UNIDIR

Safety, Unintentional Risk and Accidents in the Weaponization of Increasingly Autonomous Technologies

No. 5 UNIDIR RESOURCES

Acknowledgements

Support from UNIDIR's core funders provides the foundation for all of the Institute's activities. In addition, dedicated project funding was received from the governments of Canada, Germany, Ireland, and the Netherlands. The Institute would also like to thank Elena Finckh for her assistance with this project.

About the Project "The Weaponization of Increasingly Autonomous Technologies"

Given that governments have a responsibility to create or affirm sound policies about which uses of autonomy in weapon systems are legitimate—and that advances in relevant technologies are also creating pressure to do so—UNIDIR's work in this area is focused on what is important for States to consider when establishing policy relating to the weaponization of increasingly autonomous technologies.

See http://bit.ly/UNIDIR_autonomy for Observation Papers, audio files from public events, and other materials.

About UNIDIR

The United Nations Institute for Disarmament Research—an autonomous institute within the United Nations—conducts research on disarmament and security. UNIDIR is based in Geneva, Switzerland, the centre for bilateral and multilateral disarmament and non-proliferation negotiations, and home of the Conference on Disarmament. The Institute explores current issues pertaining to the variety of existing and future armaments, as well as global diplomacy and local tensions and conflicts. Working with researchers, diplomats, government officials, NGOs and other institutions since 1980, UNIDIR acts as a bridge between the research community and governments. UNIDIR's activities are funded by contributions from governments and donor foundations.

Note

The designations employed and the presentation of the material in this publication do not imply the expression of any opinion whatsoever on the part of the Secretariat of the United Nations concerning the legal status of any country, territory, city or area, or of its authorities, or concerning the delimitation of its frontiers or boundaries. The views expressed in this publication are the sole responsibility of UNIDIR. They do not necessarily reflect the views or opinions of the United Nations or UNIDIR's sponsors.

www.unidir.org

Safety, Unintentional Risk and Accidents in the Weaponization of Increasingly Autonomous Technologies

Recent international attention on autonomous weapon systems (AWS) has focused on the implications of what amounts to a 'responsibility gap' in machine targeting and attack in war. As important as this is, the full scope for accidents created by the development and deployment of such systems is not captured in this debate. It is necessary to reflect on the potential for AWS to fail in ways that are unanticipated and harmful to humans—a broader set of scenarios than simply those in which international humanitarian law applies.

Of course, any complex, hazardous technology carries 'unintentional' risk, and can have harmful results its designers and operators did not intend. AWS may pose novel, unintended forms of hazard to human life that typical approaches to ensuring responsibility do not effectively manage because these systems may behave in unpredictable ways that are difficult to prevent. Among other things, this paper suggests human-machine teams would, on their own, be insufficient in ensuring unintended harm from AWS is prevented, something that should bear on discussions about the acceptability of deploying these systems. This is the fifth[1] in a series of UNIDIR papers on the weaponization of increasingly autonomous technologies.[2]

Context

Advances in a variety of fields have led recently to the development of machines with increasing capacities for autonomous assessment and action, and which offer transformative potential for human society.[3] As an earlier paper in this UNIDIR series noted, 'Machines and systems that have increasing amounts of autonomy will require that we carefully examine aspects of our legal codes, privacy regulations, and health and safety policies. These technologies also raise fundamental questions about how we—as societies and individuals—perceive, interact with and relate to machines.'[4] Already, evidence of such re-examination (or calls for it) can be seen in domains as seemingly varied as the development of self-driving cars[5], drone package delivery,

1 For more information about UNIDIR's project 'The Weaponization of Increasingly Autonomous Technologies', see http://bit.ly/UNIDIR_autonomy.

2 UNIDIR would like to acknowledge the thoughtful contributions of the participants in an April 2016 expert meeting convened by UNIDIR: Darren Ansell, Gwendolyn Bakx, Aude Billard, John Borrie, David Danks, Neil Davison, Sean Legassick, Patricia Lewis, Pavel Podvig, Sabine Roeser, Heather Roff, Paul Scharre, Kerstin Vignard and Wendell Wallach. The views expressed in this paper are the sole responsibility of UNIDIR, and do not imply the endorsement of these participants.

3 For one accessible, if polemical, overview of some of these developments see M. Ford, *The Rise of the Robots: Technology and the Threat of Mass Unemployment, OneWorld*, 2015.

4 UNIDIR, *Framing discussions on the weaponization of increasingly autonomous technologies*, No. 1, 2014, p. 2: http://bit.ly/UNIDIR_autonomy.

5 P. Lin, 'The ethics of autonomous cars', *The Atlantic*, 8 October 2013: www.theatlantic.com/technology/archive/2013/10/the-ethics-of-autonomous-cars/280360.

robotic health care assistance for the elderly, household cleaning robots[6], and the delivery of lethal force by robots in certain law enforcement contexts[7].

The weaponization of increasingly autonomous technologies is one domain of growing interest and concern at the international level. Notably, AWS (Autonomous Weapon Systems) or LAWS (Lethal Autonomous Weapon Systems) have been discussed over the last several years in talks under the auspices of the 1980 United Nations Convention on Certain Conventional Weapons (CCW) in Geneva[8], and look likely to be on its agenda beyond 2016. Even though few, if any, weapon systems have yet been deployed that could be described as autonomous in any general way, some sense that the world is on the cusp of an era in which 'killer robots' could be developed in earnest and deployed on the battlefield.[9]

In terms of the hazards AWS might generate, international policy discussions have mainly revolved around whether there would be, in effect, a 'responsibility gap'[10] relating to machines deciding *when*, *where* and *how* to attack *which* objects and people in war.[11] The prospect of militaries delegating targeting and attack decisions to machines raises a number of important challenges in terms of transparency, ethics and accountability with international humanitarian law (IHL) and human rights law.[12] These questions have yet to be resolved. It has led some academics, industry experts and campaigners to call for a pre-emptive prohibition or moratorium on machines deploying such capabilities.[13] A contrasting view is that it is unrealistic to think that emerging technological capabilities like systems for autonomous targeting and attack will be abandoned, or even meaningfully constrained. Rather, their emergence is a condition that must be managed.[14] Others, including some States, suggest that issues around AWS should be handled within national processes of development of new technological capabilities, like those used for existing military weapon systems.[15] In a related vein, some policy practitioners have argued that

6 C. Metz, 'Forget Doomsday A.I.—Google is worried about housekeeping bots gone bad', *Wired*, 21 June 2016: www.wired.com/2016/06/forget-doomsday-ai-google-worried-housekeeping-bots-gone-bad/.

7 P. Tucker, 'Military Robotics Makers See a Future for Armed Police Robots', *Defense One*, 11 July 2016: www.defenseone.com/technology/2016/07/military-robotics-makers-see-future-armed-police-robots/129769/.

