BRILL
NIJHOFF

JOURNAL OF INTERNATIONAL HUMANITARIAN LEGAL STUDIES 10 (2019) 129-157

INTERNATIONAL HUMANITARIAN LEGAL STUDIES
brill.com/ihls

Innovation-Proof Global Governance for Military Artificial Intelligence?

How I Learned to Stop Worrying, and Love the Bot

Matthijs M Maas
PhD Fellow, Centre for International Law, Conflict and Crisis, Faculty of Law, University of Copenhagen, Denmark; Research Associate, Center for the Governance of AI, Future of Humanity Institute, University of Oxford, UK
matthijs.maas@jur.ku.dk

Abstract

Amidst fears over artificial intelligence 'arms races', much of the international debate on governing military uses of AI is still focused on preventing the use of lethal autonomous weapons systems (LAWS). Yet 'killer robots' hardly exhaust the potentially problematic capabilities that innovation in military AI (MAI) is set to unlock. Governance initiatives narrowly focused on preserving 'meaningful human control' over LAWS therefore risk being bypassed by the technological state-of-the-art. This paper departs from the question: how can we formulate 'innovation-proof governance' approaches that are resilient or adaptive to future developments in military AI? I develop a typology for the ways in which MAI innovation can disrupt existing international legal frameworks. This includes 'direct' disruption – as new types of MAI capabilities elude categorization under existing regimes – as well as 'indirect' disruption, where new capabilities shift the risk landscape of military AI, or change the incentives or values of the states developing them. After discussing two potential objections to 'innovation-proof governance', I explore the advantages and shortcomings of three possible approaches to innovation-proof governance for military AI. While no definitive blueprint is offered, I suggest key considerations for governance strategies that seek to ensure that military AI remains lawful, ethical, stabilizing, and safe.

Keywords

artificial intelligence – AI – military – innovation – technology governance – preventative security governance – international humanitarian law – Collingridge Dilemma

| DOI:10.1163/18781527-01001006

Technology,
I used to hope
For a break-through;
Now I wonder
What into?[1]

1 Introduction

The above poem adequately expresses a sentiment common perhaps amongst international lawyers. In recent years, new developments in military technologies have repeatedly challenged prevailing norms of international law, and extant governance instruments with them.[2] Perhaps none of these innovations has seized the attention as much as has the development of 'lethal autonomous weapons systems' (LAWS) – weapons platforms capable of selecting and engaging targets on their own, potentially with lethal force.[3]

Amidst broader fears of an emerging 'arms race' in artificial intelligence (AI),[4] concerns over the ethical and legal ramifications of machines autonomously taking kill decisions have fuelled widespread public campaigns against the use of such so-called 'killer robots'.[5] Consequently, 26 countries have now called for an explicit ban that requires some form of human control in the use of force,[6] and the UN Convention on Certain Conventional Weapons (CCW) Group of Governmental Experts (GGE) has been in ongoing discussion on

[1] W Berry, Address at the US Library of Congress, 1979, quoted in Graeme Laurie, Shawn HE Harmon and Fabiana Arzuaga, 'Foresighting Futures: Law, New Technologies, and the Challenges of Regulating for Uncertainty' (2012) 4 Law, Innovation & Technology 1.

[2] Denise Garcia, 'Future Arms, Technologies, and International Law: Preventive Security Governance' (2016) 1 European Journal of International Security 94.

[3] UNHRC, 'Report of the Special Rapporteur on extrajudicial, summary or arbitrary executions' (2013) UN Doc A/HRC/23/47.

[4] Edward Moore Geist, 'It's Already Too Late to Stop the AI Arms Race – We Must Manage It Instead' (2016) 72 Bulletin of the Atomic Scientists 318; Nicholas Thompson and Ian Bremmer, 'The AI Cold War That Threatens Us All' *Wired* (23 October 2018) <wired.com/story/ai-cold-war-china-could-doom-us-all> accessed 20 November 2018; cf Stephen Cave and Seán S Ó hÉigeartaigh, 'An AI Race for Strategic Advantage: Rhetoric and Risks', AAAI / ACM Conference on Artificial Intelligence, Ethics and Society (2018).

[5] See for instance Human Rights Watch, *Losing Humanity: The Case against Killer Robots* (Human Rights Watch 2012) <www.hrw.org/sites/default/files/reports/arms1112_ForUpload.pdf> accessed 14 January 2019.

[6] Adam Satariano, 'Will There Be a Ban on Killer Robots?' *The New York Times* (20 October 2018).

such a ban.[7] Even amongst private sector technology developers, there has been a pronounced backlash against the use of their algorithms in military applications.[8]

Yet even as governance campaigns aimed at shackling 'killer robots' are picking up steam, their focus remains on ensuring 'meaningful human control' over 'autonomous' military robots.[9] Yet as recently argued, these existing governance debates may no longer adequately reflect how technological innovation in military AI has continued even within the past few years; as such, this literature risks 'fighting the last war' – embedding a narrow view of both the problems (what military AI systems need regulation), and the optimal or adequate governance solutions to be pursued.[10] Beyond a specific problem for the UN CCW process, this instance hints at a more general pattern of 'governance-disrupting innovation', whereby technological development may continuously or iteratively challenge or bypass existing governance approaches, leaving in their wake 'jurisprudential space junk': collections of fragmented, hard-to-amend treaty regimes which 'are theoretically in force but actually simply clutter and confuse the relevant legal regime'.[11] Given that AI is a flexible, 'generally enabling' technology,[12] technological innovation and diversification in (military) AI systems is likely to continue (indeed, may even be diversified by global bans that close off certain more obvious avenues). Given this, are we always condemned to repeat the cycle? Will technology inevitably drive a

[7] See also Group of Governmental Experts of the High Contracting Parties to the CCW 'Report of the 2017 Group of Governmental Experts on Lethal Autonomous Weapons Systems (LAWS)' (2017) UN Doc CCW/GGE.1/2017/3; Group of Governmental Experts of the High Contracting Parties to the CCW 'Report of the 2018 Session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems' (2018) UN Doc CCW/GGE.1/2018/3.

[8] See 'Letter to Google C.E.O.' <static01.nyt.com/files/2018/technology/googleletter.pdf> accessed 9 April 2018.

[9] UNODA 'Perspectives on Lethal Autonomous Weapon Systems' (2017) UNODA Occasional Papers, No 3; Human Rights Watch, 'Killer Robots and the Concept of Meaningful Human Control' (*Human Rights Watch*, 11 April 2016) <www.hrw.org/news/2016/04/11/killer-robots-and-concept-meaningful-human-control> accessed 30 November 2018.

[10] cf Léonard van Rompaey, 'Shifting from Autonomous Weapons to Military Networks' (2019) 10 Journal of International Humanitarian Legal Studies 111; Hin-Yan Liu, 'From the Autonomy Framework towards Networks and Systems Approaches for "Autonomous" Weapons Systems' (2019) 10 Journal of International Humanitarian Legal Studies 89.

[11] Rebecca Crootof, 'Jurisprudential Space Junk: Treaties and New Technologies' in Chiara Giorgetti and Natalie Klein (eds), *Resolving Conflicts in the Law* (2019) 107.

[12] Michael C Horowitz, 'Artificial Intelligence, International Competition, and the Balance of Power' (2018) 1(3) Texas National Security Review 36.

sort of 'legal entropy' towards ever-increasing jurisprudential space junk? Or might we instead learn from the present experience with LAWS, and design governance approaches that can leapfrog or anticipate future innovations in military AI, or are resilient or adaptive it?

This article accordingly departs from the question of how we might formulate 'innovation-proof governance' approaches which are resilient or adaptive to future innovation in military AI. It will seek to provide an exploration of the opportunities and pitfalls of various strategies for achieving such governance for military AI (MAI); it suggests that such governance may be possible, but must involve notable changes from how international technology governance proceeds today. This argument proceeds as follows: I first discuss how ongoing innovation in military AI gives rise to a particularly severe case of the 'Collingridge Dilemma', which will challenge the efficacy of extant governance approaches. Secondly, referring to concrete potential disruptive innovations in military AI systems, I develop a typology for understanding the distinct ways by which MAI innovation may challenge or disrupt existing international legal frameworks. I distinguish between 'direct' governance disruption, whereby new types of MAI systems or capabilities elude inclusion under existing regimes, and 'indirect' governance disruption, whereby new MAI systems shift the technology's risk landscape ('problem portfolio'), or change the incentives or values of states. Thirdly, I discuss and rebut some potential objections to the idea of undertaking or even pursuing 'innovation-proof governance', and then sketch out the advantages and shortcomings of three possible approaches to achieving it in the context of military AI. Finally, I conclude with takeaways and directions.

