community reaching an early consensus around the implications of a new technology, disseminating these ideas amongst policymakers in key states, and eventually laying the foundations for a bilateral arms control agreement to curb vertical proliferation, in the shape of the 1972 Anti-Ballistic Missile (ABM) Treaty (SALT-I). This argument draws on the notion of an “epistemic community” (cf. Haas, 1992), described by Adler (1992) as:

... a network of individuals or groups with an authoritative claim to policy-relevant knowledge within their domain of expertise. The community members

share knowledge about the causation of social and physical phenomena in an area for which they have a reputation for competence, and they have a common set of normative beliefs about what will benefit human welfare in such a domain. While members are often from a number of different professions and disciplines, they adhere to the following: (1) shared consummatory values and principled beliefs; (2) shared causal beliefs or professional judgment; (3) common notions of validity based on intersubjective, internally defined criteria for validating knowledge; (4) a common policy project. (nn. 1, 101)

Adler charts how throughout the 1950s–1960s, a community of scientists and strategists—at places such as RAND, Harvard and MIT, and the Presidential Science Advisory Committee (PSAC)—developed the “imaginary” science of nuclear strategy. “Imaginary,” because this science involved theorizing about hypothetical war scenarios in the absence (or, it was feared, in advance) of actual empirical experience with nuclear war, by exploring how new weapons or technologies might deter or invite attack, and stabilize or destabilize relations (Adler, 1992, pp. 107–109). Through this work, the community developed a conviction that, contrary to prevailing faith in technological supremacy as an unalloyed strategic good, the development of effective anti-ballistic missile (ABM) systems might create immense risks. Such systems would not be able to credibly shield the United States against the might of a full Soviet surprise attack; but they might enable the United States to itself carry out a first strike, and then intercept the much-weakened Soviet retaliation. The deployment of nominally “defensive” ABM systems therefore perversely created an incentive to pre-empt, destabilizing deterrence.

As a result, the epistemic community argued that both the security interests of the United States and the global chance of avoiding nuclear war would be improved if the superpowers engaged in stabilizing arms control agreements. In their position as visible authorities with links to policymakers—and strengthened by a political climate of public rallies and grass-roots groups protesting the deployment of ABM—the community members disseminated this new understanding of deterrence dynamics to policymakers.

In this, the community was able to present arms control (and specifically the restriction of ABMs) as a politically viable “middle ground” alternative to the then-prevalent camps which advocated pursuing either full nuclear disarmament or absolute nuclear supremacy over the Soviets (Adler, 1992, p. 113). By persuading key figures such as Presidents Eisenhower and Kennedy as well as Robert McNamara, the community proved able to institutionalize norms and shift policy (Adler, 1992, p. 127). Moreover, throughout the 1950s and 1960s, the epistemic community utilized international summit meetings and scientific fora to aid in the diffusion of these arms control ideas and norms to global peers, including in the Soviet Union, setting the stage for the negotiation process that led directly towards the 1972 ABM Treaty (Adler,

1992, pp. 133–140). While some of the community’s original expectations were altered at the bargaining table, their consensus still formed the conceptual basis of the final agreement.

The above is a brief, necessarily insufficient summary of a complex historical process. Nonetheless, what lessons might we cautiously draw for governance strategies aimed at preventing the (vertical) proliferation of advanced AI weapons? Optimistically, the ABM Treaty suggests that early convergence of and action by a national epistemic community can frame the political “centre ground,” and steer policy action at an early goal—even if the risks are entirely theoretical. It suggests that such shared expectations can be diffused to epistemic communities in other key states, and that the resulting consensus can serve as a foundation for bilateral arms control efforts, even between otherwise distrustful or hostile states. It is also possible that discourse around arms control is path-dependent, and that concerted action at an early stage (i.e., before key actors have already deployed certain AI weapons as a linchpin of their strategies) can frame the terms of subsequent discourse. This suggests that there is a time premium on acting early.

However, there are also a number of pessimistic lessons, indicating the fragility of arms control efforts: The process in which an epistemic community articulates, institutionalizes and disseminates norms is slow, and appears vulnerable to sudden derailment by subsequent changes in the national administration or the international mood. By the early 1980s, missile defense was again briefly pursued as of President Reagan’s Strategic Defense Initiative. Moreover, the ABM Treaty itself was terminated after the United States decided to withdraw in 2002—which President Putin has identified as the reason behind Russia’s decision to covertly develop a range of exotic next-generation nuclear delivery systems (Putin, 2018). Moreover, the current climate for arms control agreements seems grim: as of October 2018, the United States announced that it would withdraw from the 1987 Intermediate-Range Nuclear Forces Treaty (Sanger & Broad, 2018); earlier that year, the Deep Cuts Commission cautioned that if the 2010 U.S.-Russia New Start treaty is not extended before it expires in February 2021, this will mark the first time since 1972 that both arsenals are entirely unconstrained by agreements (Deep Cuts Commission, 2018). Such trends illustrate how the prospect and stability of many arms control regimes depend on healthy international relations more than that they precede them—and that such agreements may begin to fray at exactly the tense moments they are the most needed (Fatton, 2016).

