6 Reframing Autonomous Weapons Systems

Autonomous systems that are designed to cause physical harm have additional ethical ramifications as compared to both traditional weapons and autonomous systems that are not designed to cause harm. Multi-year discussions on international agreements around autonomous systems in the context of war are occurring at the UN, but professional ethics about such systems can and should have a higher standard covering a broader array of concerns.

Broadly, we recommend that technical organizations accept that meaningful human control of weapons systems is beneficial to society, that audit trails guaranteeing accountability ensure such control, that those creating these technologies understand the implications of their work, and that professional ethical codes appropriately address works that are intended to cause harm.

Specifically, we would like to ensure that stakeholders are working with sensible and comprehensive shared definitions of concepts relevant in the space of autonomous weapons systems (AWS). We recommend designers not only take stands to ensure meaningful human control, but be proactive about providing quality situational awareness through those autonomous or semi-autonomous systems to the humans using those systems. Stakeholders must recognize that the chains of accountability backward, and predictability forward, also include technical aspects such as verification and validation of systems, as well as interpretability and explainability of the automated decision-making, both in the moment and after the fact.

A concern is that professional ethical codes should be informed by not only the law but an understanding of both direct and macro-level ramifications of products and solutions developed explicitly as, or that can be expected to be used or abused as, AWS. Some types of AWS are particularly societally dangerous because they are too small, insidious, or obfuscated to be attributable to the deploying entity, and so ethical recommendations are needed to prevent these instances from having dangerous outcomes.

6 Reframing Autonomous Weapons Systems

Issue:

Professional organization codes of conduct often have significant loopholes, whereby they overlook holding members' works, the artifacts and agents they create, to the same values and standards that the members themselves are held to, to the extent that those works can be.

Background

Many professional organizations have codes of conduct intended to align individuals' behaviors toward particular values; however, they seldom sufficiently address members' behaviors in contributing toward particular artifacts, such as creating technological innovations deemed threatening to humanity, especially when those innovations have significant probabilities of costly outcomes to people and society. Foremost among these in our view are technologies related to the design, development, and engineering of AWS.

Candidate Recommendations

- We propose that any code of conduct be extended to govern a member's choice to create or contribute to the creation of technological innovations that are deemed threatening to humanity. Such technologies carry with them a significant probability of costly outcomes to people and society. When codes of conduct are directed towards ensuring positive benefits or outcomes for humanity, organizations should ensure that members do not create technologies that undermine or negate such benefits. In cases where created technologies or artifacts fail to embody or conflict with the values espoused in a code of conduct, it is imperative that professional organizations extend their codes of conduct to govern these instances so members have established recourse to address their individual concerns. We also recommend that codes of conduct more broadly ensure that the artifacts and agents offered into the world by members actively reflect the professional organization's standards of professional ethics.
- Professional organizations need to have resources for their members to make inquiries concerning whether a member's work contravenes International Humanitarian Law or International Human Rights Law.

Further Resources

- Kvalnes, Øyvind. "Loophole Ethics," in *Moral Reasoning at Work: Rethinking Ethics in Organizations*, 55–61. Palgrave Macmillan U.K., 2015.
- Noorman, Merel. "Computing and Moral Responsibility," *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), Summer 2014 Edition.
- Hennessey, Meghan. "ClearPath Robotics Takes Stance Against 'Killer Robots'." ClearPath Robotics, 2014.
- "Autonomous Weapons: An Open Letter from AI & Robotics Researchers." Future of Life Institute, 2015.

6 Reframing Autonomous Weapons Systems

Issue:

Confusions about definitions regarding important concepts in artificial intelligence (AI), autonomous systems (AS), and autonomous weapons systems (AWS) stymie more substantive discussions about crucial issues.

Background

The potential for confusion about definitions is not just an academic concern. The lack of clear definitions regarding AWS is often cited as a reason for not proceeding toward any kind of international control over autonomous weapons.

The term autonomy is important for understanding debates about AWS; yet there may be disputes—about what the term means and whether it is currently possible—that prevent progress in developing appropriate policies to guide its design and manufacture. We need consistent and standardized definitions to enable effective discussions of AWS, free from technological considerations that are likely to be quickly outdated. As this is both a humanitarian issue and an issue of geopolitical stability, the focus in this area needs to be on how the weapons are controlled by humans rather than about the weapons technology per se.