2 Background: Military AI and the Collingridge Dilemma: Towards Governance-Disrupting Innovation?

2.1 *Whence Military AI?*

What is the scope of innovation of AI in a military context, and why is it likely to be particularly disruptive? AI is a general-purpose, 'enabling' technology;[13] with the aim of supporting, substituting for, and improving over (in terms of accuracy, speed, and/or scale) human performance in tasks such as 'pattern recognition', 'prediction', or 'decision-making'. While these tasks are individually quite bounded and narrow, the sheer domain-generality of such tasks – the range of contexts in which, say, being able to recognize patterns comes in useful – means that AI can be integrated in a wide range of military functions,

[13] Ibid.

and embedded in and distributed across a range of platforms and cloud systems. Such networked systems approaches what the US Army has called the 'Internet of Battle Things',[14] and what the Chinese PLA has referred to as the 'intelligentization' of war.[15]

The space of potential military uses of AI therefore extends far beyond just 'killer robots' alone. It includes a wide range of systems and applications, both those deployed in a direct offensive capability, whether kinetic (such as lethal autonomous weapons systems) or non-kinetic but nonetheless adversarial, such as autonomous cyberwarfare systems or adaptive radar-jamming or electronic warfare capabilities.[16] This also includes applications of AI in non-kinetic and supportive military roles, such as logistics, medevac, or tactical surveillance.[17] Finally, it includes improved (satellite or anti-submarine) sensing, intelligence, or war-planning capabilities which may directly affect the (nuclear) strategic balance.[18]

It is important to note that not all uses of AI in war will be challenging for international (legal or public) norms: compare the use of AI for improving

[14] Alexander Kott, 'Challenges and Characteristics of Intelligent Autonomy for Internet of Battle Things in Highly Adversarial Environments' <arxiv.org/ftp/arxiv/papers/1803/1803.11256.pdf> accessed 14 November 2018.

[15] cf Elsa B Kania, 'Battlefield Singularity: Artificial Intelligence, Military Revolution, and China's Future Military Power' (Center for a New American Security 2017) <s3.amazonaws.com/files.cnas.org/documents/Battlefield-Singularity-November-2017.pdf?mtime=20171129235804> accessed 28 March 2018.

[16] See, eg, Paul Tilghman, 'Adaptive Radar Countermeasures (ARC)' (*DARPA*) <www.darpa.mil/program/adaptive-radar-countermeasures> accessed 12 March 2018; Paul Tilghman, 'Behavioral Learning for Adaptive Electronic Warfare (BLADE)' (*DARPA*) <www.darpa.mil/program/behavioral-learning-for-adaptive-electronic-warfare> accessed 12 March 2018.

[17] For a survey, cf Matthijs M Maas, Tim Sweijs and Stephan De Spiegeleire, *Artificial Intelligence and the Future of Defense: Strategic Implications for Small- and Medium-Sized Force Providers* (The Hague Centre for Strategic Studies 2017) <hcss.nl/report/artificial-intelligence-and-future-defense> accessed 19 May 2017.

[18] Edward Geist and Andrew J Lohn, 'How Might Artificial Intelligence Affect the Risk of Nuclear War?' (RAND 2018) <www.rand.org/pubs/perspectives/PE296.html> accessed 22 September 2018; Keir A Lieber and Daryl G Press, 'The New Era of Counterforce: Technological Change and the Future of Nuclear Deterrence' (2017) 41 International Security 9; Kenneth Payne, 'Artificial Intelligence: A Revolution in Strategic Affairs?' (2018) 60 Survival 7; Christopher A Bidwell and Bruce W MacDonald, 'Emerging Disruptive Technologies and Their Potential Threat to Strategic Stability and National Security' (Federation of American Scientists 2018) <fas.org/wp-content/uploads/media/FAS-Emerging-Technologies-Report.pdf> accessed 5 November 2018.

logistics or medevac.[19] However, some will be. Moreover, further innovation in military AI is likely to be disruptive both strategically, conceptually – and (ergo) legally. Developments in predictive analytics, 'war-algorithms',[20] 'digitally-enabled' warfare,[21] and the increasing integration of AI systems in networks with many military and civilian nodes, and with diverse sensors and effectors, introduces novel sets of ethical, legal, strategic or safety concerns.[22] The challenges created by networked AI systems currently remain relatively under-recognized and under-addressed within the regulatory paradigm focused on 'killer robots' or 'autonomous weapons systems' alone,[23] although the most recent iteration of the CCW process has, to its credit, focused more awareness of the problems of the 'characterization' of autonomous weapons systems (AWS).[24] Yet, even if these debates are now gradually coming to terms with the challenges of networked AI, can they keep pace with the next wave(s) of 'governance-disruptive innovation' in military AI?

What might such innovation look like? This paper does not seek to offer definitive predictions about what the next breakthroughs in AI will be.[25] Indeed,

[19] Kenneth H Wong, 'Framework for Guiding Artificial Intelligence Research in Combat Casualty Care', *Medical Imaging 2019: Imaging Informatics for Healthcare, Research, and Applications* (International Society for Optics and Photonics 2019) <www.spiedigitallibrary.org/conference-proceedings-of-spie/10954/109540Q/Framework-for-guiding-artificial-intelligence-research-in-combat-casualty-care/10.1117/12.2512686.short> accessed 22 March 2019.

[20] Dustin A Lewis, Gabriella Blum and Naz K Modirzadeh, 'War-Algorithm Accountability' (Harvard Law School Program on International Law and Armed Conflict, 31 August 2016) <blogs.harvard.edu/pilac/files/2016/09/War-Algorithm-Accountability-Without-Appendices-August-2016.pdf> accessed 13 December 2018.

[21] Jacquelyn Schneider, 'Digitally-Enabled Warfare: The Capability-Vulnerability Paradox' (Center for a New American Security 2016) <www.cnas.org/publications/reports/digitally-enabled-warfare-the-capability-vulnerability-paradox> accessed 12 January 2019.

[22] Matthijs M Maas, 'How Viable Is International Arms Control for Military Artificial Intelligence? Three Lessons from Nuclear Weapons' (2019) 40 Contemporary Security Policy 286–287.

[23] See also Léonard Van Rompaey, 'Distributed and Networked Autonomy: Visualising the Legal Problems Caused by Military Networks' (presentation, Beyond Killer Robots: Networked Artificial Intelligence Disrupting the Battlefield, Copenhagen, 15 November 2018); Hin-Yan Liu, 'From the Autonomy Framework Towards Networks and Systems Approaches' (n 10).

[24] For instance, see Group of Governmental Experts of the High Contracting Parties to the CCW (n 7), Annex III para 1.

[25] Indeed, the question of how to improve our ability to forecast capability progress in AI remains an underdeveloped yet critical area of research: Allan Dafoe, 'AI Governance:

if predicting these accurately and exhaustively were so straightforward that a single scholar could do it, this would suggest that technological innovation would be not at all surprising or disruptive to governance.[26] However, given the deep and extensive historical links between the development of many distinct technologies (like the internet) with defence purposes,[27] we can expect the future path of AI to be similarly shaped by its military roots.[28] Ultimately, military or strategic applications seem set to remain a key driver of AI research; and one does not need to presume potential breakthroughs in far-future capabilities, to already consider a range of capabilities in development or in reach within the near-term, which by themselves or in combination would have far-reaching tactical, strategic and political impacts.[29] We will discuss some of these prospective innovations in later sections.

2.2 *… and Whence Governance?*

More fundamentally, one can ask, why should we expect continued innovation to prove a structural problem for governance? In a domestic context, scholars have argued that while new technologies can certainly pose a 'recurring dilemma' to existing laws,[30] eventual legal development can usually come to terms with them. Yet, in a global context, governing the continuing development in MAI may be particularly challenging, because it meets many features

A Research Agenda' (Governance of AI Program, Future of Humanity Institute 2018) <www.fhi.ox.ac.uk/govaiagenda> accessed 20 October 2018.

[26] Though for work on conditions under which forecasts of military technology can achieve reasonably accuracy, cf Alexander Kott and Philip Perconti, 'Long-Term Forecasts of Military Technologies for a 20–30 Year Horizon: An Empirical Assessment of Accuracy' [2018] arXiv:1807.08339 [cs] <arxiv.org/abs/1807.08339> accessed 18 December 2018.

[27] cf Sharon Weinberger, *The Imagineers of War: The Untold Story of DARPA, the Pentagon Agency That Changed the World* (Random House 2017); Neil deGrasse Tyson and Avis Lang, *Accessory to War: The Unspoken Alliance Between Astrophysics and the Military* (W W Norton & Company 2018).

[28] The author thanks one anonymous reviewer for this point.