8 V. Boulanin, *Mapping the debate on LAWS at the CCW: taking stock and moving forward*, SIPRI, March 2016.

9 See, for instance, Human Rights Watch, *Losing Humanity: The Case Against Killer Robots*, 19 November 2012: www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots.

10 See D.G. Johnson, 'Technology with No Human Responsibility?', *Journal of Business Ethics*, 2014, pp. 1–9: www.law.upenn.edu/live/files/3774-johnson-d-technology-with-no-responsibility.

11 There are exceptions. In April 2016, a number of expert panelists at the CCW alluded to possible unintended risks AWS development and deployment would create.

12 For example, see C. Heyns, *Report of the Special Rapporteur on extrajudicial summary or arbitrary executions*, United Nations Human Rights Council, A/HRC/23/47, 2013: www.ohchr.org/Documents/HRBodies/HRCouncil/RegularSession/Session23/A-HRC-23-47_en.pdf.

13 Future of Life Institute, *Autonomous weapons: An open letter from AI & robotics researchers*, 28 July 2015: www.futureoflife.org/AI/open_letter_autonomous_weapons.

14 B. Allenby, 'Emerging technologies and the future of humanity', *Bulletin of the Atomic Scientists*, 2015, vol. 71, no. 6, pp. 29–38, p. 36: 'we need to stop thinking of "problems" with "solutions", and think more in terms of "conditions" that will require long-term, adaptive management. Challenges such as ISIS and climate change will not be solved, but they can and must be managed in light of other relevant goals. In this, the experience with nuclear weapons is instructive: They are not a problem that can be unmade, but they are a condition that can be, and has so far been, relatively successfully managed.'

15 For example, obligations under Article 36 of the Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I), 8 June 1977.

all AWS should be subject to 'meaningful human control' (MHC)[16] in individual attacks. MHC has not been tightly defined—let alone generally agreed—by States, though some have lent their support to the concept in their CCW statements.[17]

Reflecting on responsibility

Apportioning responsibility for targeting and the consequences of individual attacks caused by AWS takes us so far. But is it enough? Features in the development of some of these systems, such as some machine learning techniques, suggest an increased potential for novel problems in complex and ambiguous real world environments, including—but not limited to—those in which IHL applies.

Experts themselves have begun to call for more studied attention to understanding and preventing machine learning-related accidents, which some have defined as 'unintended and harmful behaviour that may emerge from machine learning systems when we specify the wrong objective function, are not careful about the learning process, or commit other machine learning-related implementation errors.'[18] At present, because many of the applications of such autonomous systems are in their infancy, the consequences of such accidents are probably limited. However, this is likely to change in time, not least if there is further convergence between the development of autonomous systems based upon machine learning and efforts to weaponize them.

A second reason for critical reflection by policy makers is that no technology is fail-safe. Moreover, it is understood from the operation of existing *hazardous* technologies of various kinds that despite careful design and operation, catastrophic failure is a persistent risk. This is due, as will be shown, to complexity and tight coupling in these systems, which periodically lead to 'system accidents' at odds with the intentions of their designers and operators—even when robust accountability procedures exist.[19] At least some AWS would likely be complex and tightly coupled, or be embedded within broader complex, tightly coupled systems.

A third reason for reflection is that some experts emphasize the necessity of humans and machines working as teams—or joint cognitive systems (JCS)—with each doing what it is best at, as a satisfactory way of managing increasing autonomy in weapons. However, to what extent are JCS really a solution to preventing AWS-related accidents? Research from aerospace, for instance, has shown that increasing machine autonomy can make the user's job more complicated. The cognitive workload of the human in the system sometimes increases substantially, and they can suffer from lower situational awareness because it often leaves them out of the loop and struggling to understand what the system is doing so as to supervise and intervene in time-critical situations.[20]

---

16 For instance, see Article 36, *Structuring debate on autonomous weapons systems*, Article 36 briefing paper, November 2013: www.article36.org/wp-content/uploads/2013/11/Autonomous-weapons-memo-for-CCW.pdf.

17 UNIDIR, *The weaponization of increasingly autonomous technologies: Considering how Meaningful Human Control might move the discussion forward*, no. 2, 2014: http://bit.ly/UNIDIR_autonomy.

18 D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman and D. Mané, 'Concrete Problems in AI Safety', *arXiv*, July 2016: https://arxiv.org/pdf/1606.06565v2.pdf, pp. 1–2.

19 C. Perrow, *Normal Accidents: Living With High-Risk Technologies*, Basic Books, 1984.

20 See M.R. Endsley, 'Building resilient systems via strong human systems integration', *Defense AT&L*, (Jan–Feb 2016), pp. 6–12, p. 11: dau.dodlive.mil/files/2015/12/Endsley.pdf.

Description of terms

Autonomous systems are those that operate without human intervention in the physical world or some kind of digital or virtual environment. Autonomous systems select actions that operate upon that environment based on some kind of assessment of the environment's current state. Usually this will be in pursuit of some set of pre-specified goals. (This distinguishes them from *purely automated* systems—the latter are governed by prescriptive rules that permit no deviations.[21]) One point of distinction within the category of autonomous systems some have made is that, in broad terms, 'systems incorporating *autonomy at rest* operate virtually, in software, and include planning and expert advisory systems, whereas systems incorporating *autonomy in motion* have a presence in the physical world and include robotics and autonomous vehicles.'[22]

Autonomy in weapon systems is understood, for the purposes of this paper, as a spectrum of capability 'moving from remotely controlled systems on one side to autonomous weapon systems on the other. Autonomy increases as one moves along the spectrum from objects controlled by human operators from a distance (such as remotely piloted unmanned aerial vehicles), to automatic and automated systems, to fully autonomous ones.'[23] Currently, there is no settled definition of AWS at the international level. This paper aligns itself with common definitions of AWS as 'robot weapons that once launched will select and engage targets without further human intervention.'[24] A point to bear in mind, however, is that if autonomous systems at rest become intimately involved in targeting advice, selection and strike planning before launch, this definition may be insufficient in order to consider the breadth of potential AWS failures.

Artificial intelligence (AI) is an imprecise but commonly used term. Sometimes it refers to the long-term horizon of software systems with a quality often referred to as *general* or *general-purpose intelligence*. Yet AI is also often referred to as more immediate software techniques that can tackle problems previously thought to require human intelligence. The distinction is sometimes couched in terms of general AI and narrow AI. However, what happens to be considered 'narrow AI' changes over time. An iconic example is chess-playing: where once a computer chess player capable of defeating a human being was considered very much the province of AI, today's smartphone chess apps can beat most, if not all, humans. Yet it is clear that this does not make chess apps or the 'smart' phones they run on general-purpose intelligence. There is a lot of excitement about the autonomous systems it might be possible to build with AI technologies, for instance truly autonomous vehicles. However, autonomous systems are not necessarily AI, and systems built using AI are not necessarily autonomous.