The phrase "in the loop" also requires similar clarification. Let us assume that an automatic weapons system requests permission to fire from a human operator, and the operator gives permission. How long of a delay should be acceptable between the system request and the operator's permission take place before the situation has changed to invalidate the permission? A sub-second clearance would probably be judged as acceptable in most cases, but what about multiple minutes? It could be argued that the situation itself should be examined, but that may result in either undue cognitive load on the operator at a critical time, or for the system itself to make decisions on what is "an appropriate level of change" and possibly retract its intent to fire.

What is often also unclear in these scenarios is whether clearance to fire at a target means a system is cleared to prosecute that target indefinitely, or has predetermined limits on the amount of time or ordinance each clearance provides.

In analyzing these issues, one quickly realizes that the type of autonomy that is of concern is no more complicated than the type of autonomy that we cede to chess programs. In both cases the human has not anticipated in advance and made an appropriate decision for every situation that can possibly arise. In many cases the machine's decision in these instances will be different from what the human's decision would have been.

This notion of autonomy can be applied separately to each of the many functions of a weapons system; thus, an automatic weapons system could be autonomous in searching

6 Reframing Autonomous Weapons Systems

for targets but not in choosing which ones to attack, or vice versa. It may or may not be given autonomy to fire in self-defense when the program determines that the platform is under attack, and so on. Within each of these categories, there are also many intermediate gradations in the way that human and machine decision making may be coupled.

Candidate Recommendations

- The term *autonomy* in the context of AWS should be understood and used in the restricted sense of delegation of decision-making capabilities to a machine. Since different functions within AWS may be delegated to varying extents, and the consequences of such delegation depend on the ability of human operators to forestall negative consequences via the decisions over which they retain effective control, it is important to be precise about the ways in which control is shared between human operators and AWS.
- We recommend that various authorization scenarios be further investigated for ethical best practices by a joint workshop of stakeholders and concerned parties (including, but not limited to, international humanitarian organizations and militaries), and that those best practices be promoted by professional organizations as well as by international law.

Further Resources

- Dworkin, Gerald. *The Theory and Practice of Autonomy*. Cambridge University Press, 1988.
- Frankfurt, Harry G. "Freedom of the Will and the Concept of a Person," in T*he Importance of What We Care About*, Cambridge University Press, 1987.
- DoD Defense Science Board, The Role of Autonomy in DoD Systems, Task Force Report, July 2012, 48.
- DoD Defense Science Board, Summer Study on Autonomy. June 2016.
- Young, Robert. Autonomy: *Beyond Negative and Positive Liberty*. St. Martin's Press, 1986.
- Society of Automotive Engineers standard J3016, Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems, 2014.
- Sheridan, T. B., and W. L. Verplank. Human and Computer Control of Undersea Teleoperators.. Cambridge, MA: Man-Machine Systems Laboratory, Department of Mechanical Engineering, Massachusetts Institute of Technology, 1978.
- Sharkey, Noel. "Towards a Principle for the Human Supervisory Control of Robot Weapons." *Politica and Società* 2 (2014): 305–324.

6 Reframing Autonomous Weapons Systems

Issue:
AWS are by default amenable to covert and non-attributable use.

Background

The lack of a clear owner of a given AWS incentivizes scalable covert or non-attributable uses of force by state and non-state actors. Such dynamics can easily lead to unaccountable violence and societal havoc.

Candidate Recommendation

Because AWS are delegated authority to use force in a particular situation, they are required to be attributable to the entity that deployed them through the use of physical external and internal markings as well as within their software.

Further Resources

- Bahr, Elizabeth. "Attribution of Biological Weapons Use," in *Encyclopedia of Bioterrorism Defense*. John Wiley & Sons, 2005.
- Mistral Solutions. "Close-In Covert Autonomous Disposable Aircraft (CICADA) for Homeland Security,"2014.
- Piore, Adam. "Rise of the Insect Drones." Wired. January 29, 2014.

Issue:
There are multiple ways in which accountability for AWS's actions can be compromised.

Background

Weapons may not have transparency, auditability, verification, or validation in their design or use. Various loci of accountability include those for commanders (e.g., what are the reasonable standards for commanders to utilize AWS?), and operators (e.g., what are the levels of understanding required by operators to have knowledge of the system state, operational context, and situational awareness?).