[29] cf Kareem Ayoub and Kenneth Payne, 'Strategy in the Age of Artificial Intelligence' (2016) 39 Journal of Strategic Studies 793. See also some of the discussion under 'international security' in Dafoe (n 25); as well as Edward Parson and others, 'Artificial Intelligence in Strategic Context: An Introduction' (*AI Pulse*, 8 February 2019) <aipulse.org/artificial-intelligence-in-strategic-context-an-introduction> accessed 26 February 2019.

[30] Lyria Bennett Moses, 'Recurring Dilemmas: The Law's Race to Keep Up With Technological Change' (Social Science Research Network 2007) SSRN Scholarly Paper ID 979861 <www.austlii.edu.au/au/journals/UNSWLRS/2007/21.html> accessed 3 July 2018; see also David D Friedman, 'Does Technology Require New Law?' (2001) 71 Public Policy 16.

of the Collingride Dilemma.[31] As stated by David Collingridge: 'When change is easy, the need for it cannot be foreseen; when the need for change is apparent, change has become expensive, difficult, and time-consuming.'[32] This is because of a bifurcated problem in the context of regulating a technology; early on in its development, we face *an information problem*: the technology's critical features, uses and impacts cannot be easily predicted (or unanimously agreed upon) until it is more extensively developed and used. Yet once the technology has been more fully developed and deployed, we face *a power problem*: control or broad governance has become difficult because the technology has already become embedded in path-dependent ways; because established (unequal) stakes or interests or shifts in power have become clear and entrenched; or because extant governance approaches have begun to converge on certain path-dependent 'regulation niches' which lock in certain foci, framings, and (suboptimal or increasingly inadequate) solution portfolios.

Of course, these challenges are certainly not unique to technology governance at the global level; nor are they unique to the development of AI in the military sphere alone. Indeed, as will be discussed later on, many of the core shifts and disruptions in power which AI enables will overlap and blur the easy distinction between the military and civilian sectors.[33] Nonetheless, the information and power problems inherent in the Collingridge Dilemma may be particularly severe for military AI technologies. Moreover, in responding to extant debates focused narrowly on LAWS, it is military AI systems that can serve as the 'hard case' for establishing innovation-proof global governance approaches going forward. We therefore return to the motivating question: how, if at all, might we formulate 'innovation-proof governance' approaches that are resilient or adaptive to future innovation in military AI? What might those look like?

3 How Might MAI Innovation Disrupt Governance?

As discussed in the work of Lyria Bennett Moses, new technological innovation can challenge existing legal structures directly – by creating new entities,

[31] For a similar argument on AI, see Michael Guihot, Anne F Matthew and Nicolas Suzor, 'Nudging Robots: Innovative Solutions to Regulate Artificial Intelligence' (2017) 20 Vanderbilt Journal of Entertainment & Technology Law 385.

[32] David Collingridge, *The Social Control of Technology* (Palgrave Macmillan 1981).

[33] cf Rain Liivoja, Kobi Leins and Tim McCormack, 'Emerging Technologies of Warfare' in Rain Liivoja and Tim McCormack (eds), *Routledge Handbook of the Law of Armed Conflict* (Routledge 2016); William H Boothby (ed), *New Technologies and the Law in War and Peace* (Cambridge University Press 2018); Shoshana Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (PublicAffairs 2019).

or by enabling new behaviour.[34] In addition, innovation can also challenge governance indirectly, by changing the risk landscape or 'problem portfolio' of the technology; or by changing the incentives or values of states parties.[35] We will discuss these in turn.

3.1 *Direct Disruption – New Entities; New Capabilities*

Innovation can pose a direct challenge to existing governance instruments. In the most straightforward way, it does so by creating new 'entities' (or objects) not explicitly covered by existing treaties, leading to a need to extend existing regimes (compare the CCW Additional Protocol on Blinding Laser Weapons), or a need for the creation of entirely new regimes.

More problematically, innovation may enable new *behaviour* that renders it more difficult to apply established principles – compare the difficulty of applying international humanitarian law (IHL) principles in cyberspace, where attribution may be difficult[36] – or which enables parties to respect the letter of existing (bilateral) agreements, while violating their spirit. One example of the latter, in a different context, could be found in the recently concluded US program to refit the W76 nuclear warheads on their missile submarine force with 'superfuzes' which prevent 'overshoot', and enable a more reliable targeting of hardened installations such as (Russian) missile silos. This program, presented as a modernization measure to ensure continued reliability, has in fact effectively tripled the effective counter-force lethality of these nuclear missiles.[37] This case illustrates how even seemingly 'incremental' technological improvements can qualitatively shift military capabilities in destabilizing ways, even as the principal states formally comply with treaty caps on deployed numbers of weapons or launchers.

In an AI context, innovations that could similarly allow technology to 'route around' existing treaties might include novel human-machine integration dynamics (e.g. crowd-sourced labelling of training data; new ways to integrate

[34] cf Moses (n 30); for her broader analysis of a 'Theory of Law and Technological Change', see Lyria Bennett Moses, 'Why Have a Theory of Law and Technological Change?' (2007) 8 Minnesota Journal of Law, Science & Technology 19.

[35] This typology is loosely inspired by Colin B Picker, 'A View from 40,000 Feet: International Law and the Invisible Hand of Technology' (2001) 23 Cardozo Law Review 151.

[36] Michael J Glennon, 'The Dark Future of International Cybersecurity Regulation' (2013) 6 Journal of National Security Law & Policy 563; Garcia (n 2) 95.

[37] Hans Kirstensen, Matthew McKinzie and Theodore A Postol, 'How US Nuclear Force Modernization Is Undermining Strategic Stability: The Burst-Height Compensating Super-Fuze' (*Bulletin of the Atomic Scientists*, 1 March 2017) <thebulletin.org/how-us-nuclear-force-modernization-undermining-strategic-stability-burst-height-compensating-super10578> accessed 12 March 2018.

human operators with drone swarms,[38] or soldier-platform brain-computer interfaces),[39] which could formally re-insert a human 'in-the-loop', but in altered or distributed modes of cognitions that render this arrangement less protective or meaningful. More generally, the integration of AI functionalities in existing weapons systems[40] might enable far-reaching military capability gains in many other fields, even as IHL principles are formally respected.

In sum, these and other innovations in military AI would disrupt governance by leading to conceptual confusion or apparent difficulty of categorization. For instance, Article 36 of Additional Protocol I of the Geneva Convention mandates states' conduct reviews of any 'new weapon, means or method of warfare'.[41] Liu has previously argued that AWS straddle the existing legal categories of 'weapons' and 'combatants', with many of these systems often not inflicting violence in a direct manner, but rather serving as intermediary platforms.[42] MAI innovation – including the progressive integration of smart modules into pre-existing, legal weapons systems, and/or the integration of existing legal weapons systems into broader information networks that include AI components at the tactical or strategic levels – may also exacerbate the problem of categorizing them as discrete weapons systems, rendering yet more difficult the question as to whether (or to what degree) these individual, integrated or aggregate MAI systems can easily be classified as a 'new weapon, means or method of warfare' within the remit of Article 36.

On the other hand, it might be argued that in principle, the emergence of new technological entities or capabilities need not pose a *structural* problem to governance, especially in areas where international law is to some degree 'technology-neutral'. As Garcia has noted, it might be argued that 'extant IHL …

[38] Irving Lachow, 'The Upside and Downside of Swarming Drones' (2017) 73 Bulletin of the Atomic Scientists 96.

[39] Michael Joseph Gross, 'The Pentagon's Push to Program Soldiers' Brains' *The Atlantic* (November 2018) <www.theatlantic.com/magazine/archive/2018/11/the-pentagon-wants-to-weaponize-the-brain-what-could-go-wrong/570841> accessed 18 December 2018; see also Brad Allenby, 'Designer Warriors: Altering Conflict – and Humanity Itself?' (2018) 74 Bulletin of the Atomic Scientists 379.

[40] cf also Maaike Verbruggen, 'Breaking out of the Silos: The Need for a Whole-of-Disarmament Approach to Arms Control of AI' (presentation, Beyond Killer Robots: Networked Artificial Intelligence Disrupting the Battlefield, Copenhagen, Denmark, November 2018).

[41] Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts (adopted 8 June 1977, entered into force 7 December 1978) 1125 UNTS 3 (Additional Protocol I) art 36.