Machine learning is a specific approach to AI, and one that has achieved much success over the last five to ten years. By way of example, machine learning is currently used in applications for automatic translation, photo labelling, spam detection, and video recommendations. There are many different approaches within machine learning, although generally speaking it involves 'training' a statistical model with some input data (normally in large volumes) so that the model produces useful output when presented with novel data. How 'usefulness' is determined depends to some extent on the type of machine learning in

21 United States Department of Defense (DoD), *Report of the Defense Science Board Summer Study on Autonomy*, June 2015, p. 4.

22 United States DoD, *Report of the Defense Science Board Summer Study on Autonomy*, p. 5.

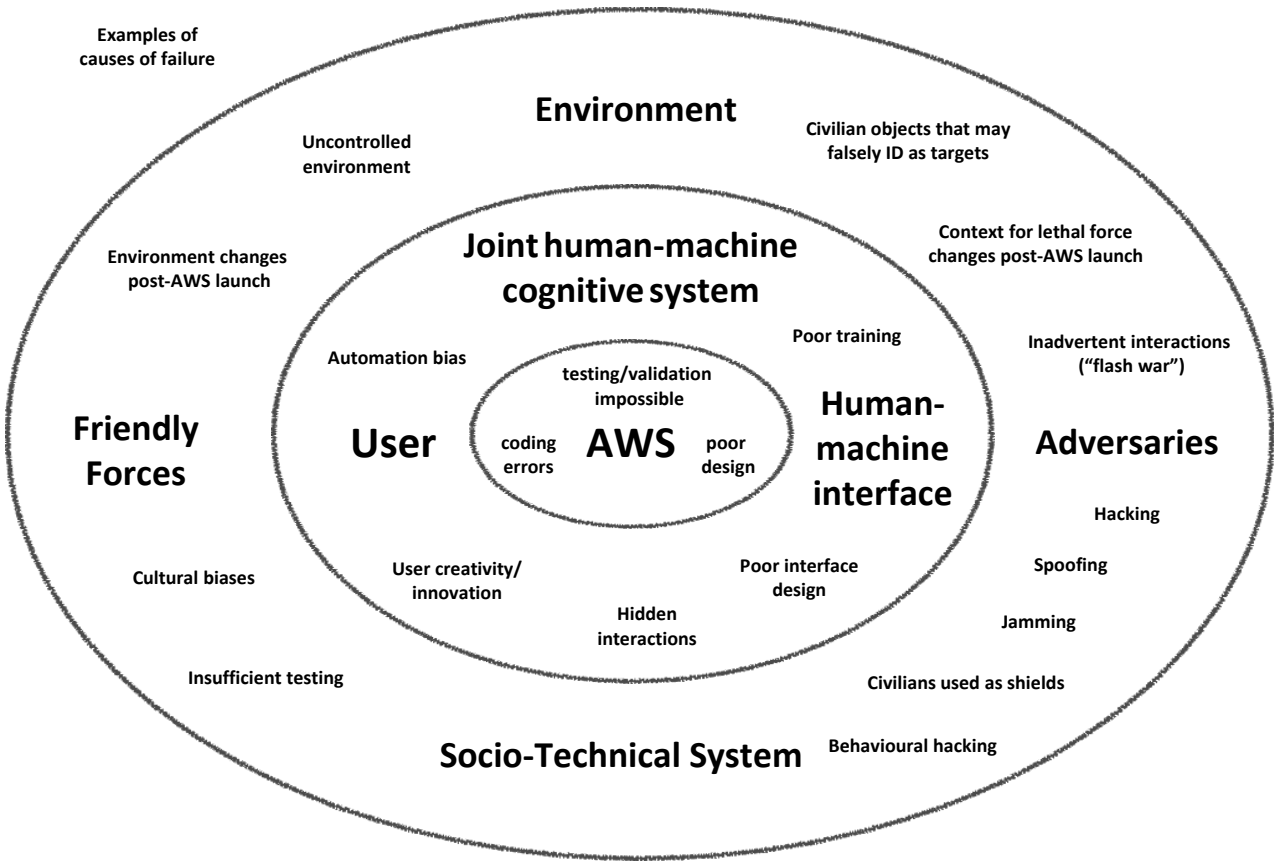
23 UNIDIR, *Framing discussions on the weaponization of increasingly autonomous technologies*, no. 1, Geneva, 2014, p. 2: http://bit.ly/UNIDIR_autonomy. This paper also sets out a list of variables that can comprise assessments of autonomy.

24 J. Altman, P. Asaro, N. Sharkey and R. Sparrow, 'Armed military robots: editorial', *Ethics and Information Technology*, vol. 15, 2013, pp. 73–76, p. 73.

use. For example, a useful output for a photograph labeller is accurate labels for the objects depicted in a photo, and for a spam detector it would be a simple binary (i.e. spam/not spam).

Failure in the context of AWS in this paper refers to 'actual or perceived degradation or loss of intended functionality or inability of the system to perform as intended or designed. Failures can result from a number of causes, including, but not limited to, human error, human-machine interaction failures, malfunctions, communications degradation, software coding errors, enemy cyber attacks or infiltration into the industrial supply chain, jamming, spoofing, decoys, other enemy countermeasures or actions, or unanticipated situations on the battlefield.'[25] Failure might not mean that a system fails to function; it might still do so, but in a very altered way. Figure 1 provides some examples of potential causes of failure in AWS.

Figure 1. Examples of potential causes of failure in AWS



Source: based on discussion at April 2016 UNIDIR expert meeting

Accidents, as used here, are failures in AWS that result in unintended and harmful outcomes for humans. By nature of their design and operation, AWS would be *intended* to cause harm to human beings *within certain parameters*. So, accidents involving AWS constitute events in which AWS cause harm to humans to which harm was not intended, or at a time or place not intended. Accident is a narrower term than 'failure' as failures are not necessarily harmful. Accident is also a narrower term than **inadvertent risk**, which could cover phenomena other than accidents such as assumptions, processes or socio-technical systems that may elevate the likelihood or consequences of failures or accidents without being the direct cause.

25 United States DoD, *DoD Directive: Autonomy in Weapon Systems (No. 3000.09)*, 21 November 2012, p. 14.

Observations: accidents will happen

We now turn to five observations about safety and responsibility challenges related to AWS.

1. Accidents are inevitable in complex and tightly coupled systems

It is likely that autonomous weapon systems will be both highly complex and very tightly coupled. (Tight coupling means that 'there is no slack or buffer or give between two items. What happens in one directly affects what happens in the other.'[26]) In complex systems there are a lot of *common mode* connections between components, which means it is not necessarily clear which of them has gone wrong when there is a failure.

History shows that certain kinds of system for the management of hazardous technology possessing significant levels of automation can fail in ways not anticipated by their human designers and operators.[27] In those contexts, apparently simple or trivial errors or failures can rapidly engender situations in which 'system' failures occur. Serious accidents sometimes result. This is because these systems' complex and tightly coupled nature hides crucial phenomena from the view of the human operators—either literally, or because things occur too quickly for them to respond sufficiently. Events such as the Chernobyl nuclear reactor meltdown in 1986, the destruction of two NASA space shuttles with the loss of their entire crews, and the Deep Water Horizon oil rig disaster in 2010 show that catastrophic failures occur despite careful technological design and planning, organizational control and training, and the addition of multiple technical redundancies.