There are currently weapons systems in use that, once activated, automatically intercept high-speed inanimate objects such as incoming missiles, artillery shells, and mortar grenades. Examples include C-RAM, Phalanx, NBS Mantis, and Iron Dome. These systems complete their detection, evaluation, and response process within a matter of seconds and thus render it extremely difficult for human operators to exercise meaningful supervisory control once they have been activated other than deciding when to switch them off. This is called supervised autonomy by the US DoD[lxii] because the weapons require constant and vigilant human evaluation and monitoring for rapid shutdown in cases of targeting errors, change of situation, or change in status of targets.

6 Reframing Autonomous Weapons Systems

Candidate Recommendations

- Trusted user authentication logs and audit trail logs are necessary, in conjunction with meaningful human control. Thorough factors-driven design of user interface and human–computer/robot interaction design is necessary for situational awareness, knowability, understandability and interrogation of system goals, reasons and constraints, such that the user could be held culpable.
- Tamper-proof the equipment used to store authorization signals and base this on open, auditable designs, as suggested by Gubrud and Altmann (2013). Further, the hardware that implements the human-in-the-loop requirement should not be physically distinct from the operational hardware of the system, to deter the easy modification of the overall weapon after the fact to operate in fully autonomous mode.
- System engineers must have higher standards and regulations of security for system design from a cybersecurity perspective than they would for other computer-controlled weapons systems. AWS ought to be designed with cybersecurity in mind such that preventing tampering, or at least undetected tampering, is a highly weighted design constraint.

Further Resources

- Gubrud, M., and J. Altmann. "Compliance Measures for an Autonomous Weapons Convention." International Committee for Robot Arms Control, May 2013.
- The UK Approach to Unmanned Aircraft Systems (UAS), Joint Doctrine Note 2/11, March 30, 2011.
- Sharkey, Noel. "Towards a Principle for the Human Supervisory Control of Robot Weapons." *Politica and Società* 2 (2014): 305–324.
- Owens, D. "Figuring Forseeability." *Wake Forest Law Review* 44 (2009): 1277, 1281–1290.
- Scherer, Matt. "Who's to Blame (Part 4): Who's to Blame if an Autonomous Weapon Breaks the Law?" *Law and AI* (blog), February 24, 2016.

Issue:

An automated weapons system might not be predictable (depending upon its design and operational use). Learning systems compound the problem of predictable use.

Background

Modeling and simulation of AWS, particularly learning systems, may not capture all possible circumstances of use or situational interaction. They are underconstrained cyberphysical systems. Intrinsic unpredictability of adaptive systems is also an issue: one cannot accurately model one's adversary's systems and how an

6 Reframing Autonomous Weapons Systems

adversary will adapt to your system resulting in an inherently unpredictable act.

Candidate Recommendation

The predictability of the overall user-system-environment combination should be striven for. Having a well-informed human in the loop will help alleviate issues that come with open-world models and should be mandated.

Further Resources

- International Committee for Robot Arms Control. "LAWS: Ten Problems for Global Security" (leaflet). 10 April 2015.
- Owens, D. "Figuring Forseeability." *Wake Forest Law Review* 44 (2009): 1277, 1281–1290.
- Scherer, Matt. "Who's to Blame (Part 5): A Deeper Look at Predicting the Actions of Autonomous Weapons." *Law and AI* (blog), February 29, 2016.

Issue:

Legitimizing AWS development sets precedents that are geopolitically dangerous in the medium-term.

Background:

The widespread adoption of AWS by major powers would destabilize the international security situation by:

A. Allowing an autonomous weapon to initiate attacks in response to perceived threats, leading to unintended military escalation or war;

B. Creating weapons that adapt their behavior to avoid predictability, thereby reducing humans' ability to foresee the consequences of deployment;

C. Creating a fragile strategic balance that depends largely on the capabilities of autonomous weapons, which can change overnight due to software upgrades or cyber-infiltration; and,

D. Allowing the dynamics of constant incursions, similar to those faced in the cyberwarfare sphere, where offense is asymmetrically easier than defense, to enter the kinetic sphere.