More generally, this complex record indicates that the success of an epistemic community is highly sensitive to specific and contingent historical circumstances—such as the degree to which a technology is still in development or already deployed and relied upon in the field; the distribution of power amongst the key parties, domestic political sentiment, or prevailing popular

and cultural perceptions of the technology. The anti-ABM arms control community proved victorious in the 1960s and 1970s; but earlier quests for nuclear arms control, such as the 1946 Baruch Plan, had failed to get traction (Baratta, 2004), and the anti-ABM community might not have had such success had they missed their narrow window of opportunity.

Therefore, it bears asking: Given the high sensitivity to historically contingent circumstances, are there any transferable lessons in the ABM Treaty for an epistemic community which would pursue the restriction of military AI today? While a military AI epistemic community cannot choose its “external” historical context, the successful experience of nuclear arms control can at least provide lessons for “internal” contributors to success. This includes ways in which a community can organize itself to articulate and converge on a shared policy project; as well as strategies to further the intermediate goals of intellectual innovation, domestic norm institutionalization, and global norm dissemination, which lay the fertile ground for later cooperation.

This comparison also helps assess the viability of an AI military epistemic community today. One critical consideration for the viability of an epistemic community is the degree to which relevant (technical and policy) experts in the field already possess shared principled or causal beliefs about the technology—that is, whether there exists a sufficient consensus (whether on ethical, legal, strategic or operational grounds) against the use of (specific) AI systems in warfare. The effective threshold may be high: for instance, Herman Kahn claimed that throughout the 1960s, roughly 90% of all U.S. experts consulting for the government opposed ABM systems (as cited in Adler, 1992, n. 117; Kahn, 1969, p. 285). It is unclear if such consensus is in reach for military AI, though it may exist amongst the private AI industry, as indicated by their participation in the 2017 Open Letter to the UN CCCW, and the 2018 internal employee protests at Google over its involvement in the MAVEN program (“2017 Open Letter to the United Nations Convention on Certain Conventional Weapons,” 2017; “Letter to Google C.E.O.,” 2018). However, it remains unclear to what degree these convictions are shared by relevant experts within policymaking circles themselves.

In sum, a coherent and sufficiently well-situated epistemic community on military AI does not yet appear to exist today—in fact, Payne has suggested that “the [AI] epistemic community is physically fragmented, notwithstanding efforts to coordinate resistance to the weaponisation of AI” (Payne, 2018a, p. 16). For those seeking to rally these stakeholders and build an effective community that can effectively shift the policy Overton window on military AI, reviewing the historical success of the ABM Treaty can at least suggest some of the salient steps to take in terms of mobilization, intellectual innovation, domestic norm institutionalization, and global norm dissemination.

“Normal accidents” in complex systems render “meaningful human control” meaningless as governance goal

The third, more pessimistic point concerns the limits of operational safety for certain technologies, which has implications for what kind of goals or criteria governance regimes should aim at.

In recent years, much of the discussion around military uses of AI—specifically lethal autonomous weapons systems—has converged around the notion of ensuring “meaningful human control” over these systems (Crootof, 2015; Human Rights Watch, 2016; cf. also Horowitz & Scharre, 2015). Some have previously suggested that ensuring this might not always be tactically feasible, since “the security dilemma and the evident tactical advantages of rapid automated decisions make keeping a ‘man in the loop’, or even ‘on the loop’, problematic” (Payne, 2018a, p. 18). Yet as the history of nuclear weapons shows, there is a more fundamental challenge to the idea of ensuring meaningful human control, in the form of emergent “normal” accidents.

In spite of the extremely high stakes involved, nuclear weapons and their command and control systems have a history of frequent accidents and close calls, both during and since the Cold War. Some of these—such as the 1961 Goldsboro B-52 crash (Jones, 1969) or the 1980 “Damascus Incident” Titan-II explosion—carried risks of an accidental nuclear detonation (Sagan, 1993; Schlosser, 2014). More frighteningly, on both the United States and USSR/Russian sides, human- and computer errors on repeated occasions generated false positive signals of an incoming nuclear strike—credible signals which could (and, in some cases, by normal operational protocol “should”) have led to nuclear war (Baum, de Neufville, & Barrett, 2018; Borrie, 2014; Lewis, Williams, Pelopidas, & Aghlani, 2014; Schlosser, 2014).

In an influential 1993 study, Sagan (1993) explained the track record of repeated U.S. nuclear weapon incidents throughout the Cold War by arguing that nuclear forces operate their technologies and organizations in a manner that makes them vulnerable to so-called “normal accidents.” Developed by sociologist Perrow (1984), normal accident theory (NAT) explores how catastrophic accidents inevitably—as a “normal” consequence of how the system is set up—emerge at the interface of mechanical-, software-, operator- and organizational failures. It is this systemic perspective that makes normal accident theory transferable: It does not focus on vulnerabilities in one specific technology, but instead looks at key operational features which render a wide range of systems—from nuclear reactors to aircraft and from spacecraft to algorithmic trading systems (Perrow, 1984, p. 23; Scharre, 2016a, pp. 30–33)—susceptible to catastrophic accidents.

It has been previously argued that Autonomous Weapons might be susceptible to normal accidents (Borrie, 2016; Scharre, 2016a)—or indeed, that many types of military AI systems may meet the operational criteria of NAT,

perhaps more so even than past “textbook” NAT technologies (Maas, 2018). What are these features, and how do they apply to military AI?