Hidden interactions in complex systems can create feedback loops that are invisible to operators for a time because their existence cannot be directly detected, only inferred. Operators may not understand these feedback loops—or even realize they exist—if these are situated outside their mental models of how the system works. As Perrow observed of the 1979 Three Mile Island nuclear accident, 'seeing is not necessarily believing: sometimes we must believe before we can see'.[28] In that case, the human operators of the nuclear power plant literally did not believe some of the readings their instrumentation gave them, because it did not fit with their belief about how the plant could technically operate. Hidden interactions can happen despite careful design and testing of the system.

Handing controls and decisions off to machines can sometimes help in the management of hazardous technologies. But experts have also observed that such delegation can decrease the flexibility of the operator to correct minor failures before they can become major ones, and masks the hidden interactions discussed above. Autonomous systems such as remote submersibles[29] or stealthy unmanned aerial vehicles[30], for instance, might be accorded

26 Perrow, *Normal Accidents*, pp. 89–90.

27 *Operator* here means the human agent(s) who establish the parameters for the autonomous functions of a technology to operate, who monitor its operation—however distantly—and who, presumably, would be held accountable for accidents or mishaps involving that system. See presentation by J. Borrie, 'Safety aspects of "meaningful human control": Catastrophic accidents in complex systems', at UNIDIR event "Weapons, Technology and Human Control", New York, 16 October 2014: http://bit.ly/UNIDIR_autonomy. A human in a JCS would constitute an operator.

28 Perrow, *Normal Accidents*, p. 9.

29 UNIDIR, *Testing the Waters: The weaponization of increasingly autonomous technologies in the maritime environment*, no. 4, 2015: http://bit.ly/UNIDIR_autonomy.

30 See also United Nations Office for Disarmament Affairs, *Study on Armed Unmanned Aerial Vehicles Prepared on the Recommendation of the Advisory Board on Disarmament Matters*, United Nations, October 2015, pp. 47–8: www.

considerable contextual latitude to decide and act by their operators because of challenges to continuous or even regular communication. It is thus possible to envisage AWS interactions being obscured. Pointing to instances of 'flash crashes' in financial systems—where unintended interactions between autonomous trading algorithms were exacerbated by the high speed of the transactions themselves—some experts have raised concerns about unanticipated and emergent behaviour in future AWS.[31]

2. Hidden interactions have particular risk potential in systems reliant on machine learning

The kinds of capabilities militaries desire from autonomous machine systems are likely to require the application of machine learning techniques in order to be realized.

Machine learning can be immensely powerful when applied to suitable problems. Moreover, machine learning systems can exhibit capabilities that were previously intractable for computer software. One recent public demonstration of success in this field was AlphaGo, an AI built by DeepMind (owned by Google), which defeated a top human Go player in 2016.[32] Another example of machine learning's application is for improving medical diagnosis, something to which Watson, a system developed by IBM and which earlier achieved fame in winning at a television game show called *Jeopardy*, has been turned to.[33]

Meanwhile, many of the tasks of a soldier are dirty, dull or dangerous. Increasingly autonomous machines offer humans the prospect of relief from some of these tasks. As a consequence, robots with some degree of *autonomy in motion* are already making their way into military roles[34] whether as 'pack-bots', kits to allow supply trucks to 'self-drive' in supply convoys, or swarming small marine vessels to protect naval vessels from suicide attacks. But not all of these systems depend on machine learning techniques. And, for now, outside of very constrained contexts such as close-in air defence (systems sometimes described as being *semi-autonomous*), decisions about targeting and attack in war remain in human hands.

However, as increasingly autonomous machines and systems of machines become an accustomed part of the fabric of life in some militaries—and indeed in the wider civilian world—the prospect of their use for targeting and attack functions may well grow, especially as war becomes ever more kinetic, and novel roles are envisaged.

- For instance, unmanned combat aerial vehicles (UCAVs) are becoming more sophisticated, and future systems may be able to operate in differing modes with different levels of

---

un.org/disarmament/publications/more/drones-study/.

31 See presentation by P. Scharre, 'Flash War: Autonomous weapons and strategic stability', at UNIDIR event "Understanding Different Types of Risk", Geneva, 11 April 2016: http://bit.ly/UNIDIR_autonomy; see also Australian Broadcasting Corporation, 'Interview: Robotics—Professor Noel Sharkey', 21 May 2015: www.abc.net.au/lateline/content/2015/s4240313.htm.

32 C. Metz, 'How Google's AI Viewed the Move No Human Could Understand', *Wired*, 14 March 2016: www.wired.com/2016/03/googles-ai-viewed-move-no-human-understand.

33 N. Leske, 'Doctors seek help on cancer treatment from IBM supercomputer', *Reuters*, 9 February 2013: in.reuters.com/article/ibm-watson-cancer-idINDEE9170G120130208.

34 See for instance, P. Tucker, 'These Are the Decisions the Pentagon Wants to Leave to Robots' *Defense One*, 14 December 2015: www.defenseone.com/technology/2015/12/these-are-decisions-pentagon-wants-leave-robots/124480.

autonomy—from remotely piloted to mostly or even fully autonomous.[35] It has been proposed, for example, that the U.S. Navy incorporate UCAVs into its carrier airwings over the next decade to remedy perceived deficiencies in littoral deep-strike capabilities.[36]

- Autonomous marine systems of various kinds are being developed and tested now by navies for tasks such as surveillance, mapping and mine detection, and there is interest in using autonomous systems for area denial, or tracking and destroying submarines in the future.[37]

Machine learning techniques are relevant to enabling or improving the performance of all of these kinds of systems. This is because automation governed by prescriptive rules that permit no deviations will be insufficient to ensure they can achieve the goals of their human operators.

Meanwhile, machine learning-based techniques in *autonomy at rest* systems conceivably have no less important roles to play in the military sphere than *autonomy in motion* systems like robot weapons. Systems that process large amounts of sensory and intelligence data in order to aid military decision making and logistics planning hold obvious appeal for the military advantage this might convey[38]—capabilities that might eventually extend to pre-vetting and even selecting targets in ways that shape decision makers' perceptions.

This has major implications for safety because, for all of their promise, machine learning-based systems present challenges:

- Machine learning systems, and in particular neural networks and similar architectures, are **complex**. Their effectiveness is the result of their mathematical properties and complex relationships between opaque internal parameters. It means that even the operators running the systems do not have a complete understanding of the underlying learned logic of, say, a trained deep learning network.
- Currently it is not possible to produce **formal proofs** of the behaviour of machine learning systems. This poses challenges for attaining the levels of formal verification that are demanded for many software code-based systems, especially for systems performing critical functions on which human lives may rely.
- Machine learning systems are stochastic, and **so predictability is a challenge**. The machine is not constrained by human experience or expectations.[39] Systems can be tested, and their behaviour observed in a range of scenarios—but this is a long way from formal verification.
- **Interpretability** (the ability to analyse and assess the 'learnt logic' on a machine learning system) is in its infancy. At present, techniques are crude.