Due to the iterative and competitive nature of weapons acquisition and use, the development and deployment of AWS creates incentives for further use and development of more sophisticated AWS in all domains. This cycle incentivizes faster decision-making in critical situations and conflicts, and more complex and less scrutable or observable processes, thereby excluding human participation in decision-

6 Reframing Autonomous Weapons Systems

making. Hence, by decoupling the number of weapons that can be deployed in an attack from the number of humans required to manage the deployment, AWS lead to the possibility of scalable weapons of mass destruction whose impact on humanity is likely to be negative.

AWS use will likely give rise to rapid escalation of conflict due to their purpose of increasing operational efficiency and tempo. Thus, there will likely be little or no opportunity for human commanders to deliberate and perform de-escalation measures in scenarios where such weapons are deployed on multiple sides of a conflict or potential conflict.

Use of AWS by two parties (or more) to a conflict will likely lead to complex interactions that are difficult to model, understand, and control. AWS also enable oppression through suppression of human rights, both in domestic and international settings, by enabling new scalabilities in enacting potentially illegal or unethical orders that human soldiers might reject.

AWS's ability to decouple the number of weapons that can be deployed in an attack from the number of humans required to manage their deployment leads to the possibility of scalable weapons of mass destruction whose impact on humanity is likely to be negative.

There is, thus, a dual, and interactive concern with regards to AWS:

1. The nature of inter-state competition in arms races yields escalatory effects with regards to arms development, deployment and proliferation; and
2. The very nature of AI in competitive and cyclical environments drives toward goal-maximizing behavior that without sufficient safeguards enables "flash crash"–type scenarios.[lxiii]

Candidate Recommendation

Autonomy in functions such as target selection, attack, and self-defense leads to negative consequences for humanity, and therefore should be curtailed by designing systems which require human involvement in such decisions. There must be meaningful human control over individual attacks.

Design, development, or engineering of AWS beyond meaningful human control that is expected to be used offensively or kill humans is to be unethical. Such systems created to act outside of the boundaries of "appropriate human judgment," "effective human control," or "meaningful human control," undermine core values technologists adopt in their typical codes of conduct.

Further Resources

- Scharre, P., and K. Sayler. "Autonomous Weapons and Human Control" (poster). Center for a New American Security, April 2016.
- International Committee for Robot Arms Control. "LAWS: Ten Problems for Global Security" (leaflet). April 10, 2015.

6 Reframing Autonomous Weapons Systems

Issue:

Exclusion of human oversight from the battlespace can too easily lead to inadvertent violation of human rights and inadvertent escalation of tensions.

Background

The ethical disintermediation afforded by AWS encourages the bypassing of ethical constraints on people's actions that should require the consent of multiple people, organizations, or chains of commands. This exclusion concentrates ethical decision making into fewer hands

Candidate Recommendation:

Design, development, or engineering of AWS for anti-personnel or anti-civilian use or purposes are unethical. An organization's values on respect and the avoidance of harm to persons precludes the creation of AWS that target human beings. If a system is designed for use against humans, such systems must be designed as semi-autonomous where the control over the critical functions remains with a human operator, (such as through a human-in-the-loop hardware interlock). Design for operator intervention must be sensitive to human factors and increasing—rather than decreasing—situational awareness. Under no circumstances is it morally permissible to use predictive or anticipatory AWS against humans. "Preventive self-defense" is not a moral justification in the case of AWS.
Ultimately, weapons systems must be under meaningful human control. AWS operating without meaningful human control should be prohibited, and as such design decisions regarding human control must be made so that a commander has meaningful human control over direct attacks during the conduct of hostilities. In short, this requires that a human commander be present and situationally aware of the circumstances on the ground as they unfold to deploy either semi-autonomous or defensive anti-materiel AWS. Organizational members must ensure that the technologies they create enhance meaningful human control over increasingly sophisticated systems and do not undermine or eliminate the values of respect, humanity, fairness, and dignity.

Further Resources

- International Committee for Robot Arms Control. "LAWS: Ten Problems for Global Security" (leaflet), April 10, 2015.
- Heller, Kevin Jon. "Why Preventive Self-Defense Violates the UN Charter." *Opinio Juris* (blog), March 7, 2012.
- Scherer, Matt. "Who's to Blame (Part 5): A Deeper Look at Predicting the Actions of Autonomous Weapons." *Law and AI* (blog), February 29, 2016.