[42] cf Hin-Yan Liu, 'Categorization and Legality of Autonomous and Remote Weapons Systems' (2012) 94 International Review of the Red Cross 627, 628.

applies to new technology because it is not the nature of the means and methods that triggers IHL, but the context and the humanitarian consequences'.[43] Others likewise contend that IHL is in principle applicable to all new military technologies,[44] such that conceptual difficulties in categorizing the technology's 'true nature' should not arise or be legally relevant.[45] Yet as Garcia also admits, there are some problems with this approach, the first pragmatic problem being that despite being required under international law, in practice only some states regularly carry out weapon reviews, and particularly drone programs have been surrounded by secrecy rather than transparency.[46] Moreover, the IHL framework is applicable only in situations of armed conflict, missing the fact that many of these technologies may be used during nominal 'peace time' or outside of warzones (e.g. in cyberspace) – or that, indeed, these capabilities may shift or blur both the distinction between war and peace that is pivotal to IHL, as well as the 'dual-use' distinction pivotal to many arms control regimes.[47] Thirdly, and related to this, is the problem that relying on IHL as the 'hammer' by which to constrain military AI may not be appropriate or sufficient in governing a wider portfolio of dangers or problems introduced by MAI innovation, many of which are not 'nails' under the traditional conception of warfare.

3.2 *Indirect Disruption: MAI Innovation Could Alter the 'Problem Portfolio'*

This suggests an additional, indirect way in which ongoing MAI innovation may challenge the efficacy or sufficiency of existing governance approaches; this is by shifting the prevailing 'distribution of dangers' arising from the technology. The current concerns about military usage of AI, in public debates and in the CCW process, have mainly (and for good reason) focused on *ethics* and *legality,* respectively. Ethical concerns revolve around the moral horror of

[43] Garcia (n 2) 105.

[44] Justin McClelland, 'The Review of Weapons in Accordance with Article 36 of Additional Protocol I' (2003) 85 International Review of the Red Cross 397. This requirement was also recently reiterated in the guiding principles (d) issued by the CCW GGE. Group of Governmental Experts of the High Contracting Parties to the CCW (n 7) 4.

[45] For a more general examination that seeks to distill the 'essential features' of robotics, see Ryan Calo, 'Robotics and the Lessons of Cyberlaw' (2015) 103 California Law Review 513; and for a critique, see Jack B Balkin, 'The Path of Robotics Law' (2015) 6 California Law Review Circuit 17.

[46] Garcia (n 2) 105.

[47] cf Tara Mahfoud and others, 'The Limits of Dual Use' (*Issues in Science and Technology*, 31 July 2018) <issues.org/the-limits-of-dual-use> accessed 8 January 2019.

'machine killing' and its effects on human dignity; legal concerns pertain to the effects of LAWS on IHL principles or human rights.

However, there is little reason to expect that the 'risk portfolio' of military AI will remain stable under continued or iterative innovation – that 'machine killing' or 'proportionality under the laws of war' will remain the sole, main, or even dominant governance problems created by this technology. By comparison, this past decade the debate around the use of (piloted) drones was hardly limited to their ethical or legal implications for the just war tradition, but also came to encompass concerns that they might for instance lower the threshold to war.[48]

Similarly, many argue that as MAI continues to develop, additional risks will manifest in the domains of *strategic stability* or *safety*. In terms of strategic stability, AI might destabilize conflicts, for instance, by putting a tactical premium on offense; by increasing the opacity of the new relative balance between force capabilities;[49] or through new sensing and anti-submarine warfare (ASW) capabilities that erode nuclear deterrence stability. In terms of safety, Scharre and Borrie have previously suggested that LAWS may be susceptible to emergent 'normal accidents';[50] more recent work has suggested such vulnerability may be an intrinsic feature of not just LAWS, but nearly all AI systems operating in highly competitive (e.g. military) environments.[51]

[48] Daniel Brunstetter and Megan Braun, 'The Implications of Drones on the Just War Tradition' (2011) 25 Ethics & International Affairs 337; on the 'lowered threshold to war' argument, see also: James Igoe Walsh and Marcus Schulzke, 'The Ethics of Drone Strikes: Does Reducing the Cost of Conflict Encourage War?' (US Army War College & Strategic Studies Institute 2015) <ssi.armywarcollege.edu/pdffiles/PUB1289.pdf>; as well as Michael C Horowitz, Sarah E Kreps and Matthew Fuhrmann, 'The Consequences of Drone Proliferation: Separating Fact from Fiction' (2016) 41 International Security 7.

[49] cf Matthew Kroenig and Bharath Gopalaswamy, 'Will Disruptive Technology Cause Nuclear War?' (*Bulletin of the Atomic Scientists*, 12 November 2018) <thebulletin.org/2018/11/will-disruptive-technology-cause-nuclear-war> accessed 22 November 2018.

[50] Paul Scharre, 'Autonomous Weapons and Operational Risk' (Center for a New American Security 2016) <s3.amazonaws.com/files.cnas.org/documents/CNAS_Autonomous-weapons-operational-risk.pdf> accessed 24 January 2018; John Borrie, 'Safety, Unintentional Risk and Accidents in the Weaponization of Increasingly Autonomous Technologies' (UNIDIR 2016) UNIDIR Resources 5 <www.unidir.org/files/publications/pdfs/safety-unintentional-risk-and-accidents-en-668.pdf> accessed 7 March 2018.

[51] Matthijs M Maas, 'Regulating for "Normal AI Accidents" – Operational Lessons for the Responsible Governance of AI Deployment' *Proceedings of the 2018 AAAI / ACM Conference on Artificial Intelligence, Ethics and Society* (Association for the Advancement of Artificial Intelligence 2018) <www.aies-conference.com/wp-content/papers/main/AIES_2018_paper_118.pdf> accessed 22 February 2018; cf also Richard Danzig, 'Technology Roulette:

The resulting potential for emergent accident cascades may create risks of emergent, inadvertent 'flash wars' between interacting AI systems.[52] Worse, as the record of normal accidents in other technologies shows, this accident risk is often exacerbated rather than addressed by fail-safes or nominal human operator involvement,[53] rendering 'meaningful human control' illusory or at least very problematic.

It might be countered that, even if these problems will emerge, this is surely not a problem for IHL, which has a clear focus and jurisdiction, and was never meant to concern itself much or at all with questions of strategic stability, which instead are in the purview of arms control regimes. That is fair, and indeed the point is less that one should try to expand or stretch IHL to cover these new species of risk – which might stretch the regime to its breaking point. Instead, it is a reminder that the ongoing evolution in the MAI risk portfolio, as a result of innovation, will erode the efficacy or sufficiency of governance foci (e.g. 'meaningful human control') that target only a subset of these four sources of danger.

3.3 *Indirect Disruption: MAI Innovation Could Shift State Incentives*

Moreover, further innovation in AI may come to enable new MAI capabilities that shift the incentives of states parties, in ways that may come to put considerable pressure on extant agreements or regimes, or even undercut their legitimacy.

3.3.1 Increasing Military Appeal

In the first place, innovation may come to radically increase the perceived (absolute or relative) military *benefits* that a state perceives it forgoes by full compliance with bans. For instance, it might be argued that the introduction of unmanned but tele-operated drones offered large tactical advantages over manned aircraft – reducing or eliminating casualties amongst human pilots, and enabling militaries to have weapon platforms maintain a constant

Managing Loss of Control as Many Militaries Pursue Technological Superiority' (Center for a New American Security 2018) <s3.amazonaws.com/files.cnas.org/documents/CNAS Report-Technology-Roulette-DoSproof2v2.pdf?mtime=20180628072101> accessed 15 July 2018.

[52] Paul Scharre, 'Flash War: Autonomous Weapons and Strategic Stability' (Understanding Different Types of Risk, Geneva, 11 April 2016) <www.unidir.ch/files/conferences/pdfs/-en-1-1113.pdf> accessed 17 November 2017.

[53] Charles Perrow, *Normal Accidents: Living with High Risk Technologies* (Princeton University Press 1984); Scott D Sagan, *The Limits of Safety: Organizations, Accidents, and Nuclear Weapons* (Princeton University Press 1993).

loitering presence over an operational area. By comparison, fully autonomous LAWS, as conceived today, might offer militaries only relatively modest *additional* tactical benefits over piloted drones – benefits mostly in terms of savings in labour time and costs; in operational resilience to electromagnetic warfare and scrambling; and in advantages in reaction speed which may be critical in certain specific but very circumscribed contexts (such as point-defence anti-missile cannons). Of course, to a military force such benefits are certainly not to be scoffed at, but they might not be so large that fully autonomous LAWS become (perceived as) an irrevocable and irreplaceable military necessity – a *sine qua non* to prevail in any conflict.