In the first place, the system is opaque and has high *interactive complexity*—which is less a feature of the system’s size or its number of component units, and more a product of the way these components are connected in a manner that ensures they interact in unexpected, non-linear ways that are not visible or immediately comprehensible (Sagan, 1993, p. 15). Interactive complexity ensures operators are unable to meaningfully understand or anticipate the system’s behavior at a given moment, especially in new environments or to unanticipated inputs; it also reduces operator ability to detect or isolate an error when it does emerge. Like the complex nuclear command and control networks plugged into diverse radar, sensor and communication arrays, leading AI architectures such as deep learning can be complex, opaque and unpredictable; indeed, it may be impossible to either produce a formal proof of their behavior, or exhaustively test for all real-world scenarios during training (Borrie, 2016, pp. 8–9).

In the second place, the system is *tightly coupled*, which means that “there is no slack or buffer or give between two items. What happens in one directly affects what happens in the other” (Perrow, 1984, pp. 89–90). This ensures that errors cascade rapidly through a system, and are difficult to see or contain before they trigger catastrophic outcomes. Moreover, tight coupling also entails that, perversely, adding more redundancies or fail-safes may increase operational risk. This is because adding components further increases a system’s interactive complexity, and creates vulnerability to “common-mode failures,” where a single external shock—say, a hack—both triggers an error cascade, and disables nominally “independent” alarm or backup systems which the operator relies upon. Providing such safety measures can also induce operator risk homeostasis—where they are willing to behave more recklessly with a “safe” system—as well as automation bias—where the operator trusts the system overmuch, and defers to it, undercutting the efficacy of maintaining a human “in the loop.” Many military AIs—particularly in cyberspace—will operate at superhuman speed or scale, and rely on tightly coupled networks of sensors, modules, actuators, making them tightly coupled—and teetering on the edge of “a million mistakes a second” (Scharre, 2018b).

In the third place, the organizations operating the technology have *multiple competing objectives* beyond “safety.” Some strategic or tactical goals can directly increase the risk of accidents by requiring tight coupling: in the case of U.S. nuclear forces, the strategic objective of “launch-on-warning” within minutes of receiving indications of an incoming strike cuts short the time window within which to verify that alert. Other goals, such as the military pursuit of secrecy or the bureaucratic pressure to downplay safety incidents lest this threatens the organizational reputation, inhibit an

organization's ability to acknowledge and learn from past incidents—or, if it does learn, inhibits their willingness to share these experiences and best practices with other military branches, let alone with adversaries, operating the same risk-prone weapons. The designers, trainers and military users of AI weapon systems will likewise have a range of operational (external) and bureaucratic (internal) objectives, many of which will be orthogonal or even anti-correlated to “safety.”

Fourthly, all of these risks are exacerbated in a *competitive context*, where there is a premium on rapid operational speed and pre-emption, and a large space for miscommunication or miscalculation between parties. This might hold even more for deployed tactical AI weapons than for strategic nuclear forces. After all, in a battlefield context military AIs must base decisions on incomplete, messy, potentially unreliable data; there is a tactical premium on increasing speed or system reaction to adversary AI systems; and adversaries have an incentive to hack or spoof the system, directly or through adversarial input (Goodfellow et al., 2017). These factors increase the risk of unexpected behavior or accident cascades amongst interacting military AI systems—particularly in cyberspace (Schneider, 2016)—with the risk of a potential “flash war” (Scharre, 2016b), analogous to the algorithmic flash crashes observed in the financial sector.

In spite of the extremely high stakes, the history of nuclear weapons shows that we are unable to avert emergent accident cascades in technologies that have high interactive complexity and tight coupling, when their handlers have multiple objectives and operate in competitive environments. And yet, these nuclear command and control systems will look quaint and simple compared to the sheer scope and range of uses in which different AI systems will come to be employed, the ways in which they will be networked, and their interactions with new (including malicious) inputs or environments, exponentially increasing both the number and range of system interactions.

The governance and policy implications of normal accident risks in military AI systems are diverse. One takeaway however concerns the aforementioned notion, central to the ongoing debates around the regulation of autonomous weapon, that retaining a human “-in-the-loop” might enable us to preserve “meaningful human control” over military AI systems, and operate these (more) safely. The propensity of military AI systems towards “normal” accident, and the propensity of human operators towards automation bias, offer a much deeper challenge to the viability, utility, or sufficiency of such claims or aims. They suggest that much of the current focus of the epistemic community, on ensuring that kinetic “killer robots” remain subject to “meaningful human control,” misses the mark, twice—once with regards to the actual range of military AI systems which may pose hazards; and again in putting overdue faith in the solution of “human control.” Assurances of such control are often spurious in practice, and

should not serve as cornerstone of arms control agreements or governance efforts.

Conclusion

How viable is international arms control for military AI? Recent years have seen no shortage of alarmism or pessimism regarding the imminence of AI arms races, and yet the historical track record in slowing and containing the spread of “the ultimate deterrent” suggests a surprisingly optimistic answer: Arms races are not inevitable, but can be managed, channeled or even stopped. This may be facilitated through direct engagement with domestic political coalitions, or indirectly, by shaping norms top-down (through international regimes) or bottom-up (through epistemic communities).