---

35 See United Nations Office for Disarmament Affairs, *Study on Armed Unmanned Aerial Vehicles*, Chapter 3.

36 J. Hendrix, *Retreat from Range: The Rise and Fall of Carrier Aviation*, Center for a New American Security, October 2015: https://s3.amazonaws.com/files.cnas.org/documents/CNASReport-CarrierAirWing-151016.pdf.

37 D. Hambling, *The Inescapable Net: Unmanned Systems in Anti-Submarine Warfare* (BASIC Parliamentary Briefings on Trident Renewal), British-American Security Information Council, 2016.

38 See United States DoD, *Defense Science Board Summer Study on Autonomy*, especially Chapter 4 ('Strengthening operational pull for autonomy').

39 For a discussion, see J. Tapson, 'Google just proved how unpredictable artificial intelligence can be', Business Insider UK, 19 March 2016: uk.businessinsider.com/google-just-proved-how-unpredictable-artificial-intelligence-can-be-2016-3.

- Machine learning systems tend to be **tightly coupled**. Many applications involve a single deep learning network that is largely opaque once it is trained.

In view of the challenges above, most current civilian applications of machine learning systems provide humans with tools—complementing and augmenting human expertise, rather than replacing it. Humans playing a key role in control of the overall system (whatever that may be) is generally seen as vital to its safe and reliable functioning. In particular, *observability* of the internal processes of machine learning-based systems, and *directability* if these make errors, are important principles for safety.

Where machine learning is being used in civil uses of autonomous systems such as autonomous cars, a hybrid of machine learning, handcrafted rule sets and 'fail-safes' are used. Even then, formally verifying the behaviour of machine learning subcomponents is a risk mitigation challenge. For instance, the autonomous vehicle project at Google has completed more than two million kilometres of road testing in real driving conditions, and Google has said that considerably more testing will be needed before it puts such vehicles into production.[40]

It is hard to envisage what an equivalent level of testing and fail-safes would look like for AWS making targeting and attack decisions based on machine learning. After all, AWS would be designed to kill, which could worsen the consequences of lethal accidents when they occur. The outcomes of AWS failures could include fratricide, civilian casualties, or unintended escalation in a crisis as machines pursue emergent yet inexplicable goals such as area denial (to friendlies as well as hostiles) that could have strategic consequences.[41]

It also seems to have been widely assumed that developments in autonomy ultimately depend on algorithms expressed in software code that can be inspected and altered by the human designers of the system in question—and that this constitutes a robust safeguard. However, it is not clear yet to what extent this safeguard can effectively prevent hidden interactions in highly autonomous systems, particularly those employing self-optimization routines.

3. Safety is always one of a number of competing objectives

One influential perspective on managing the safety of hazardous technologies in complex organizations like militaries is so-called high reliability theory 'whose proponents argue that extremely safe operations are possible, even with extremely hazardous technologies, if appropriate organizational design and management techniques are followed.'[42] (See Table 1 for its characteristics.) High reliability organizations such as aircraft carrier crews, nuclear power plants, and air traffic control centres—which have less than their fair share of accidents—recognize that averting errors requires reducing human variability.[43] These organizations are constantly preoccupied with the possibility of failures that could become catastrophic.

40 A. Davies, 'Google's self-driving cars aren't as good as humans—yet', *Wired*, 12 January 2016: www.wired.com/2016/01/google-autonomous-vehicles-human-intervention. For latest information on Google's tests, including miles driven autonomously, see Google Self-Driving Car Project: www.google.com/selfdrivingcar/reports.

41 See presentation by P. Scharre, 'Flash War: Autonomous weapons and strategic stability', at UNIDIR event "Understanding Different Types of Risk", Geneva, 11 April 2016: http://bit.ly/UNIDIR_autonomy.

42 S. D. Sagan, *The Limits of Safety: Organizations, Accidents, and Nuclear Weapons*, Princeton University Press, 2013, p. 13.

43 J. Reason, 'Human error: models and management', *British Medical Journal,* vol. 320, 2000, pp. 768–770.

However, others have characterized high reliability theory as an optimistic view. In reality, safety is just one of a number of priorities for large, complex organizations. Nuclear weapon early warning and launch systems, for instance, represent a compromise between making sure that arsenals are kept safe and secure, and ensuring they are usable and ready to launch at very short notice if needed. In practice, designers and operators of AWS are likely to face analogous dilemmas between ensuring such systems do not cause accidents, and giving these systems the level of autonomy intended to convey military advantage. A military commander, for instance, might feel that operational signals silence is more important in a given situation (say, to ensure tactical surprise against an adversary) than an operator being able to monitor the processes and behaviour of an AWS in the field in order to shut it down or override it if it begins to act unsafely. Nor would a military commander presumably want 'friendly' AWS acting too predictably, if it makes it easier for the adversary to counter their capabilities.

'Normal accidents theory' (see Table 1) presents a much more pessimistic prediction of safety—that 'serious accidents with complex high technology systems are inevitable'.[44] Despite the best efforts of high reliability organizations, the world has come within a whisker of the use of nuclear weapons on more than one occasion, for instance, with only the intuition and initiative of individual human beings avoiding disaster.[45]

Table 1. Competing perspectives on safety with hazardous technologies

High Reliability Theory	Normal Accidents Theory
Accidents can be prevented through good organizational design and management	Accidents are inevitable in complex and tightly coupled systems
Safety is the priority organizational objective	Safety is one of a number of competing objectives
Redundancy enhances safety: duplication and overlap can make 'a reliable system out of unreliable parts'	Redundancy often causes accidents: it increases interactive complexity and opaqueness and encourages risk-taking
Decentralized decision-making is needed to permit prompt and flexible field-level responses to surprises	Organization contradiction: decentralization is needed for complexity, but centralization is needed for tightly coupled systems
A 'culture of reliability' will enhance safety by encouraging uniform and appropriate responses by field-level operators	A military model of intense discipline, socialization, and isolation is incompatible with democratic values
Continuous operations, training, and simulations can create and maintain high reliability operations	Organizations cannot train for unimagined, highly dangerous, or politically unpalatable operations
Trial and error learning from accidents can be effective, and can be supplemented by anticipation and simulations	Denial of responsibility, faulty reporting, and reconstruction of history cripples learning efforts

Reproduced from Sagan, *The Limits of Safety*, p. 46.

44 Sagan, *The Limits of Safety*, p. 13.

45 See, for instance, E. Schlosser, *Command and Control*, Allen Lane, 2013, and P. Lewis et al, *Too Close for Comfort*: *Cases of near nuclear use and options for policy*, Chatham House, 2014.