However, the history of AI shows that, while progress can sometimes stall or plateau during extended 'AI winters', disjunctive advances in performance can be achieved suddenly. In non-adversarial contexts (e.g. face recognition, where the AIs do not directly 'fight their like'), such performance increases offer marginal but non-decisive improvements over existing systems. However, other capability gains offer qualitative advantages – compare the performance of DeepMind's AlphaZero, which in a series of 100 games utterly defeated Stockfish, the previous reigning computer chess world champion, by 28 wins, 72 draws, and not a single loss.[54] This may not even require deep breakthroughs; as OpenAI's GPT-2 text generation AI demonstrated, sometimes 'qualitative' breakthroughs in (near-)human equivalent performance can be achieved through 'grind' – the simple scaling up of existing (unsupervised) machine learning approaches to a sufficiently large dataset.[55] In a similar manner, continuing (fundamental, but also incremental) innovations in MAI systems are likely to yield qualitative, decisive advantages in key domains, for instance by enabling (cyberwarfare, pilot) capabilities that render all rival human or AI capabilities in that given domain obsolete. In these cases, as Payne has argued, 'marginal quality might prove totally decisive' because 'other things being equal, we can expect higher-quality AI to comprehensively defeat inferior

[54] David Silver and others, 'A General Reinforcement Learning Algorithm That Masters Chess, Shogi, and Go through Self-Play' (2018) 362 Science 1140; for commentary, see Steven Strogatz, 'One Giant Step for a Chess-Playing Machine' *The New York Times* (8 January 2019) <www.nytimes.com/2018/12/26/science/chess-artificial-intelligence.html> accessed 11 January 2019.

[55] Alec Radford and others, 'Language Models Are Unsupervised Multitask Learners' (14 February 2019) <d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf> accessed 17 February 2019; Alec Radford and others, 'Better Language Models and Their Implications' (*OpenAI Blog*, 14 February 2019) <blog.openai.com/better-language-models> accessed 17 February 2019.

rivals'.[56] Such gains are not easy to anticipate, but once conceivable would steeply increase the perceived (counterfactual) military costs states consider they incur by continued compliance with bans that they acceded to when the (prospective) military advantages appeared more modest.

3.3.2 Reducing Barriers to Access and Proliferation

Secondly, innovation might reduce *barriers* to access, facilitating noncompliance with governance regimes. This could happen because certain technical barriers are overcome, which enables wider access by many more parties to a certain minimum threshold level of useful AI capabilities. To give one example, recent years have seen the emergence of new and increasingly efficient paradigms in AI training approaches – such as the use of synthetic data; 'simulation transfer' learning;[57] 'meta-learning' and 'one-shot learning' from smaller data-sets of examples;[58] or new neural network designs that are increasingly able to work with messy, continuous, and irregularly measured datasets.[59] These powerful AI training capabilities could help ensure that achieving some level of 'functional' AI performance is increasingly restricted less by access to sufficient training data or even hardware, and more by (easily disseminable) software. This could increase proliferation risk,[60] and lower the use threshold

[56] Payne (n 18) 24.

[57] Compare the OpenAI 'Dactyl' robotic hand, which was trained entirely in simulation to solve real-world tasks, without physically accurate modelling of the world. OpenAI and others, 'Learning Dexterous In-Hand Manipulation' [2018] arXiv:1808.00177 [cs, stat] <arxiv.org/abs/1808.00177> accessed 11 January 2019. See also OpenAI. 'Learning Dexterity' (30 July 2018) <openai.com/blog/learning-dexterity> accessed 11 January 2019.

[58] Joaquin Vanschoren, 'Meta-Learning: A Survey' [2018] arXiv:1810.03548 [cs, stat] <arxiv.org/abs/1810.03548> accessed 11 January 2019; Natalie Ram, 'One Shot Learning In AI Innovation' (*AI Pulse*, 25 January 2019) <aipulse.org/one-shot-learning-in-ai-innovation> accessed 26 February 2019; Brenden M Lake and others, 'One Shot Learning of Simple Visual Concepts' *Proceedings of the Annual Meeting of the Cognitive Science Society* (2011) <cims.nyu.edu/~brenden/LakeEtAl2011CogSci.pdf> accessed 26 February 2019.

[59] cf Ricky TQ Chen and others, 'Neural Ordinary Differential Equations' [2018] arXiv:1806.07366 [cs, stat] <arxiv.org/abs/1806.07366> accessed 11 January 2019; for a discussion, see Karen Hao, 'A Radical New Neural Network Design Could Overcome Big Challenges in AI' [2018] *MIT Technology Review* <www.technologyreview.com/s/612561/a-radical-new-neural-network-design-could-overcome-big-challenges-in-ai> accessed 11 January 2019.

[60] cf Michael C Horowitz, 'The Algorithms of August' *Foreign Policy* (12 September 2018) <foreignpolicy.com/2018/09/12/will-the-united-states-lose-the-artificial-intelligence-arms-race> accessed 20 November 2018.

to non-state actors, potentially challenging the efficacy or sufficiency of state-based governance regimes.

3.3.3 Inhibiting Detection of Noncompliance, or Reducing Its Political Costs

Thirdly, along with increasing the prospective *gains* of development and removing barriers, MAI innovation could also reduce the anticipated effective *costs* of noncompliance with AI governance regimes. Whereas past arms control regimes were facilitated by the difficulty of hiding, say, (uranium) enrichment facilities, missile launch sites, or nuclear tests, it has been argued that the development of AI is, amongst other things, 'discreet', and 'discrete'.[61] Moreover, incidents such as the Volkswagen emissions scandal illustrate how certain algorithmic features or software capabilities may be easily disguised from inspectors. In this way, innovation towards increasingly complex AI systems, and the cross-integration of nominally civilian nodes or modules with military platforms, could inhibit effective, reliable or meaningful Article 36 Reviews. This would make it harder for others to detect or prove (in advance of armed conflict, or even during one) that a state is engaging in developing or deploying circumscribed MAI systems.

More fundamentally, MAI innovation could also reduce the reputational or political costs a state expects to incur for developing, deploying, or integrating into existing weapons new types of military AI, even if this is 'revealed' (indeed, states may no longer even feel the need to actively hide these capabilities). It is easy to see, for instance, why the public might rally against 'killer robots', which (appear to) have a clear threshold in use, have a visceral, violent impact – and which come with a long legacy of featuring in popular culture as terrifying antagonists to boot. By contrast, it may prove much harder to achieve a comparable degree of public opprobrium for, say, 'military cloud systems' that algorithmically process and mediate all military decision-making, but which are (or appear) two or more direct causal steps removed from the 'kill decisions'.

Critically, the public perceptions of military AI will matter not just from a political standpoint, but also a legal one. In recent years, both scholars and activists have suggested that the Martens Clause might be invoked to outlaw

[61] Matthew U Scherer, 'Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies' (2016) 29 Harvard Journal of Law & Technology 353. On the complexities of regulating AI broadly, see also Guihot, Matthew and Suzor (n 31).

LAWS.[62] The Martens Clause, as phrased in the 1977 Additional Protocol to the Geneva Conventions, states that:

> In cases not covered by this Protocol or by other international agreements, civilians and combatants remain under the protection and authority of the principles of international law derived from established custom, from the principles of humanity and from the dictates of public conscience.[63]

Some have suggested that 'the dictates of public conscience' might be established by reference to global public opinion on the use of such weapons.[64] Yet is there a coherent global public opinion on this matter? For instance, while a majority of respondents in a 2017 global IPSOS survey oppose deployment of LAWS (54% opposed, versus 24% in favour), respondents in China and India (representing a sizeable chunk of the world population) see majorities of public support (47% and 60% respectively).[65] Even in the West, opposition to LAWS is contextual, and can be both strengthened by popular culture exposure,[66] as well as weakened, when their use is framed as the protection of domestic forces.[67] That does not mean other concerns could not buttress or inform

[62] Human Rights Watch (n 5) 35–36; for a critical evaluation, see Tyler D Evans, 'At War with the Robots: Autonomous Weapon Systems and the Martens Clause' (2014) 41 Hofstra Law Review 39; Rob Sparrow, 'Ethics as a Source of Law: The Martens Clause and Autonomous Weapons' (*Humanitarian Law & Policy*, 14 November 2017) <blogs.icrc.org/law-and-policy/2017/11/14/ethics-source-law-martens-clause-autonomous-weapons> accessed 13 January 2019.

[63] Additional Protocol I art 1(2).

[64] Heather M Roff, 'What Do People Around the World Think About Killer Robots?' *Slate* (8 February 2017) <www.slate.com/articles/technology/future_tense/2017/02/what_do_people_around_the_world_think_about_killer_robots.html> accessed 5 April 2018; for a discussion of some of the problems with this, see Sparrow (n 64).