The study of the road to the 1972 ABM Treaty likewise suggests, optimistically, that small communities of experts, appropriately organized and mobilized, can have a disproportionate effect in framing global arms control cooperation through bottom-up norm institutionalization, and could do so again for military AI. At the same time, this reading comes with a note of caution: Organizing an effective epistemic community in the field of military AI may be difficult, given the limited consensus; moreover, the window of opportunity to coordinate this community and institutionalize global norms on military AI may already be closing. This is specifically problematic since many of the current debates in the nascent epistemic community remain aimed at overtly narrow goals such as “human-in-the-loop” arrangements or assurances of “meaningful human control”—arrangements which are rendered problematic in the context of networked AI systems that are deployed in ways that render them susceptible to “normal accidents.”

This argument has practical implications for actors seeking to pursue arms control. It suggests that there are real prospects of halting proliferation, or channeling it in less hazardous directions. However, it also suggests that the nascent epistemic community seeking to combat military AI must look beyond its focus on banning “killer robots” on ethical and legal grounds alone. While valid and worthwhile, a more effective and comprehensive arms control effort against military AI might see this epistemic community re-organize and readjust in two ways: First, it should consider—and engage with the far broader range of prospective military AI systems than kinetic autonomous weapons systems alone. Secondly, it should explicitly bring to bear the full portfolio of rationales that have driven arms control historically—rationales that include not just ethics and legality, but also strategic stability and safety. Such a portfolio approach has a higher chance of affecting policy, by offering rationales (and arguments) more salient to domestic political actors sympathetic to controlling military AI. This can aid in the

dissemination and institutionalization of norms, as well as targeting control efforts on the full spectrum of risks posed by military AI systems.

Finally, this argument suggests some ways by which to update and extend established theories in international relations—whether those examining the causes of state decision-making; the top-down and bottom-up processes of norm creation, dissemination and institutionalization, and literatures on organizational safety in the context of complex technologies. This can help keep these fields current in the era of AI; if the arms control challenge is greater this time, that provides all the more reason for AI governance scholars and advocates to understand the successes and failures of the past. This article has only offered a preliminary examination of such lessons, and much more work is needed to adequately identify which avenues of global governance are potentially fruitful and which might prove dead ends. The ability to learn from experience may be one clear advantage AI arms control advocates have over the nuclear arms control advocates of the past. They would do well to make full use of this history, whether in preparing to study or make this next chapter in arms control.

Acknowledgements

The author would like to thank the participants of the *Cambridge Conference on Catastrophic Risk (CCCR)* 2016, as well as Hin-Yan Liu, Sophie-Charlotte Fischer, Peter Cihon, Jade Leung, the editor, and two anonymous reviewers for valuable feedback on the argument throughout the process. He would also like to thank Emma Dam Olesen for support in preparing the manuscript.

Disclosure statement

No potential conflict of interest was reported by the author.

Notes on contributor

Matthijs M. Maas is a PhD Fellow at the University of Copenhagen (Centre for International Law, Conflict and Crisis, Faculty of Law), and a Research Affiliate with the Center for the Governance of AI (Future of Humanity Institute, University of Oxford). His research focuses, amongst others, on exploring effective and legitimate global governance strategies for emerging technologies, particularly artificial intelligence.

