The United Nations Institute for Disarmament Research

Phone +41 (0)22 917 34 28
Fax +41 (0)22 917 0176
jborrie@unog.ch
Palais des Nations
CH-1211 Geneva 10
Switzerland

www.unidir.org

UNIDIR

# SECURITY, UNINTENTIONAL RISK, AND SYSTEM ACCIDENTS

**John Borrie, Chief of Research, UNIDIR**

Geneva, 15 April 2016

Thank you very much for inviting me to speak on this panel today.

When policy practitioners consider security and the increasing autonomization of weapon systems (AWS), the natural preoccupation is with intentional actions and their consequences.

At the level of humanitarian and human rights law, for instance, a major concern is with accountability if machine systems are enabled to select and attack targets without a human decision in each individual attack.[1]

At the level of non-proliferation and arms control, some experts are concerned that developments in autonomy will lead to arms races (that is, vertical proliferation) and horizontal proliferation—including to non-state armed groups.

Moreover, Paul Scharre explained in a UNIDIR side-event earlier this week how autonomous weapon systems might be causes both of strategic stability and instability in crisis situations.[2] The use or threatened use of lethal autonomous systems might conceivably deter if it convinces an adversary that one's hands are tied in advance. It might allow a user to move up and down the escalation ladder more reliably and more visibly. But autonomy could also compound strategic fragility, for example due to inadvertent risk of various kinds.[3]

It's this subject of inadvertent risk that I want to explore briefly now because it's under-explored in the context of security.[4]

- Risk, simply stated, is the possibility of some bad event happening, something commonly quantified as the probability of an event multiplied by its consequences.[5]
- By unintentional risk I mean a subset of total risk in which machine systems that have targeting and attack functions fail to behave in ways intended—or necessarily even predicted—by their designers and operators. There are many potential causes of inadvertent risk, and thus firm reason to believe such failures would be a question of 'when' and not 'if' such systems were to be developed and deployed.[6]

**Inadvertent risk**

Let's consider some examples of potential failure in autonomous weapon systems that would contribute to inadvertent risk. We can categorize these possibilities in various ways. Last week, at a UNIDIR meeting of experts that I moderated, we came up with several approaches. Due to constraints on time, I've displayed just one here (in Figure 1), with examples displayed in red. In due course, UNIDIR will publish an observation paper that explores these and related issues in greater depth.

Now the diagram doesn't presume that a given system would necessarily fail in these ways. One thing the diagram displays well, I think, is that causes of failure may reflect interactions with the user or operator, as well as the overall context including the environment, the behaviour of adversaries and friendly forces, as well as the socio-technical system in which any technology emerges and is used. Indeed, a key point I want to convey to you is that any of these many factors might interact with any other—or others—to compound the initial failure. So when we talk about autonomous 'systems', it's really not just the machine or the code itself that's solely relevant.

In some cases, failures might have negligible consequences. In others, these sources of inadvertent risk could undermine the safe design and operation of AWS to the extent that they conceivably cause mass lethality accidents, or/and the creation of crises that undermine strategic stability.