AWS pose the question of whether established techniques for achieving high reliability standards of operation in complex human organizations are meaningful goals for systems that by their very description will operate largely or almost entirely separate from direct human control or oversight. In this context, a 'culture of reliability' among human operators may not be enough. Low observability and complexity of the internal processes of AWS may undermine predictable operation in which accidents can be reliably avoided.

A related point is that the adversary is actively conspiring against success in ways that may be unanticipated. This would plausibly extend to active countermeasures against the other side's AWS, such as jamming communication links. These measures and their consequences may undermine safety, or at least commitment to safety as a prime goal (for instance, due to 'military necessity').

4. Redundancy is a cause of accidents

Redundancies are features added to a system that are intended by the designer to compensate for failures in the system's components. (A simple illustration of this is a reserve parachute, carried by a skydiver in case their main parachute fails.) Redundancies might also be added to a system to try to compensate for errors or shortcomings of human operators.

Perhaps counterintuitively, redundancy has been shown to be a cause of accidents in the management of hazardous technologies because it increases interactive complexity and opaqueness, and the chance of hidden interactions. With respect to the system, more 'moving parts' means more things to fail, or interact with other aspects of the system in unanticipated ways. This can make it more difficult for humans operating the system to gain an accurate picture of what is going on. Programmable autopilot systems, which have many benefits in aviation, are a case in point. Historically, accidents involving autopilots have occurred—even when the system was functioning correctly—due to a variety of reasons, including interface problems, distracted operators, or human pilots possessing an inaccurate mental picture of what the system is doing.

In this respect, developers and operators of high-technology systems can fall into a trap called the *substitution myth*—that adding automation will substitute for human input or failings. In fact, it usually changes the nature of the required human input, and can instead place higher demands on humans in the system in order to ensure that 'automation surprises' such as accidents do not occur. This is especially a problem when three factors converge:

1. 'Automated systems act on their own without immediately preceding directions from their human partner
2. Gaps occur in user's models of how their machine partners work in different situations
3. Weak feedback is given about the activities and future behaviour of the agent relative to the state of the world'.[46]

This is obviously relevant to the discussion about autonomy in weapon systems. In the CCW, for example, there has been discussion about humans 'in the loop' as a form of redundancy for AWS, or—conversely—machine autonomy as a form of redundancy for human decision-making. This may not reliably be the case because for all three factors listed above, humans are likely

46 D. Woods, 'Automation surprises' in D. Woods and E. Hollnagel, *Joint Cognitive Systems: Patterns in Cognitive Systems Engineering*, CRC Press, 2006, pp. 113–142, p. 121.

to have a highly incomplete picture of what the systems they are operating are doing. Such low observability in autonomous systems could thus increase the risk of accidents, not least because what cannot be observed and understood is not *directable*. Accidents could occur simply because operators cannot diagnose and respond effectively to 'automation surprises'[47] as quickly as is necessary to prevent an unsafe deviation from the intended activity.

Redundancy can also encourage risk taking by operators. According to risk compensation theory, people typically adjust their behaviour in response to a perceived level of risk, and so could be lulled into over-complacency by the autonomous features of a system. Indeed, *automation bias* (complacency or over-reliance on automated systems) has been a cause of accidents across a range of fields from aviation to clinical decision support systems used in medicine.[48] In observing that 'humans have a tendency to disregard or not search for contradictory information in light of a computer-generated solution that is accepted as correct and be exacerbated in a time critical domain', experts argue designers of 'intelligent' decision aids 'must be mindful that higher levels of automation combined with unreliable systems can actually cause new errors in system operation if not designed with human cognitive limitations and biases in mind.'[49]

5. Development and operation of autonomous systems is not value- or bias-free

Even though machine systems might behave, for practical intents and purposes, independently of humans in some contexts, they ultimately derive from human policy and design choices. Those choices, which stem from various factors including the values of designers and policy makers, in turn affect the ability of humans overseeing or operating autonomous systems to foresee or prevent accidents from occurring. Badly designed aspects of systems such as inefficient human-machine interfaces or inappropriate uses for highly autonomous technology exacerbate the risk of accidents.

Design and values

Design and testing are rarely linear in the development of any technical system. Designers always have to design relative to constraints of various kinds (e.g. cost, time, available technology) in trying to respond to a specified goal or function of a system. Moreover, there are almost always many possible designs. However, the choices of the designer also reflect factors such as their values and those of the socio-technical system they design within (for example, what level of risk is acceptable in a given design? How much can be left to chance in terms of validating that the design works as intended—and *only* as intended?). An important related question is what designers and operators *do not* want a system to do, and which values this reflects. Ultimately, the choices of designers are based on a variety of factors, including their values, biases and assumptions about the role autonomy plays.

In general, designers find it undesirable to have machines acting contrary to their values. However, as ethicists and computer scientists have discovered, it is difficult to define human values and translate them into the design of machine learning systems because often these

47 See Woods, 'Automation surprises'.

48 K. Goddard, A. Roudsari, and J.C. Wyatt, 'Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators', *Journal of the American Medical Informatics Association*, vol. 19, no. 1, 2012, pp. 121–127.

49 M. Cummings, 'Automation Bias in Intelligent Time Critical Decision Support Systems', *American Institute of Aeronautics and Astronautics 1st Intelligent Systems Technical Conference*, 2004, p. 1 and p. 5: wayback.archive.org/web/20141101113133/http://web.mit.edu/aeroastro/labs/halab/papers/CummingsAIAAbias.pdf.

depend on an understanding of context that is beyond the relatively 'narrow' intelligence these systems have. This issue becomes especially acute in autonomous systems because, to a greater or lesser extent, they may be involved in their own proximate goal setting.

In this regard AI researchers recently highlighted five problems or possible failure modes that designers should take account of in applying machine intelligence-using AI for use in real world circumstances. To illustrate these problems, they used the simple hypothetical example of a cleaning robot:

- '**Avoiding Negative Side Effects:** How can we ensure that our cleaning robot will not disturb the environment in negative ways while pursuing its goals, e.g. by knocking over a vase because it can clean faster by doing so? Can we do this without manually specifying everything the robot should not disturb?
- **Avoiding Reward Hacking:** How can we ensure that the cleaning robot won't game its reward function? For example, if we reward the robot for achieving an environment free of messes, it might disable its vision so that it won't find any messes, or cover over messes with materials it can't see through, or simply hide when humans are around so they can't tell it about new types of messes.
- **Scalable Oversight:** How can we efficiently ensure that the cleaning robot respects aspects of the objective that are too expensive to be frequently evaluated during training? For instance, it should throw out things that are unlikely to belong to anyone, but put aside things that might belong to someone (it should handle stray candy wrappers differently from stray cellphones). Asking the humans involved whether they lost anything can serve as a check on this, but this check might have to be relatively infrequent—can the robot find a way to do the right thing despite limited information?
- **Safe Exploration:** How do we ensure that the cleaning robot doesn't make exploratory moves with very bad repercussions? For example, the robot should experiment with mopping strategies, but putting a wet mop in an electrical outlet is a very bad idea.
- **Robustness to Distributional Shift:** How do we ensure that the cleaning robot recognizes, and behaves robustly, when in an environment different from its training environment? For example, heuristics it learned for cleaning factory workfloors may be outright dangerous in an office.'[50]

These safety problems are generalizable to a variety of 'learning' autonomous systems. It is easy to see how this would extend to AWS. For example, how could designers ensure that an AWS recognizes an environment change, say from one in which it may target any unidentified moving vehicle, to one in which it may not, for instance because an ambulance or a bus filled with civilians is present (robustness to distributional shift)? Could designers ensure that in 'training' a system to, say, recognize tanks in target images, they do not inadvertently train a system that distinguishes targets based on some extraneous factor, like the difference in light conditions on cloudy days from sunny ones[51] (avoiding reward hacking)?