ORCID

Matthijs M. Maas  <http://orcid.org/0000-0002-6170-9393>

Reference list

- 2017 Open Letter to the United Nations Convention on Certain Conventional Weapons. (2017). Retrieved from <https://www.dropbox.com/s/g4ijcaqq6ivq19d/2017%20Open%20Letter%20to%20the%20United%20Nations%20Convention%20on%20Certain%20Conventional%20Weapons.pdf?dl=0>
- Adler, E. (1992). The emergence of cooperation: National epistemic communities and the international evolution of the idea of nuclear arms control. *International Organization*, 46, 101–145. doi:10.1017/S0020818300001466
- Allen, G., & Kania, E. (2017, September 8). China is using America’s own plan to dominate the future of artificial intelligence. *Foreign Policy*. Retrieved from <https://foreignpolicy.com/2017/09/08/china-is-using-americas-own-plan-to-dominate-the-future-of-artificial-intelligence/>
- Allison, G. (2010, January/February). Nuclear disorder. *Foreign Affairs*. Retrieved from <https://www.foreignaffairs.com/articles/pakistan/2010-01-01/nuclear-disorder>
- Amodei, D., & Hernandez, D. (2018, May 16). AI and compute. Retrieved from <https://blog.openai.com/ai-and-compute/>
- Auslin, M. (2018, October 19). Can the pentagon win the AI arms race? *Foreign Affairs*. Retrieved from <https://www.foreignaffairs.com/articles/united-states/2018-10-19/can-pentagon-win-ai-arms-race>
- Ayoub, K., & Payne, K. (2016). Strategy in the age of artificial intelligence. *Journal of Strategic Studies*, 39, 793–819. doi:10.1080/01402390.2015.1088838
- Baratta, J. P. (2004). *The politics of world federation. Vol. 1: The United Nations, U.N. reform, atomic control*. Westport: Praeger.
- Barnes, J. E., & Chin, J. (2018, March 2). The new arms race in AI. *Wall Street Journal*. Retrieved from <https://www.wsj.com/articles/the-new-arms-race-in-ai-1520009261>
- Baum, S., de Neufville, R., & Barrett, A. (2018). *A model for the probability of nuclear war* (SSRN Scholarly Paper No. ID 3137081). Rochester, NY: Social Science Research Network. Retrieved from <https://papers.ssrn.com/abstract=3137081>
- Borghard, E. D., & Lonergan, S. W. (2018, January 16). Why are there no cyber arms control agreements? Retrieved from <https://www.cfr.org/blog/why-are-there-no-cyber-arms-control-agreements>
- Borrie, J. (2014). A limit to safety: Risk, “normal accidents”, and nuclear weapons. *ILPI-UNIDIR Vienna Conference Series*. Retrieved from <http://www.isn.ethz.ch/Digital-Library/Publications/Detail/?ots591=0c54e3b3-1e9c-be1e-2c24-a6a8c7060233&lng=en&id=186094>
- Borrie, J. (2016). *Safety, unintentional risk and accidents in the weaponization of increasingly autonomous technologies* (UNIDIR Resources No. 5). Geneva. UNIDIR. Retrieved from <http://www.unidir.org/files/publications/pdfs/safety-unintentional-risk-and-accidents-en-668.pdf>
- Boulanin, V., & Verbruggen, M. (2017). *Mapping the development of autonomy in weapon systems*. Stockholm: Stockholm International Peace Research Institute. Retrieved from https://www.sipri.org/sites/default/files/2017-11/siprireport_mapping_the_development_of_autonomy_in_weapon_systems_1117_1.pdf
- Carranza, M. E. (2018). Deterrence or taboo? Explaining the non-use of nuclear weapons during the Indo-Pakistani post-tests nuclear crises. *Contemporary Security Policy*, 39, 441–463. doi:10.1080/13523260.2017.1418725
- Cave, S., & Ó hÉigeartaigh, S. S. (2018). An AI race for strategic advantage: Rhetoric and risks. In *AAAI / ACM Conference on Artificial Intelligence, Ethics and Society*.

- Retrieved from http://www.aies-conference.com/wp-content/papers/main/AIES_2018_paper_163.pdf
- China's State Council. (2017). *A next generation artificial intelligence development plan* (R. Creemers, G. Webster, P. Triolo, & E. Kania, Trans.). New America Cybersecurity Initiative. Retrieved from <https://na-production.s3.amazonaws.com/documents/translation-fulltext-8.1.17.pdf>
- Crootof, R. (2015). A meaningful floor for “meaningful human control”. *Temple International and Comparative Law Journal*, 30, 53–62.
- Dafoe, A. (2018). *AI governance: A research agenda*. Oxford: Governance of AI Program, Future of Humanity Institute. Retrieved from <https://www.fhi.ox.ac.uk/govaiagenda/>
- Danzig, R. (2018). *Technology roulette: Managing loss of control as many militaries pursue technological superiority*. Washington, DC: Center for a New American Security. Retrieved from <https://s3.amazonaws.com/files.cnas.org/documents/CNASReport-Technology-Roulette-DoSproof2v2.pdf?mtime=20180628072101>
- Davison, N. (2017). A legal perspective: Autonomous weapon systems under international humanitarian law. In *Perspectives on lethal autonomous weapon systems* (pp. 5–18). New York: UNODA.
- Deep Cuts Commission. (2018, March 19). Urgent steps to avoid a new nuclear arms race and dangerous miscalculation – Statement of the deep cuts commission. Retrieved from https://www.armscontrol.org/sites/default/files/files/documents/DCC_1804018_FINAL.pdf
- Drum, K. (2018, August). Tech world: Welcome to the digital revolution. *Foreign Affairs*.
- European Parliament. (2018, September 12). P8_TA-PROV(2018)0341: European Parliament resolution of 12 September 2018 on autonomous weapon systems (2018/2752(RSP)). European Parliament. Retrieved from <http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//NONSGML+TA+P8-TA-2018-0341+0+DOC+PDF+V0//EN>
- Fatton, L. P. (2016). The impotence of conventional arms control: Why do international regimes fail when they are most needed? *Contemporary Security Policy*, 37, 200–222. doi:10.1080/13523260.2016.1187952
- Fischer, S.-C. (2017). The role of the private sector in the governance of autonomous weapon systems: A principal-agent perspective. Presented at the We Robot 2017. Retrieved from <http://www.werobot2017.com/wp-content/uploads/2017/03/Fischer-The-role-of-the-private-sector-in-the-governance-of-autonomous-weapon-systems-1.pdf>
- Fuhrmann, M., & Lupu, Y. (2016). Do arms control treaties work? Assessing the effectiveness of the nuclear nonproliferation treaty. *International Studies Quarterly*, 60, 530–539. doi:10.1093/isq/sqw013
- Garnett, S. W. (2012). The “Model” of Ukrainian Denuclearization. In Jeffrey W. Knopf (Ed.), *Security assurances and nuclear nonproliferation* (pp 246–274). Stanford, CA: Stanford University Press.
- Geist, E., & Lohn, A. J. (2018). *How might artificial intelligence affect the risk of nuclear war?* (p. 28). Web. RAND. Retrieved from <https://www.rand.org/pubs/perspectives/PE296.html>
- Geist, E. M. (2016). It's already too late to stop the AI arms race—We must manage it instead. *Bulletin of the Atomic Scientists*, 72, 318–321.