6 Reframing Autonomous Weapons Systems

Issue:

The variety of direct and indirect customers of AWS will lead to a complex and troubling landscape of proliferation and abuse.

Background

Use of AWS by a myriad of actors of different kinds, including states (of different types of regime) and non-state actors (militia, rebel groups, individuals, companies, including private military contractors) would lead to such systems becoming commonplace anywhere anyone favors violence due to the disintermediation and scalability afforded by their availability.

There will be incentives for misuse depending upon state of conflict and type of actor. For example, such misuse may include, but is not limited to, political oppression, crimes against humanity, intimidation, assassination, and terrorism. This can lead to, for example, a single warlord targeting an opposing tribe based on their respective interests as declared on Facebook, their DNA, their mobile phones, or their looks.

Candidate Recommendations

- There is an obligation to know one's customer. One must design AWS in such a way that avoids tampering for unintended use. Further work on technical means for nonproliferation should be explored, for example, cryptographic chain authorization.
- There is an obligation to consider the foreseeable use of the system, and whether there is a high risk for misuse.
- There is an obligation to consider, reflect on, or discuss possible ethical consequences of one's research and/or the publication of that research.

Issue:

By default, the type of automation in AWS encourage rapid escalation of conflicts.

Background

One of the main advantages cited regarding autonomous weapons is that they can make decisions faster than humans can, enabling rapid defensive and offensive actions. When opposing autonomous weapons interact with each other, conflict will be able to escalate more quickly than humans on either side will be able to understand.

Candidate Recommendation

- Consider ways of limiting potential harm, for example, limited magazines, munitions, or maximum numbers of platforms in collaborative teams. Explore other technological means for limiting escalation, for example, "circuit breakers," as well as features that can support confidence-building measures between adversaries, for example, methods to communicate. All such

6 Reframing Autonomous Weapons Systems

solution options ought to precede the design, development, deployment, and use of AWS.

- Perform further research on how to temper such dynamics when designing these systems.

Issue:

There are no standards for design assurance verification of AWS.

Background

Standards for guaranteeing the compliance of autonomous and semi-autonomous weapons systems with relevant ethical and legal standards are lacking. Comprehensive international standards are needed to ensure this complex topic receives the critical evaluative process it merits.

Candidate Recommendation

It should be feasible to discern and verify that a system meets the relevant ethical and legal standards, such as international humanitarian law. We recommend efforts to standardize a comprehensive suite of verification and validation protocols for AWS and semi-autonomous weapons. Stakeholders including humanitarian organizations and AI safety concerns should contribute to the technical requirements for this.

Further Resources

- International Standards Organization. ISO 13849-1:2015: Safety of Machinery—Safety-Related Parts of Control Systems, General Principles for Design.

Issue:

Understanding the ethical boundaries of work on AWS and semi-autonomous weapons systems can be confusing.

Background

While national laws may differ on what constitutes responsibility or liability for the design of a weapons' system, given the level of complicity or the causal contribution to the development of a technology, ethics looks for lines of moral responsibility. Determining whether one is morally responsible requires us to establish relevant facts in relation to a person's acts and intentions.

Candidate Recommendation

How one determines the line between ethical and unethical work on AWS requires that one address whether the development, design, production, and use of the system under consideration is itself ethical. It is incumbent upon a member to engage in reflective judgment to consider whether or not his or

6 Reframing Autonomous Weapons Systems

her contribution will enable or give rise to AWS and their use cases. Members must be aware of the rapid, dynamic, and often escalatory natures of interactions between near-peer geopolitical adversaries or rivals. It is also incumbent upon members of a relevant technical organization to take all reasonable measures to inform themselves of the funding streams, the intended use or purpose of a technology, and the foreseeable misuse of their technology when their contribution is toward AWS in whole or in part. If their contribution to a system is foreseeably and knowingly to aid in human-aided decisions—that is, as part of a semi-autonomous weapons system—this may act as a justification for their research.

Further Resources

- Sharkey, N. "Cassandra or the False Prophet of Doom: AI Robots and War." *IEEE Intelligent Systems* 28, no. 4 (2008): 14–17.
- Noorman, Merel. "Computing and Moral Responsibility," in *The Stanford Encyclopedia of Philosophy* (Summer 2014 Edition), edited by Edward N. Zalta.