50 D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman and D. Mané, 'Concrete Problems in AI Safety', p. 3.

51 Yudkowsky used this (possibly apocryphal) story of a machine learning system called a neural network, although actually to illustrate what he called 'the context problem', which relates to distributional shift: 'This form of failure is especially dangerous because it will *appear* to work within a fixed context, then fail when the context changes.' Yudkowsky, E., 'Artificial intelligence as a positive and negative factor in global risk', in N. Bostrom and M. Ćirković (eds.), *Global Catastrophic Risks*, Oxford University Press, 2008, pp. 308–345, p. 321.

Operators and their limits

Proponents of joint cognitive systems (JCS) often point to the value of human beings as 'fail safes' in human-machine systems due to our capacity for contextual thinking, which among other things is often vital for moral decision-making. Machines can process information logically and, by human standards, very rapidly. But the ability of human beings to see the big picture, so the thinking goes, acts as an important safeguard against unintended behaviours.

In this way, some experts and policy makers feel that—for a variety of reasons—it makes most sense to see AWS as parts of human-machine teams rather than as a distinct category. 'Like a centaur, the hybrid would have the strength of each of its components: the processing power of a large logic circuit and the intuition of a human brain's wetware.'[52] Proponents of this approach note that autonomy is relative, and argue it is unlikely, at least for the foreseeable future, that machines will have the capacity for targeting and attack totally divorced from human decision making and accountability.[53] JCS must be designed that ensure humans are not left out of the loop of decision making, and can intervene to avert unintended behaviour from the system, if necessary.[54]

Centaur-like human-machine teams are already becoming a feature of systems of various kinds that have highly sophisticated levels of autonomy. However, the historical evidence for their effectiveness is not wholly favourable. Bad surprises including accidents involving automated systems such as aircraft autopilots, for instance, tend to be caused not by the behaviour of the machine or the human operator per se, but due to the ways in which they relate to one another, leading one expert to conclude:

> Our fascination with the possibilities afforded by new technological powers often obscures the fact that new computerized and automated devices also create new burdens and complexities for the individuals and teams of practitioners responsible for operating, trouble shooting, and managing high-consequence systems.[55]

The vaunted capacity of human beings for contextual thinking is not without constraint, which has implications both for human designers and operators. All human beings are subject to a range of cognitive and perceptual biases that affect our decision-making. Automation bias has already been mentioned. However, some other common, relevant biases or limits include:

- **Simplification**: humans have an inherent tendency to simplify phenomena they encounter. This can have an influence on the accurate assessment of risk.
- **Risk perception bias**: people tend to be notoriously poor at accurate risk estimation, particularly if the phenomena or situation they encounter is a novel one. (Actual experience—particularly long experience—can improve judgement of risk.)
- **Confirmation bias**: the tendency to look for evidence that confirms a hypothesis, while failing to look for evidence that would disconfirm it.

---

52 W. Isaacson, 'Brain gain: Review of "Smarter Than You Think" by Clive Thompson', *New York Times*, 1 November 2013: www.nytimes.com/2013/11/03/books/review/smarter-than-you-think-by-clive-thompson.html.

53 For instance, see United States DoD, *Report of the Defense Science Board Summer Study on Autonomy*.

54 For a thorough analysis, see P. Scharre, *Autonomous Weapons and Operational Risk*, (Ethical Autonomy Project) Center for a New American Security, 2016: www.cnas.org/publications/reports/autonomous-weapons-and-operational-risk.

55 Woods, 'Automation surprises', p. 119.

- **Projection bias**: projecting onto others what our own, current preference is. Projection bias also can influence our assessment of the preferences of our own future selves.
- **Illusion of explanatory depth:** we all think we understand how systems work better than we actually do. Much of our understanding of cause and effect is based on association, without an understanding of how events are related to one another. This may lead to serious mistakes in understanding the reasons for AWS behaviour and trying to predict how AWS will behave in the future.

Human cognitive limits are challenging to overcome through design. And, indeed, the whole point of increasing autonomy is to transfer decision-making responsibilities away from humans in the system to the degree possible. It means that having humans 'in the loop' is not a solution in itself to safety-related risks in autonomous systems. For example, Wallach observed that one issue in the development and use of remotely piloted UAVs using a JCS approach for safety is that the task of adapting to the unexpected is presumed to lie with the human member(s) of the team. Similar problems could be seen with AWS:

> Anticipating the actions of a smart system becomes more and more challenging for a human operator as the system and the environments in which it operates become more complex. Expecting operators to understand how a sophisticated computer thinks, and anticipate its actions so as to coordinate the activities of a team, actually increases their responsibility.[56]

Beside the practical hazard this kind of practice can create, it is also raises the question of whether human beings should be placed in situations in which they are responsible for the behaviour of armed machines they might not be able to understand or fully control. Even if humans are held responsible for developing AWS, it is unclear under what conditions the responsibility for succeeding events transfers to those agents. An observation about the development of cyber warfare also holds for AWS, and is sobering: 'If humans were rational in their planning and decisions, then assigning responsibility to originating agents might be reasonable, but neither people nor groups exhibit the necessary rationality or self-knowledge. In particular, people are frequently unable to accurately predict their future preferences, goals, and values.'[57]

Policy makers and responsibility

Human operators are themselves constrained in significant ways that may compound rather than reduce the risk of accidents in JCS. This is not simply a problem of design and operation—it is a ultimately a broader problem of responsibility that falls on the shoulders of those considering policies that allow for AWS:

> 'A responsibility gap will not arise merely from the technological complexity of artificial agents. Artificial agents can be designed so that no human can understand or control what they do or they can be designed so that they are transparent and well within human control or they can be designed so that certain aspects or levels of the machine behavior are in human control and others are not. Which way they are designed depends on the humans involved in their development and acceptance.