- Goodfellow, I., Papernot, N., Huang, S., Duan, Y., Abbeel, P., & Clark, J. (2017, February 16). Attacking machine learning with adversarial examples. Retrieved from <https://openai.com/blog/adversarial-example-research/>
- Guterres, A. (2018, November). *Allocution du Secrétaire général au Forum de Paris sur la paix*. Paris. Retrieved from <https://www.un.org/sg/en/content/sg/statement/2018-11-11/allocution-du-secr%C3%A9taire-g%C3%A9n%C3%A9ral-au-forum-de-paris-sur-la-paix>
- Haas, P. M. (1992). Introduction: Epistemic communities and international policy coordination. *International Organization*, 46(1), 1–35. doi:10.1017/S0020818300001442
- Hogarth, I. (2018, June 13). AI nationalism. Retrieved from <https://www.ianhogarth.com/blog/2018/6/13/ai-nationalism>
- Horowitz, M. C. (2016). Public opinion and the politics of the killer robots debate. *Research & Politics*, 3(1). doi:10.1177/2053168015627183
- Horowitz, M. C. (2018a, May 15). Artificial intelligence, international competition, and the balance of power. *Texas National Security Review*. Retrieved from <https://tnsr.org/2018/05/artificial-intelligence-international-competition-and-the-balance-of-power/>
- Horowitz, M. C. (2018b, September 12). The algorithms of august. *Foreign Policy*. Retrieved from <https://foreignpolicy.com/2018/09/12/will-the-united-states-lose-the-artificial-intelligence-arms-race/>
- Horowitz, M. C., & Scharre, P. (2015). *Meaningful human control in weapon systems: A primer*. Center for a New American Security. Retrieved from https://s3.amazonaws.com/files.cnas.org/documents/Ethical_Autonomy_Working_Paper_031315.pdf?mtime=20160906082316
- Human Rights Watch. (2012). *Losing humanity: The case against killer robots*. Amsterdam/Berlin: Human Rights Watch. Retrieved from https://www.hrw.org/sites/default/files/reports/arms1112_ForUpload.pdf
- Human Rights Watch. (2016, April 11). Killer robots and the concept of meaningful human control. Retrieved from <https://www.hrw.org/news/2016/04/11/killer-robots-and-concept-meaningful-human-control>
- Hwang, T. (2018). *Computational power and the social impact of artificial intelligence* (SSRN Scholarly Paper No. ID 3147971). Rochester, NY: Social Science Research Network. Retrieved from <https://papers.ssrn.com/abstract=3147971>
- IPSOS. (2019). *Six in ten (61%) respondents across 26 countries oppose the use of lethal autonomous weapons systems*. Retrieved from https://www.ipsos.com/sites/default/files/ct/news/documents/2019-01/human-rights-watch-autonomous-weapons-pr-01-22-2019_0.pdf
- Jo, D.-J., & Gartzke, E. (2007). Determinants of nuclear weapons proliferation. *Journal of Conflict Resolution*, 51, 167–194. doi:10.1177/0022002706296158
- Jones, P. F. (1969, October 22). Goldsboro revisited: Account of hydrogen bomb near-disaster over North Carolina – declassified document. Retrieved from <http://www.theguardian.com/world/interactive/2013/sep/20/goldsboro-revisited-declassified-document>
- Joshi, S. (2019, January 17). Autonomous weapons and the new laws of war. *The Economist*. Retrieved from <https://www.economist.com/briefing/2019/01/19/autonomous-weapons-and-the-new-laws-of-war>
- Kahn, H. (1969). The missile defense debate in perspective. In J. J. Holst & W. Schneider (Eds.), *Why ABM: Policy issues in the missile defense controversy* (pp. 285–294). Pergamon. doi:10.1016/B978-0-08-015625-5.50017-1