[65] Roff (n 64); IPSOS, 'Three in Ten Americans Support Using Autonomous Weapons' (*Ipsos*, 7 February 2017) <www.ipsos.com/en-us/news-polls/three-ten-americans-support-using-autonomous-weapons> accessed 13 January 2019; see also Open Roboethics Initiative, 'The Ethics and Governance of Lethal Autonomous Weapons Systems: An International Public Opinion Poll' (Open Roboethics Initiative 2015) <www.openroboethics.org/wp-content/uploads/2015/11/ORi_LAWS2015.pdf> accessed 5 April 2018.

[66] Kevin L Young and Charli Carpenter, 'Does Science Fiction Affect Political Fact? Yes and No: A Survey Experiment on "Killer Robots"' (2018) 62 International Studies Quarterly 562.

[67] Michael C Horowitz, 'Public Opinion and the Politics of the Killer Robots Debate' (2016) 3 Research & Politics 1; Darrell M West, 'Brookings Survey Finds Divided Views on Artificial Intelligence for Warfare, but Support Rises If Adversaries Are Developing It' (*Brookings*,

public opposition; for instance, a recent survey of the American public by the Center for the Governance of AI found slight support for the US investing more in AI military capabilities – although such support fell slightly as respondents were provided with information about the risks of a US-China AI arms race.[68]

Yet even if consistent public condemnation were to be achieved for LAWS, it seems unsure that such public condemnation (and thereby the 'dictates of public conscience' in the Martens Clause) would easily extend to future, non-kinetic MAI systems. For instance, advances in AI capabilities in modelling of societal dynamics, pattern detection, or prediction might yet enable a shift to more pre-emptive, yet less kinetic or visceral interventions without clear violence thresholds, in ways that make it harder to rally a comparative public political outcry. For instance, in the domestic context, scholars are already warning of the so-called 'Disneyfication' of smart cities to describe a type of technology-enabled embedded policing. Such preventative architecture of social control may feel less coercive – and spark less public condemnation – while achieving the same or superior effects for authorities.[69] Likewise, in the context of military uses of AI, a shift away from high-profile kinetic interventions towards increasingly 'invisible wars' may make some of the more problematic trends in the militarization of AI harder to oppose – even as IHL might record a formal success in containing the use of fully autonomous killer robots.

3.4 *Indirect Disruption: AI Innovation Could Change Values, Lead to Greater Unilateralism*

Finally, and more speculatively, general AI innovation – even beyond the battlefield – could contribute to a trend towards global unilateralism. In recent years, scholars have commented on an apparent period of tension for

29 August 2018) <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> accessed 21 September 2018.

[68] Baobao Zhang and Allan Dafoe, 'Artificial Intelligence: American Attitudes and Trends' (Center for the Governance of AI, Future of Humanity Institute, University of Oxford 2019) 26–28 <governanceai.github.io/US-Public-Opinion-Report-Jan-2019> accessed on 15 February 2019.

[69] Elizabeth E Joh, 'Policing the Smart City' (Social Science Research Network 2018) SSRN Scholarly Paper ID 3189089 <papers.ssrn.com/abstract=3189089> accessed 19 October 2018; for a discussion, see John Danaher, 'The Automation of Policing: Challenges and Opportunities' (*Philosophical Disquisitions*, 12 October 2018) <philosophicaldisquisitions.blogspot.com/2018/10/the-automation-of-policing-challenges.html> accessed 15 October 2018.

international law. Incidents of state backlash against international courts,[70] dismissive statements from world leaders, and (threatened or actual) withdrawals from treaties and multilateral arrangements[71] have raised concern over the sustained health of the multilateral rules-based order. Of course, this trend should not be over-exaggerated, yet its possible intersection with general advances in AI-enabled capabilities is nonetheless salient.

After all, while certain AI capabilities could be used to facilitate diplomatic processes,[72] strengthen the monitoring and enforcement of international law,[73] or support the pursuit of shared public good such as the Sustainable Development Goals,[74] disruptive innovations in AI capabilities both on and beyond the battlefield might also enable or drive a general shift to unilateralism, or shift the values of states in ways erosive to the norms of multilateralism, reciprocity, and legal recourse on which international law depends.[75]

It has been previously suggested that AI surveillance capabilities and LAWS might be attractive to authoritarian regimes, strengthening their ability to monitor their citizens and maintain centralized control over military force projection capability.[76] Others have suggested that AI could be a strong tool

[70] Karen J Alter, James Thuo Gathii and Laurence R Helfer, 'Backlash Against International Courts in West, East and Southern Africa: Causes and Consequences' (2016) 27 European Journal of International Law 293.

[71] James Crawford, 'The Current Political Discourse Concerning International Law' (2018) 81 Modern Law Review 1.

[72] Katharina E Höne, 'Mapping the Challenges and Opportunities of Artificial Intelligence for the Conduct of Diplomacy' (DiploFoundation 2019) <www.diplomacy.edu/sites/default/files/AI-diplo-report.pdf> accessed 14 March 2019.

[73] Berenice Boutin, 'Technologies for International Law & International Law for Technologies' (*International Law under Construction: Blog of the Groningen Journal of International Law*, 22 October 2018) <grojil.org/2018/10/22/technologies-for-international-law-international-law-for-technologies> accessed 31 October 2018. cf Danny Gold, 'Saving Lives with Tech Amid Syria's Endless Civil War' *Wired* (16 August 2018) <www.wired.com/story/syria-civil-war-hala-sentry> accessed 31 October 2018.

[74] António Guterres, 'UN Secretary-General's Strategy on New Technologies' (United Nations 2018) <www.un.org/en/newtechnologies/images/pdf/SGs-Strategy-on-New-Technologies.pdf> accessed 4 October 2018.

[75] For an informal discussion, cf 'Maas on AI and the Future of International Law' (*Algocracy and the Transhumanist Project*, 2 December 2018) <algocracy.wordpress.com/2018/12/02/episode-49-maas-on-ai-and-the-future-of-international-law> accessed 4 December 2018.

[76] Michael C Horowitz, 'Who'll Want Artificially Intelligent Weapons? ISIS, Democracies, or Autocracies?' (*Bulletin of the Atomic Scientists*, 29 July 2016) <thebulletin.org/who%E2%80%99ll-want-artificially-intelligent-weapons-isis-democracies-or-autocracies9692> accessed 13 May 2017.

for asymmetric (information) warfare against democracies,[77] facilitating types of informational attacks that are disproportionately effective against the distinct informational attack surfaces and threat models that plague democracies more than autocracies.[78]

In fact, Harari has argued that, since machine learning systems get more accurate and powerful the more data they have access to, AI might structurally ‘favor tyranny’ at a deep level, by making centralized information processing (and therefore centralized power) more efficient than the diffuse or decentralized information processing typical of democracies.[79] In fact, in a recent US expert survey, a majority of scholars ‘agreed’ or ‘strongly agreed’ with the premise that ‘Technological change today is strengthening authoritarianism relative to democracy’.[80] In this way, AI innovation could change the balance of power between actors, by empowering those already unsympathetic to international norms, enabling them to flout or even challenge the international legal order.[81]

Moreover, even for states which have previously been supportive of international norms which they held to be in their interest, broad AI innovation could be(come) seen as a tempting substitute to the assurances once offered by give-and-take compromise. That is, whichever concrete national interests or goals – domestic security, global stability, prosperity, domestic or global legitimacy, soft power – states previously believed they could only or best achieve through reciprocal engagement in or compliance with international norms, new AI capabilities may arise – from AI-enhanced surveillance to AI improvements in military force projection capabilities, commercial AI applications, computational propaganda at home and abroad, the use of AI to model and

[77] cf Alina Polyakova, ‘Weapons of the Weak: Russia and AI-Driven Asymmetric Warfare’ (*Brookings*, 15 November 2018) <www.brookings.edu/research/weapons-of-the-weak-russia-and-ai-driven-asymmetric-warfare> accessed 13 January 2019.

[78] Henry Farrell and Bruce Schneier, ‘Common-Knowledge Attacks on Democracy’ (Berkman Klein Center 2018) Research Publication 2018–7 <papers.ssrn.com/abstract=3273111> accessed 13 January 2019.

[79] Yuval Noah Harari, ‘Why Technology Favors Tyranny’ *The Atlantic* (October 2018) <www.theatlantic.com/magazine/archive/2018/10/yuval-noah-harari-technology-tyranny/568330> accessed 12 September 2018.

[80] Foreign Affairs, ‘Does Technology Favor Tyranny?’ (*Foreign Affairs*, 13 February 2019) <www.foreignaffairs.com/ask-the-experts/2019-02-12/does-technology-favor-tyranny> accessed 11 March 2019.