56 W. Wallach, *A Dangerous Master: How to Keep Technology From Slipping Beyond Our Control*, Basic Books, 2015, p. 32.

57 D. Danks and J.H. Danks, 'Beyond Machines: Humans in Cyber Operations, Espionage and Conflict', in F. Allhoff, A. Henschke, and B.J. Strawser (eds), *Binary Bullets: The Ethics of Cyberwarfare*, Oxford University Press, 2016, Chapter 9, pp. 177–200, p. 27 of pre-print version: pdfs.semanticscholar.org/dc13/df07adca23157de149cf80f66d85381a6ee9.pdf.

> In the past people have chosen technologies that have some degree of risk though we have also set up mechanisms to pressure those who make and use these technologies to operate them safely and to take responsibility when something goes wrong and the fault can be traced back to them. The future may be different, but it seems there are good reasons why we might resist any future in which no humans are responsible for technologies that have a powerful role in our lives.[58]

Among the reasons for resisting such a future, quite arbitrary decisions would probably have to be made in terms of assigning responsibility for the behaviour of AWS—behaviour based on processes that could be low in observability and directability. Thus, rather than vaunted human-machine teams, JCS could create situations in which human designers and operators of AWS are, in effect, designated scapegoats for systems they cannot reasonably understand, let alone control. If these people would be little more than hostages to fortune then it is not much of a system of responsibility, and is also not going to contribute to reducing and preventing accidents involving AWS in the longer run.

Concluding thoughts

As noted earlier, policy debate until now in forums like the CCW has tended to revolve around how to apportion legal responsibility in situations in which, for instance, AWS violate IHL rules. However, it is clear that AWS pose broader responsibility challenges, which need to be considered alongside those specifically concerning targeting and attack in conflict, due to the nature of these systems, and the prospect they might cause unintended harm to humans in a broader range of situations. Forums like the CCW do not seem to have grappled with this yet. Yet, the prospect of armed autonomous systems prone to accidents must surely be an alarming one, including for militaries contemplating the roles for autonomous systems in their future operations.

So, what can be done for dealing with these uncertainties created by the development of AWS? One option would be for the precautionary principle to be applied, and for AWS systems to be prohibited on the basis that their risks—including the risk of accidents—are too great for their introduction to be considered. This might be a legally binding prohibition, or be achieved by means of an international moratorium until the international community agrees on clear standards for determining how accidents would be prevented, and how agreed standards would be verified. Admittedly, the prospects for this seem dim at the present time, but a general-purpose criterion like those in the Biological and Chemical Weapons Conventions might allow for the development of schedules that could be changed and updated. As a basic principle of such a system, it would need to put the onus on those wishing to introduce AWS, however defined, to demonstrate their safety in order to be effective.

A second alternative, suggested by some legal scholars, would be the creation of a framework for effective technical standards and procedures for verification and evaluation of AWS, in addition to protocols to guide risk assessment, end-user information and training, and what they describe as ‘fail-safe’ measures. However, the likelihood of the development of such a framework within the ambit of Article 36 weapons review is unclear, at best. Also, while certainly vital to explore, the possibility of fail-safes may prove to be a pipe dream, including for reasons outlined in this report. Nevertheless, these scholars put their finger on an important point when they conclude that the risks AWS would pose ‘are diverse, and they demand an approach that adequately

58 D.G. Johnson, ‘Technology with No Human Responsibility?’, p. 714.

identifies and controls the risk arising from behavioural uncertainty and generates information that can be used to better evaluate, manage and prevent the risk of unlawful behaviour.'[59]

Short of this, policy makers in the CCW and more broadly could factor consideration of inadvertent risks, especially accidents, into their ongoing talks on issues around AWS. This might not solve anything, but it would at least sensitize national decision makers and experts to the major points of concern. Concepts like 'appropriate human judgement' or 'meaningful human control' are entry points for discussion. At the moment, the way these concepts are talked about in the CCW mostly reflects a nest of assumptions about current IHL and its sufficiency. But questions about appropriate judgement, meaningful control, and responsibility logically extend to safety and accidents. Basic questions to be answered are these: who would the international community hold accountable for permitting the introduction of armed machine systems that are inherently unpredictable and prone to accidents with lethal consequences? And how would this be done?

To date, some national policy makers in the CCW have preferred to 'wait and see' before committing themselves to a particular view on the acceptability of AWS. No doubt some would like more complete information to emerge before reaching a position. At the same time, there appears to be a view held by some that an AWS arms race is basically inevitable, driven by technological developments that offer irresistible new capabilities. In this respect, it is important to underline that the development of technology is shaped by social forces, not just the other way around. Some rationalists have argued that widespread public revulsion (the 'ick factor') about AWS can and should be ignored. But this flies in the face of significant evidence that emotions are an indispensable normative guide in judging the moral acceptability of technological risks.[60] Hesitancy to do anything in the absence of a definitive picture of the shape of things to come on AWS may become a self-fulfilling prophecy—bringing about unanticipated and harmful consequences that policy makers could have avoided.

In closing, this paper has sought to reflect on the potential for AWS to fail in ways that are unanticipated and harmful to humans. This is a broader set of scenarios than simply those in which IHL applies. Of course, any hazardous technology carries 'unintentional' risk, and can have harmful results its designers and operators did not intend. AWS may pose novel, unintended forms of hazard to human life that typical approaches to ensuring responsibility do not prevent because these systems act in complex and unpredictable ways. Suggested paradigms in which human-machine teams are the answer to ensuring unintended harm from AWS would, on their own, likely be insufficient in preventing lethal accidents. Preventing accidents is not solely a question of design, human supervision or operation of AWS: it is only possible if policy makers make responsible choices about what are acceptable boundaries for development and use.

59 N. Bhuta and S. Pantazopoulos, 'Autonomy and uncertainty: increasingly autonomous weapon systems and the international legal regulation of risk', in N. Bhuta et al, *Autonomous Weapons Systems: Law, Ethics, Policy*, Cambridge University Press, 2016, p. 299.

60 S. Roeser, 'The role of emotions in judging the moral acceptability of risks', *Safety Science*, vol. 44, 2006, pp. 689–700.

UNIDIR

Safety, Unintentional Risk and Accidents in the Weaponization of Increasingly Autonomous Technologies

Recent international attention on autonomous weapon systems (AWS) has focused on the implications of what amounts to a 'responsibility gap' in machine targeting and attack in war. As important as this is, the full scope for accidents created by the development and deployment of such systems is not captured in this debate. It is necessary to reflect on the potential for AWS to fail in ways that are unanticipated and harmful to humans—a broader set of scenarios that simply those in which international humanitarian law applies.

Of course, any complex, hazardous technology carries 'unintentional' risk, and can have harmful results its designers and operators did not intend. AWS may pose novel, unintended forms of hazard to human life that typical approaches to ensuring responsibility do not effectively manage because these systems may behave in unpredictable ways that are difficult to prevent. Among other things, this paper suggests human-machine teams would, on their own, be insufficient in ensuring unintended harm from AWS is prevented, something that should bear on discussions about the acceptability of deploying these systems. This is the fifth in a series of UNIDIR papers on the weaponization of increasingly autonomous technologies.