- Kania, E. (2018, April 19). The pursuit of AI is more than an arms race. *Defense One*. Retrieved from <https://www.defenseone.com/ideas/2018/04/pursuit-ai-more-arms-race/147579/>
- Kania, E. B. (2017). *Battlefield singularity: Artificial intelligence, military revolution, and China's future military power*. Washington, DC: Center for a New American Security. Retrieved from <https://s3.amazonaws.com/files.cnas.org/documents/Battlefield-Singularity-November-2017.pdf?mtime=20171129235804>
- Knopf, J. W. (2018). After diffusion: Challenges to enforcing nonproliferation and disarmament norms. *Contemporary Security Policy*, 39, 367–398. doi:10.1080/13523260.2018.1431446
- Krasner, S. D. (1982). Structural causes and regime consequences: Regimes as intervening variables. *International Organization*, 36, 185–205. doi:10.1017/S0020818300018920
- Kroenig, M., & Gopalaswamy, B. (2018, November 12). Will disruptive technology cause nuclear war? Retrieved from <https://thebulletin.org/2018/11/will-disruptive-technology-cause-nuclear-war/>
- Lantis, J. S. (2018). Nuclear cooperation with non-NPT member states? An elite-driven model of norm contestation. *Contemporary Security Policy*, 39, 399–418. doi:10.1080/13523260.2017.1398367
- Lee, K.-F. (2018). *AI superpowers: China, silicon valley, and the new world order*. Boston: Houghton Mifflin Harcourt.
- Letter to Google C.E.O. (2018). Retrieved from <https://static01.nyt.com/files/2018/technology/googleletter.pdf>
- Lewis, P., Williams, H., Pelopidas, B., & Aghlani, S. (2014). *Too close for comfort: Cases of near nuclear use and options for policy*. London: Chatham House.
- Lieber, K. A., & Press, D. G. (2017). The new era of counterforce: Technological change and the future of nuclear deterrence. *International Security*, 41(4), 9–49. doi:10.1162/ISEC_a_00273
- Maas, M. (2018). Regulating for ‘normal AI accidents’—Operational lessons for the responsible governance of AI deployment. Presented at the AAAI / ACM Conference on Artificial Intelligence, Ethics and Society, New Orleans: Association for the Advancement of Artificial Intelligence. Retrieved from http://www.aies-conference.com/wp-content/papers/main/AIES_2018_paper_118.pdf
- Maas, M., Sweijs, T., & De Spiegeleire, S. (2017). *Artificial intelligence and the future of defense: Strategic implications for small- and medium-sized force providers*. The Hague, The Netherlands: The Hague Centre for Strategic Studies. Retrieved from <http://hcss.nl/report/artificial-intelligence-and-future-defense>
- MacKenzie, D., & Spinardi, G. (1995). Tacit knowledge, weapons design, and the uninvention of nuclear weapons. *American Journal of Sociology*, 101(1), 44–99. doi:10.1086/230699
- McNamara, R. (1963, February 12). “The diffusion of nuclear weapons with and without a test ban agreement.” *Memorandum for the President, Washington, DC, Department of Defense*, 12 February 1963, Digital National Security Archive (DNSA), document no. NP00941.
- Müller, H., & Schmidt, A. (2010). The little known story of de-proliferation: Why states give up nuclear weapon activities. In W. C. Potter & G. Mukhatzhanova (Eds.), *Forecasting nuclear proliferation in the 21st century. The role of theory* (Vol. 1, pp. 124–158). Stanford, CA: Stanford University Press.
- Nehal, B., Beck, S., Geiss, R., Liu, H.-Y., & Kress, K. (Eds.). (2016). *Autonomous weapons systems: Law, ethics, policy*. Cambridge: Cambridge University Press.

- Open Roboethics Initiative. (2015). *The ethics and governance of lethal autonomous weapons systems: An international public opinion poll*. Open Roboethics Initiative. Retrieved from http://www.openroboethics.org/wp-content/uploads/2015/11/ORi_LAWS2015.pdf
- Payne, K. (2018a). Artificial intelligence: A revolution in strategic affairs? *Survival*, 60(5), 7–32. doi:10.1080/00396338.2018.1518374
- Payne, K. (2018b). *Strategy, evolution, and war: From apes to artificial intelligence*. Washington, DC: Georgetown University Press. Retrieved from <http://ebookcentral.proquest.com/lib/kbdk/detail.action?docID=5394997>
- Pelopidas, B. (2011). The oracles of proliferation: How experts maintain a biased historical reading that limits policy innovation. *The Nonproliferation Review*, 18, 297–314. doi:10.1080/10736700.2011.549185
- Perrow, C. (1984). Normal accidents: Living with high risk technologies. Retrieved from <http://press.princeton.edu/titles/6596.html>
- Picker, C. B. (2001). A view from 40,000 feet: International law and the invisible hand of technology. *Cardozo Law Review*, 23, 151–219.
- Pinker, S. (2011). *The better angels of our nature: Why violence has declined*. New York, NY: Viking.
- Putin, V. (2017, September 1). Открытый урок «Россия, устремлённая в будущее» | Open Lesson “Russia moving towards the future.” Retrieved from <http://kremlin.ru/events/president/news/55493>
- Putin, V. (2018). 2018 presidential address to the federal assembly. President of Russia. Retrieved from <http://en.kremlin.ru/events/president/news/56957>
- Rickli, J.-M. (2017). Artificial intelligence and the future of warfare. In *WEF global risks report 2017* (p. 49). Retrieved from http://www3.weforum.org/docs/GRR17_Report_web.pdf
- Roff, H. M. (2014). The strategic robot problem: Lethal autonomous weapons in war. *Journal of Military Ethics*, 13, 211–227. doi:10.1080/15027570.2014.975010
- Roff, H. M. (2017, February 8). What do people around the world think about killer robots? *Slate*. Retrieved from http://www.slate.com/articles/technology/future_tense/2017/02/what_do_people_around_the_world_think_about_killer_robots.html
- Rublee, M. R., Bertsch, G., & Wiarda, H. (2009). *Nonproliferation norms: Why states choose nuclear restraint*. Athens, GA: University of Georgia Press.
- Rublee, M. R., & Cohen, A. (2018). Nuclear norms in global governance: A progressive research agenda. *Contemporary Security Policy*, 39, 317–340. doi:10.1080/13523260.2018.1451428
- Sagan, S. D. (1993). *The limits of safety: Organizations, accidents, and nuclear weapons*. Princeton, NJ: Princeton University Press.
- Sagan, S. D. (1996). Why do states build nuclear weapons?: Three models in search of a bomb. *International Security*, 21(3), 54–86. doi:10.2307/2539273
- Sagan, S. D. (2011). The causes of nuclear weapons proliferation. *Annual Review of Political Science*, 14, 225–244. doi:10.1146/annurev-polisci-052209-131042
- Sanger, D. E., & Broad, W. J. (2018, October 20). U.S. to tell Russia it is leaving landmark I.N.F. treaty. *The New York Times*. Retrieved from <https://www.nytimes.com/2018/10/19/us/politics/russia-nuclear-arms-treaty-trump-administration.html>
- Scharre, P. (2016a). *Autonomous weapons and operational risk* (Ethical Autonomy Project. 20YY Future of Warfare Initiative). Washington, DC: Center for a New American Security. Retrieved from https://s3.amazonaws.com/files.cnas.org/documents/CNAS_Autonomous-weapons-operational-risk.pdf