[81] Richard Danzig, ‘An Irresistible Force Meets a Moveable Object: The Technology Tsunami and the Liberal World Order’ (2017) 5 Lawfare Research Paper Series <assets.documentcloud.org/documents/3982439/Danzig-LRPS1.pdf> accessed 1 September 2017.

predict other states' negotiation strategies[82] – which may be perceived as enabling them to increasingly achieve these goals unilaterally, without compromise. Of course, many states worldwide are not solely or even mostly driven by such pragmatic considerations, in their support for the values enshrined in a norms-based global legal order. Yet the capabilities may well shift the values of the major leaders in AI development – such as the US, China, or Russia – on whose buy-in or acquiescence international bodies such as the UN have often been reliant.

4 Towards 'Innovation-Proof' Governance for Military AI?

Is there any hope for governance strategies that could adequately deal with such a wide range of challenge vectors generated by ongoing innovation in military AI? In this section, I first briefly sketch some preliminary caveats and considerations about whether or *why* we should even pursue the goal of 'innovation-proof' governance. I then discuss considerations for *how* we might achieve such governance, and the degree to which different strategies manage to address the disruptions detailed above.

4.1 *Why Innovation-Proof? Caveats and Merits*

In the first place, it is important to caveat the following discussion by recognizing that anticipatory or forward-looking governance may not be an unalloyed good – or a governance 'free lunch' – without its own distinct risks or drawbacks. I will therefore enumerate a number of counter-arguments, and offer conditional replies.

In the first place, while recent years have seen some work on 'legal foresighting' for new technologies,[83] the fundamental insight of the Collingridge Dilemma – that it is hard to anticipate the path or impact of new technologies in advance, and therefore hard to legislate appropriately ahead of time – is of course key. Indeed, empirically, (international) law scholars have a poor historical track record of jumping the gun on designing regulation for technologies which they (erroneously) thought to be imminent – from 1960s academic proposals for a 'Center of the Earth Treaty'; 1970s proposals to regulate weather

[82] cf Stephen Chen, 'Robots, Immune to Fear or Favour, Are Making China's Foreign Policy' *South China Morning Post* (30 July 2018) <www.scmp.com/news/china/society/article/2157223/artificial-intelligence-immune-fear-or-favour-helping-make-chinas> accessed 12 September 2018.

[83] Laurie, Harmon and Arzuaga (n 1).

control technology; or the 1982 Deep Sea Bed Mining Provisions in the 1982 Law of the Sea.[84] Critics might therefore argue that forward-looking initiatives may simply court irrelevance at best.[85]

Yet does innovation-proof governance require prediction? Indeed, the difficulty of accurate technological foresight arguably underscores rather than erodes the utility of pursuing governance strategies or instruments that are versatile even if – or precisely when – they cannot accurately predict developmental pathways. Moreover, even if accurate prediction of technological development were possible, this ability might in some circumstances, somewhat perversely, prove *counterproductive* to achieving governance consensus. After all, the more clearly a technology's future impacts are envisioned, the more clearly states today are able to project and articulate how their and other states' relative proficiency in the relevant areas of science and technology will eventually translate into concrete differential strategic advantages – and the more reluctant they may be to accede to governance regimes that would see them forgo such advantages. In other words, it clarifies the different stakes states have in the technology, which as noted by Picker, has historically proven an inhibition to effective international lawmaking.[86]

One example of this dilemma in an adjacent area of technology can be found in the area of quantum computing – future breakthroughs in which, it is anticipated, will yield large strategic advantages in a range of economic and military areas, including intelligence and information security. In the last few years, China has achieved notable successes in this realm, and some have argued that the country is well on the path to 'quantum hegemony'.[87] In this context, providing credible and concrete assessments of exactly what strategic capabilities quantum computing is set to unlock in the future might inhibit Chinese willingness to enter into global compacts on the technology. In that context, uncertainty may have its benefits; and governance regimes which are 'blind' but adaptive might prove more effective than governance regimes which are based on clear roadmaps of what interests and stakes, exactly, will be on the table.

[84] Picker (n 34) 184–187; Edith Brown Weiss, 'International Responses to Weather Modification' (1975) 29 International Organization 805.

[85] However, it is unclear if such initiatives were actually harmful, beyond perhaps squandering some academic energy and effort.

[86] Picker (n 35) 191–194.

[87] See Elsa B Kania and John K Costello, 'Quantum Hegemony? China's Ambitions and the Challenge to U.S. Innovation Leadership' (Center for a New American Security 2018) <s3.amazonaws.com/files.cnas.org/documents/CNASReport-Quantum-Tech_FINAL.pdf?mtime=20180912133406> accessed 11 January 2019.

A second, distinct counterargument is that trying to pursue 'innovation-proof' governance may constitute simple overreach, as trying to make current agreements too demanding or flexible to future hypotheticals might render them politically stillborn today. To be sure, negotiating an 'innovation-proof' governance regime on any given technology will indeed be politically difficult compared with negotiating one that narrowly responds to the particular problem at hand. But that may not be the relevant comparison. In another, longer-term perspective, investing in broader 'innovation-proof' governance early on may also be more achievable than intractable attempts to deal with 'jurisprudential space junk'[88] – to expand or reconfigure too-narrow regimes at a later stage, once unevenly shared interests in the technology are clear and entrenched, and when the regulatory window of opportunity has closed. Avoiding the second prong of the Collingridge Dilemma may certainly require additional up-front effort – but that governance investment may well pay off. Developing approaches to enable governance to track innovation cuts to the heart of questions over the simple effectiveness and relevance of international governance regimes, and their ability to achieve the spirit rather than just the letter of international law.

4.2 *Considerations for 'Innovation-Proof' Governance*

Designing innovation-proof governance regimes for any technology is clearly a complex and precarious endeavour. As such, the following is an exploratory and early discussion about different innovation-proofing strategies for the distinct governance challenges generated by ongoing MAI innovation. It is a discussion that is not about providing blueprints, but for the moment about simply sketching pitfalls and 'desiderata'[89] for governance instruments that could best facilitate or accommodate ongoing technological innovation.

4.2.1 Tackling Direct Disruption: From Firefighting to Open-Ended Adaptation?

In the first place, regarding the direct disruption created by direct new 'entities' or by new types of behaviour, three types of governance responses appear possible, which are not all equally promising.

[88] Crootof (n 11).

[89] For a similar, more extensive approach to formulating broad 'desiderata' that are salient across a wide range of policy issues (albeit in the context of governing eventual future general AI systems), see also Nick Bostrom, Allan Dafoe and Carrick Flynn, 'Public Policy and Superintelligent AI: A Vector Field Approach' in SM Liao (ed), *Ethics of Artificial Intelligence* (Oxford University Press 2019).

The simple reactive approach to innovation, which we might call 'firefighting', is to formulate particular new specific (*sui generis*) governance regimes to new technologies as they emerge. By and large, this has been the default mode for how international law has dealt with new technologies,[90] and there are areas, such as chemical weapons, in which this has perhaps been broadly effective. Yet as seen through the challenges discussed above, this mode will be inadequate for governing ongoing MAI innovation. Picker has argued that while the treaty process should usually be preferred over the slow customary international law process, even narrow treaties are rapidly made obsolete by the very speed with which technology changes,[91] for the reasons discussed above. A 'firefighting' approach falls victim to all of the political problems and path-dependencies described by the second prong of the Collingridge Dilemma.

Moreover, it remains questionable whether 'firefighting' can be applied prospectively, in a way that tracks inevitable changes in technology and ensures governance is 'innovation-proof'. While Garcia has called for the formulation of 'preventative security governance' regimes – defined as 'the codification of specific or new global norms, arising from existing international law that will clarify expectations and universally agreed behaviour on a given issue-area'[92] – she also notes that 'preventive' bans have only succeeded for just two weapons systems (expanding bullets in 1899 and blinding lasers in 1998).[93] Given the likely rapidity with which MAI innovation will continue, and the likely immense tactical and strategic benefits – far beyond those of lasers – which these systems will offer states, 'firefighting' has critical drawbacks, and appears wholly insufficient as a framework for innovation-proof governance.

In the second place, one might seek to rework and expand some version of the notion of 'technology neutrality' to the international legal frameworks around MAI. Such frameworks would then seek to emphasize (and prohibit) additional isomorphic functions or impacts of AI, rather than one or another specific technological architecture. Indeed, this is what scholars in IHL already work towards when they argue that it is the 'the context and the humanitarian consequences'[94] of a weapon that matter for IHL purposes. Innovation-proof governance regimes of MAI could therefore try to expand the set of relevant

[90] Picker (n 35).

[91] Ibid 184–186.

[92] Garcia (n 2) 95; cf also Jürgen Altmann, *Military Nanotechnology: Potential Applications and Preventive Arms Control* (Routledge 2007).

[93] Garcia (n 2) 102.