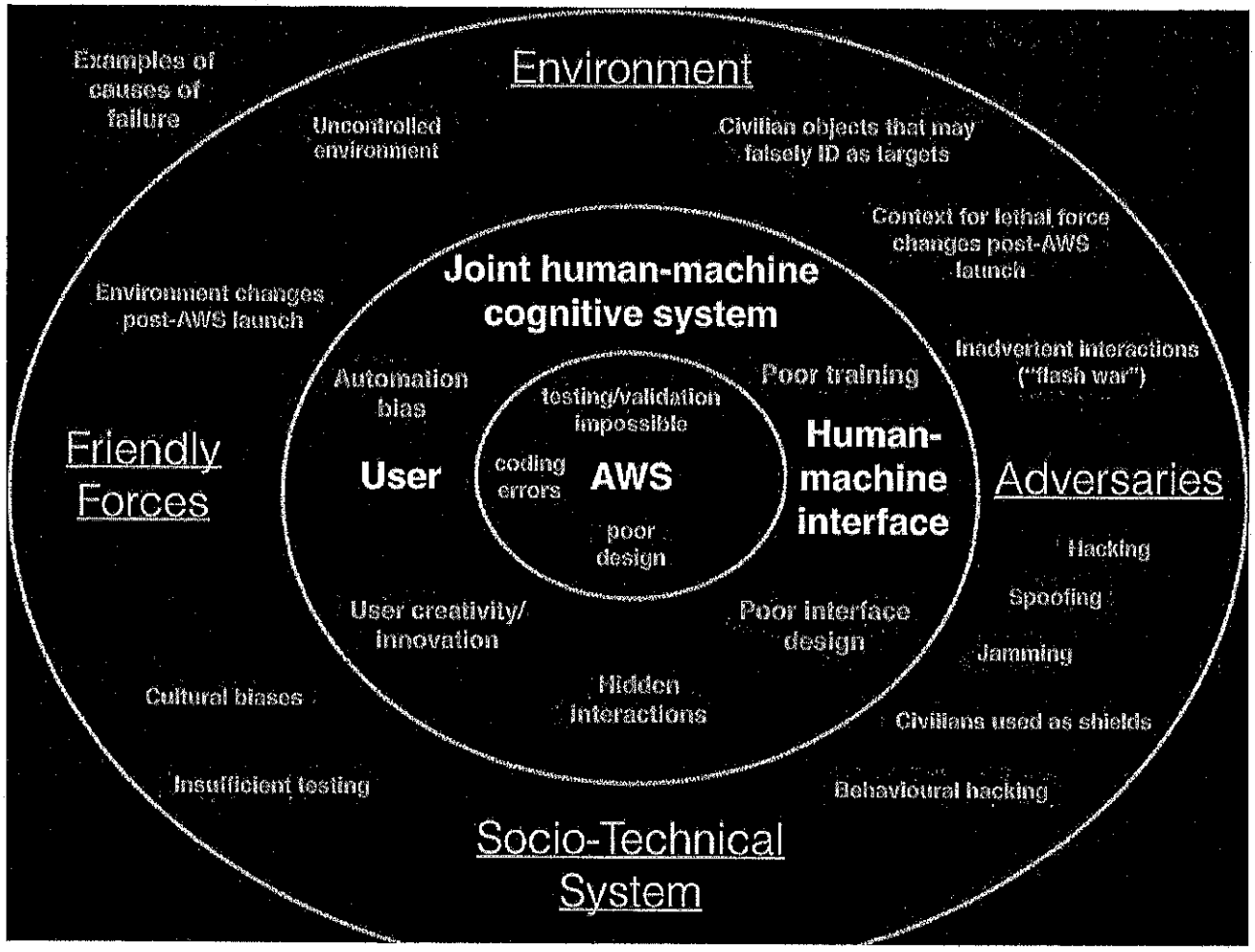


Figure 1: Examples of cause of failure in AWS (courtesy of UNIDIR expert group on technology, safety and unintentional risk, 8 April 2016)

**System accidents**

Moreover, so-called 'system accidents' are a special phenomenon to consider. I think these are very important to your discussions, as they constitute a source of risk that can't be entirely 'designed out' or quashed through rigorous operational standards.

I spoke on Monday at some length about system accidents at UNIDIR's side-event on risk, and I don't intend to do so again now. (You'll be able to find that presentation online at UNIDIR's website shortly.[7]) But here are a few points derived from the work of experts such as Charles Perrow[8] and Scott Sagan[9] working in other areas that I would suggest are helpful in starting to think about the inadvertent risks of increasingly autonomous weapon systems:

1. **Accidents are inevitable in complex and tightly coupled systems.** In that vein, the experts that UNIDIR gathered last week concluded that autonomous weapon systems are likely to be both highly complex and very tightly coupled. In complex systems there are a lot of

common mode connections between components, which means it's not necessarily clear which of them has gone wrong when there's a failure. And there are many feedback loops. Tight coupling means that the consequences of a change or failure in one part of the system will rapidly propagate—perhaps quicker than human operators can respond.

2. **Safety is always one of a number of competing objectives (not least in war).**
3. **Redundancy is often a cause of accidents: it increases interactive complexity and opaqueness. It also can encourage risk taking.** This is relevant to the discussion about autonomy, because in the CCW you often seem to be talking about humans 'in the loop' as a form of ultimate redundancy, or machines as a form of redundancy for human decision-making. This may not reliably be the case.
4. **The development of autonomous systems won't be value free:** the preferences, biases and assumptions of designers will shape machine systems even if these are intended to be highly predictable, reliable and autonomous. It means interactions with catastrophic risk potential may in practice be hidden from designers and operators.
5. **Hidden interactions might have particular risk potential in systems reliant on machine learning.** These processes can't necessarily be easily inspected or tested, but would be highly attractive in some military contexts in which, for instance, machines may have to function *incommunicado* from human decision makers for extended periods.[10] Problems such learning autonomous systems encounter in interpreting context could lead to highly unorthodox outputs.[11] These might include the potential for unintended mass lethality (if they are armed and permitted to attack in at least some circumstances), or the pursuit of emergent yet inexplicable goals such as area denial (to friendlies, as well as hostiles) that might have strategic consequences.