- Scharre, P. (2016b, April). *Flash war – Autonomous weapons and strategic stability*. Presented at the Understanding Different Types of Risk, Geneva. Retrieved from <http://www.unidir.ch/files/conferences/pdfs/-en-1-1113.pdf>
- Scharre, P. (2018a). *Army of none: Autonomous weapons and the future of war* (1st ed.). New York, NY: W. W. Norton & Company.
- Scharre, P. (2018b, September 12). A million mistakes a second. *Foreign Policy*. Retrieved from <https://foreignpolicy.com/2018/09/12/a-million-mistakes-a-second-future-of-war/>
- Scherer, M. U. (2016). Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies. *Harvard Journal of Law & Technology*, (2). Retrieved from <http://jolt.law.harvard.edu/articles/pdf/v29/29HarvJLTech353.pdf>
- Schlosser, E. (2014). *Command and control: Nuclear weapons, the damascus accident, and the illusion of safety* (Reprint ed.). New York, NY: Penguin Books.
- Schneider, J. (2016). *Digitally-enabled warfare: The capability-vulnerability paradox* (p. 15). Washington, DC: Center for a New American Security. Retrieved from <https://www.cnas.org/publications/reports/digitally-enabled-warfare-the-capability-vulnerability-paradox>
- Shultz, G. P. (1984, November 1). *Preventing the proliferation of nuclear weapons*. Washington, DC: U.S. Department of State, Bureau of Public Affairs, Office of Public Communication, Editorial Division.
- Solingen, E. (1994). The political economy of nuclear restraint. *International Security*, 19(2), 126–159. doi:10.2307/2539198
- Thompson, N., & Bremmer, I. (2018, October 23). The AI cold war that threatens us all. *Wired*. Retrieved from <https://www.wired.com/story/ai-cold-war-china-could-doom-us-all/>
- van der Meer, S. (2011). Not that bad: Looking back on 65 years of nuclear non-proliferation efforts. *Security and Human Rights*, 22(1), 37–47. doi:10.1163/187502311796365862
- van der Meer, S. (2014). Forgoing the nuclear option: States that could build nuclear weapons but chose not to do so. *Medicine, Conflict and Survival*, 30(S1), s27–s34. doi:10.1080/13623699.2014.930238
- Verbruggen, M. (2018, November). *Breaking out of the silos: The need for a whole-of-disarmament approach to Arms Control of AI*. Conference talk presented at the Beyond Killer Robots: Networked Artificial Intelligence Disrupting the Battlefield, Copenhagen, Denmark.
- West, D. M. (2018, August 29). Brookings survey finds divided views on artificial intelligence for warfare, but support rises if adversaries are developing it. Retrieved from <https://www.brookings.edu/blog/techtank/2018/08/29/brookings-survey-finds-divided-views-on-artificial-intelligence-for-warfare-but-support-rises-if-adversaries-are-developing-it/>
- Williamson, R. (2003). Hard law, soft law, and non-law in multilateral arms control: Some compliance hypotheses. *Chicago Journal of International Law*, 4(1). Retrieved from <https://chicagounbound.uchicago.edu/cjil/vol4/iss1/7>
- Work, B. (2015, December 14). Deputy secretary of defense bob work's speech at the CNAS defense forum. Retrieved from <http://www.defense.gov/News/Speeches/Speech-View/Article/634214/cnas-defense-forum>
- Yusuf, M. (2009). *Predicting proliferation: The history of the future of nuclear weapons*. Washington, DC: Brookings Institute.
- Zhang, B., & Dafoe, A. (2019). *Artificial intelligence: American attitudes and trends*. Oxford: Center for the Governance of AI, Future of Humanity Institute,

University of Oxford. Retrieved from <https://governanceai.github.io/US-Public-Opinion-Report-Jan-2019/>

Zwetsloot, R., Toner, H., & Ding, J. (2018, November 16). Beyond the AI Arms Race: America, China, and the dangers of zero-sum thinking. *Foreign Affairs*. Retrieved from <https://www.foreignaffairs.com/reviews/review-essay/2018-11-16/beyond-ai-arms-race>