[94] Ibid 105.

impacts of AI to beyond a narrow notion of 'autonomy', to include criteria such as 'susceptibility towards systemic failure' (though this would open up new definitional debates). Technological neutrality has the advantage of not requiring clear advance foresight of the future pathways of MAI technological development – a degree of foresight which might, as discussed, be both very difficult to achieve, and potentially politically counterproductive.

However, one drawback with technology-neutral regulation is that the functions served by many AI systems are of a sufficiently high level of generality or abstraction – 'prediction', 'pattern recognition', 'autonomous reaction' – that any governance regime seeking to define terms at that level is almost bound to become enormously (over-)inclusive. In a political sense, that may make states parties loath to enter overtly broad (and potentially restrictive) regimes in advance. Nonetheless, the example of technological neutrality illustrates how, in a general sense, innovation-proof governance might seek to emphasize the formulation and use of standards rather than of rules, since the latter allow more flexibility.[95]

Thirdly, one can consider varieties of open-ended, '*adaptive*' governance instruments, which allow progressive, modular, or iterative expansion of governance regimes. In fact, the design of the UN Convention on Certain Conventional Weapons (CCW) offers one early example of such – with its Additional Protocols on non-detectable fragments, landmines, incendiaries, booby traps, and blinding lasers. However, in spite of its ongoing discussions on banning Autonomous Weapons, it might be argued that the CCW itself would be a suboptimal regime that could not scale well with the challenges introduced by ongoing MAI innovation. This is both because of the Convention's general problems (its lack of mechanisms for verification or enforcement of compliance), as well as specifically because of its narrow inclusion criteria (indiscriminate or excessively injurious weapons), which cannot extend well to the changing problem portfolio of MAI in areas such as strategic stability and safety.

Unfortunately, by themselves, it appears that none of these three approaches is entirely satisfactory. 'Firefighting' treaties are too piecemeal and reactive; 'technologically neutral' governance may be perceived as prohibitively over-inclusive by states. Taken in general, the idea of a modular treaty regime for military AI – directed by broader criteria than is the CCW – would appear at least somewhat promising, combining as it would the political feasibility of specific treaties with the open-ended flexibility of technological neutrality. Such designs could and should be explored much further.

[95] Picker (n 35) 185.

4.2.2 Tackling Indirect Disruption: Aligning Interests and Values, Preserving the Framework

Underlining the first-order question of what governance instruments can best evolve along the pace and path of innovation, is a second-order question of measures that must be taken to preserve the practical and international conditions for such governance in the first place, in the context of changes wrought by that AI. Rather than suggesting one or another specific governance approach or instrument, the indirect challenges created by AI innovation instead produce general considerations about the rationale and scope of governance, and underlying political perceptions that must be shifted.

In the first place, the risk that new MAI capabilities might significantly shift or expand the 'problem portfolio' beyond the scope of existing legal instruments, calls for governance that likewise evoke all four areas at risk. Governance campaigns or initiatives should therefore pursue AI that is not just 'Lawful' and 'Ethical', but also 'Stabilizing' and 'Safe' (perhaps yielding the amusing acronym 'L.E.S.S. AI'). Indeed, non-state actors can take meaningful steps, in this regards. For instance, it has been argued that during the Cold War, 'epistemic communities' of experts played a large role in the successful reframing of nuclear non-proliferation and arms control as goals shared to some extent between rival great powers[96] – and that properly organized, contemporary epistemic communities, appropriately organized, could repeat this feat in the context of emerging military AI.[97] Much of the work of such communities should be aimed at reframing AI, not just in the context of public debate or legal scholarship, but also to states or national security establishments.

Related to this, there are distinct points of intervention for epistemic communities to counter the changes in incentives which MAI innovation might produce for states. For instance, efforts could aim at reducing, downplaying or qualifying the prospective *benefits* of new MAI capabilities and at framing unrestrained MAI development as not just unethical or unlawful, but also simply risky, unreliable or destabilizing. To give one example; former US Secretary of the Navy Richard Danzig has recently questioned the narrow military equivalence of (AI) technological supremacy with national security; instead, he suggests the unrestrained pursuit of military technological advantage drives a risky 'technology roulette', where militarization by one state drives the proliferation and early deployment of accident-prone military AI systems all

[96] Emanuel Adler, 'The Emergence of Cooperation: National Epistemic Communities and the International Evolution of the Idea of Nuclear Arms Control' (1992) 46 International Organization 101.

[97] Maas (n 22).

round.[98] The importance of reiterating the shared fundamental interests states have in a stable international system – to draw attention to the strategic goods they gamble and stand to lose in the pursuit of narrow tactical benefits – has also been articulated in other contexts such as the governance of cyberwar.[99]

In addition, governance measures might aim to maintain or increase *barriers* to access and proliferation of new (military) AI innovations. Although this might not come without its own spate of (legitimacy) problems and drawbacks, it might contain the number of actors that could slip through the net of compliance verification. Alternatively, such measures could at least reduce the 'speed (to market)' of new MAI innovations – the rate at which new innovations may diffuse and see deployment before governance instruments have had time to take stock.

Moreover, we can seek to improve capabilities for the detection of treaty compliance violations, potentially even through new monitoring capabilities offered by AI itself.[100] Efforts might also aim to increase the public political *cost* of pursuing or developing a broad range of military AI, though governance advocates may soon need to articulate new and additional sources of public opprobrium beyond the old trope of 'killer robots'.

Finally, regarding the risk that AI might structurally empower illiberal states over democratic ones, or might even shift the core *values* of its leading developers towards unilateralism, and thereby erode the fabric of international law, is so fundamental and constitutive that no one should attempt a meaningful answer within a single paragraph. Suffice it to say that a necessary condition for addressing this shift may be the articulation, in the coming years, of compelling and encompassing vision of how states can instead harness 'AI for global good'.[101]

5 Conclusion

Is it possible to design 'innovation-proof governance' approaches to deal with future advances in a technology? Or must we resign ourselves to a perpetual

[98] Danzig (n 52).

[99] Garcia (n 2) 109.

[100] cf also Miles Brundage, 'Scaling Up Humanity: The Case for Conditional Optimism About Artificial Intelligence' Should we fear artificial intelligence? (European Parliamentary Research Service, Scientific Foresight Unit (STOA) 2018) <www.europarl.europa.eu/RegData/etudes/IDAN/2018/614547/EPRS_IDA(2018)614547_EN.pdf> accessed 27 November 2018.

[101] Dafoe (n 25) 48–51; Cave and Ó hÉigeartaigh (n 4).

regulatory firefighting amidst increasing clouds of less and less workable 'jurisprudential space junk'?

This is a question of general relevance, to any rapidly developing and broadly applicable technology – and especially to technologies which we seek to regulate at a global level. It is also a question of particular urgency in the debates around lethal autonomous weapons. This is because 'killer robots', I have argued, are not the final or biggest 'robotic' challenge to international law; instead, they will likely prove merely one in a series – or rather a web – of future innovations in military AI: innovations which – barring a good answer to the above question – we may expect to repeatedly and structurally challenge the efficacy, resilience or relevance of extant international law and governance.

I have argued that future MAI innovations or systems will challenge and disrupt these legal frameworks both directly (as new types of MAI systems or capabilities elude inclusion under existing regimes), and indirectly (as new MAI systems shift the technology's risk landscape, or change the incentives or values of the states parties developing them). This identifies a set of critical boundary conditions – but also policy levers – for approaches aiming to preserve important international norms and values.

Ultimately, this paper has not managed to sufficiently resolve or dissolve the Collingridge Dilemma posed by military AI, or to sketch out definitive blueprints for innovation-proof governance of the same. In that sense, this analysis is still incomplete: it suggests that innovation-proof governance may be achievable, but that it should involve notable changes, both from the piecemeal and reactive way international technology governance has proceeded so far, but also, specifically, from the narrow way we still approach the regulation and control of military AI. Nonetheless, it is hoped that this initial exploration has helped sketch the nature of the challenges, and offered some considerations for governance strategies that can ensure that military AI remains lawful, ethical, stabilizing, and safe.

Acknowledgements

An initial version of this argument was presented at the conference Beyond Killer Robots – Networked Artificial Intelligence Disrupting the Battlefield, organized on 15–16 November 2018 by the University of Copenhagen (Faculty of Law, Centre for International Law, Conflict and Crisis and the AI and Legal Disruption research unit). I would like to thank the attendees of the conference – in particular Léonard van Rompaey, Hin-Yan Liu, and Sophie-Charlotte

Fischer – for valuable feedback on the original idea, and Maaike Verbruggen and Merel Noorman for thought-provoking discussion during the panel session. I would also like to thank two anonymous reviewers for key input, and Thomas Basbøll for invaluable writing advice. Any remaining errors or inaccuracies are my own.