**Final thoughts**

In sum, the propensity of complex, tightly coupled systems for 'system accidents' is relevant to security because these could undermine strategic stability in ways unintended by operators.

If 'normal accident' theory holds, then system accidents are a cause of inadvertent risk and unpredictability that cannot be eradicated. Of course, this is the case with many contemporary hazardous technologies from nuclear

power to nuclear weapon control systems and manned spaceflight. Accidents may be rare, but they do occur. And when these complex, tightly coupled systems fail they tend to do so dramatically, if not catastrophically.

What arguably makes our problem different and even more challenging to consider is this: *failures with direct lethal consequences* could result from the decision making of systems that are not human cognitive and moral agents *in addition* to resembling the kinds of complex, tightly coupled system we already struggle with in safety terms.

Machine learning further suggests that the potential for hidden interactions in those systems would be greater—not lesser—because we could not necessarily predict in advance the processes by which these sense and interpret the world, and seek to achieve their goals within it.

All of this also implies that there is a safety aspect to security, as well as to 'meaningful human control' or whatever we wish to call it. Among other things, the nature of systems with autonomous targeting and attack functions prompts the question of whether the requisite level of predictability, system transparency, training and testing in order to allow the reassertion of human control before catastrophic consequences occur is realistically achievable—now or in the future. This should be part of our debate about security.

I thank you for your time.

---

[1] For example, see Heyns, C. (A/HRC/23/47) *Report of the Special Rapporteur on extrajudicial summary or arbitrary executions*, Geneva, United Nations Human Rights Council, 2013. Online: http://www.ohchr.org/Documents/HRBodies/HRCouncil/RegularSession/Session23/A-HRC-23-47_en.pdf. See also UNIDIR, *Framing discussions on the weaponization of increasingly autonomous technologies*, (No. 1), Geneva, 2014, p. 2.

[2] Scharre P. 'Center for a New American Security Presentation on 'Flash War': Autonomous Weapons and Strategic Stability', UNIDIR CCW lunchtime side-event on *The Weaponization of Increasingly Autonomous Technologies: Understanding Different Types of Risks*, Geneva, 11 April 2016. These and other presentations (including the author's) are to be found online at www.unidir.org.

[3] Strategic stability is an important lens through which to see security. However, it is important to note that strategic stability is, in itself, insufficient for security, even if it might be a necessary precondition in a multipolar world in which one major power cannot dominate the others.

[4] Although there are notable exceptions. In particular, see Scharre, P. (2016), *Autonomous Weapons and Operational Risk* (Ethical Autonomy Project), Washington D.C., Center for a New American Security. See also Borrie, J., 'Safety aspects of "meaningful human control": Catastrophic accidents in complex systems', New York, UNIDIR, 16 October 2014. Wendell Wallach's book *A Dangerous Master: How to Keep Technology from Slipping Beyond Our Control*, New York, Basic Books, 2015, is an important contribution that touches on these issues in the context of broader concerns about AWS.

[5] See European Commission (2010), *Risk Assessment and Mapping Guidelines for Disaster Management*, pp. 15-16.

[6] Issues around unintentional risk are to be explored further in a UNIDIR observation report on *Technology, Safety, and Unintentional Risk* (2016, forthcoming) following a UNIDIR meeting of experts on these topics from 7 to 8 April 2016 in Geneva.

[7] Borrie, J. "Unintentional Risks", UNIDIR CCW lunchtime side-event on *The Weaponization of Increasingly Autonomous Technologies: Understanding Different Types of Risks*, Geneva, 11 April 2016.

[8] Perrow, C. (1984), *Normal Accidents: Living With High-Risk Technologies* New York, Basic Books.

[9] Sagan, S. D. (2013), *The Limits of Safety: Organizations, Accidents, and Nuclear Weapons.* Princeton, New Jersey, Princeton University Press.

[10] For instance, in maritime environments. See UNIDIR, *Testing the Waters: The weaponization of increasingly autonomous technologies in the maritime environment,* (No. 4), Geneva, 2015. See also Hambling, D. (2016). *The Inescapable Net: Unmanned Systems in Anti-Submarine Warfare* (Parliamentary Briefings on Trident Renewal), London, BASIC.

[11] See, for instance, Woods D., 'Chapter 10: Automation surprises' in *Joint Cognitive Systems: Patterns in Cognitive Systems Engineering*, CRC Press, 2006, pp. 113-142. For more specifically on the 'context problem' see Yudkowsky, C., 'Artificial intelligence as a positive and negative factor in global risk', in Bostrom N. and Ćirković M. (eds.) *Global Catastrophic Risks* (Oxford, Oxford University Press, 2008, pp. 308-345, p. 321.