Autonomous weapons as a solution to war crimes?

Article36

WWW.ARTICLE36.ORG
INFO@ARTICLE36.ORG
@ARTICLE36

JUNE 2022
DISCUSSION PAPER

This discussion paper is written by James Dawes, Director of the Program in Human Rights and Humanitarianism at Macalester College. He is the author of The Novel of Human Rights (Harvard University Press, 2018); Evil Men (Harvard University Press, 2013 - winner of the International Human Rights Book Award); That the World May Know: Bearing Witness to Atrocity (Harvard University Press, 2007); and The Language of War (Harvard University Press, 2002).

Introduction

> "Autonomous weapons, Arkin has argued, could be programmed to never break the laws of war. They would be incapable of doing so. They wouldn't seek revenge. They wouldn't get angry or scared. They would take emotion out of the equation. They could kill when necessary and then turn killing off in an instant, if it was no longer lawful."
>
> Paul Scharre, Army of None, pp. 282-283, characterising the thinking of robotocist Ron Arkin.

Article 36 argues for prohibitions on autonomous weapons that cannot be used with meaningful human control or that would target people directly (as opposed to, say, military vehicles). However, some opponents of constraint on autonomous weapon systems assert that any prohibitions would be morally counterproductive. They argue that war crimes and atrocities are the result of specific personality traits, like an individual propensity for cruelty or racism, along with generalized human shortcomings under conditions of extreme stress, including the tendency to lose moral regulation when experiencing panic, hate, or rage. Artificial intelligence and machines, by contrast, would be immune to the emotions of war, lacking a personality that could, for instance, be sadistic. Far from being a moral worry, proponents argue, having machines make decisions about the application of force is *a potential solution* to the problem of war crimes.

This paper demonstrates why these arguments for the superior morality of autonomous weapons fail. The paper proceeds by developing three primary claims.

- Atrocities are generally the result of systems, not personalities – and AI and machines are as much a product of systems as any individual soldier.
- Autonomous weapons systems magnify and extend the capacity of human actors to do harm, and so they extend the capacity of 'bad actors' also.
- The architecture of autonomous systems intended to target people reproduces the dehumanization of others which is a principal precursor to atrocity.

Recognition of war crimes and atrocities is important to our collective constraint on behavior in war. Yet when marshalled in support of autonomous weapons systems we tend to be given only tokenistic representations of these phenomena: as if they are the primary drivers of harm in conflict and flow from the 'weakness' of humans experiencing 'emotions'. In reality, the great majority of death, injury and deprivation to civilians results from the accumulation of 'collateral damage', mistakes and, ultimately, the systemic normalization of harms that should be considered unacceptable. Far from saving us from future atrocities, proposals for allowing autonomous systems to target people reinforces the structures of dehumanization and distancing upon which that normalization of harm depends.

Atrocities are the result of systems, not personalities

After the Holocaust, multiple studies were conducted to identify the predisposing characteristics of perpetrators of atrocity. Decades of research have led to one firm conclusion: all the features purportedly common to war criminals are also common to "millions of other individuals who may have done nothing more criminal in their lives than commit a parking meter violation."[1] Perpetrators are, in Christopher Browning's phrase, "ordinary men."[2]

Today, scholars generally agree that personality is a less important variable in understanding atrocity than the information environment that embeds personality. When soldiers are trained in a system that characterizes enemy soldiers and civilians as inferior or morally tainted due to their race or gender, indiscriminate violence is amplified. Racism and sexism, in this sense, are structural rather than personal. And as study after study has revealed, artificial intelligence is as vulnerable to structural racism and sexism as any group of soldiers. Google's racist search engine that identified African-Americans as gorillas and Amazon's sexist hiring algorithm that systematically devalued women applicants are only two of the most prominent recent examples.[3]

1 James Waller, *Becoming Evil: How Ordinary People Commit Genocide and Mass Killing* (Oxford: Oxford University Press, 2007), pp. 86-87.

2 Christopher Browning, *Ordinary Men: Reserve Police Battalion 101 and the Final Solution in Poland* (New York: Harper, 1998).

3 See for links: https://eu.usatoday.com/story/tech/2015/07/01/google-apologizes-after-photos-identify-black-people-as-gorillas/29567465/ and https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

These algorithms did not produce racist and sexist results because they were malfunctioning. Malfunctioning is a separate category of concern. These algorithms produced racist and sexist outcomes because they were operating as intended, reflecting a racist and sexist information environment. Artificial intelligence does not eliminate discrimination; it absorbs discrimination into the institutional structure, making it perhaps less immediately visible - but no less harmful.

Autonomous systems magnify and extend the capacity of (bad) actors to do harm

An underpinning aspiration behind greater autonomy in weapons systems is the potential to amplify the lethal capacities of their users. Proponents imagine launching swarm attacks upon enemies using collaborative drones that calculate decisions at great speed - or sending weapons on seek-and-destroy missions over great ranges and long periods of time, and with less traceability than existing guided missiles or loitering munitions.

Weapons developers argue that these attacks will be carefully controlled and targeted, minimizing civilian casualties. Similar arguments have been made about current drone warfare, which is often ascribed with the characteristics of "surgical precision." Agnes Callamard, former UN Special Rapporteur, dismisses this justification as a "myth."[4] Due to lack of effective oversight, information about drones and civilian casualties is currently unreliable, but it is likely that official reports systematically underreport civilian harm. It has been estimated that drone strikes in non-battlefield settings have caused significant civilian deaths when compared with crewed weapon systems in conventional battlefields.[5] Whatever the hypothetical potential of armed drones, the actual patterns of use and of harm do not match the rhetoric.

Where the rhetoric around autonomous weapons is that they will be more morally sound than human soldiers; this stance tends to neglect the fact that such systems *will still be used by people*. In so far as these systems extend the capability of people to do harm, they also extend the potential for inadvertent harms – not only from the sorts of embedded bias discussed above but also as a result of 'accidents' and technical vulnerabilities, and the capability of bad actors to create bad effects.

Autonomous weapons can be usefully viewed through the lens of "normal accident" theory. Sociologist Charles Perrow argues that in complex, tightly-coupled systems with catastrophic potential (his primary case study is the 1979 Three Mile Island nuclear accident), operator errors are often less important than predictable, operator-independent system errors that can cascade into major disasters. The amplified capabilities of autonomous weapons (to identify and strike more targets, whilst being allowed to operate independently over wider areas and longer periods of time) will amplify the problem of normal accidents and resultant civilian harms.

4 https://news.un.org/en/story/2020/07/1068041

5 https://foreignpolicy.com/2016/04/25/drones-kill-more-civilians-than-pilots-do/

This problem is exacerbated by specific issues of concern beyond the inevitable problem of mistakes and malfunctioning. Autonomous systems may be vulnerable to the same sort of hacking that plagues all technology today, including "spoofing" attacks, data "poisoning," and input attacks. Arthur Holland Michel **argues** that failures in autonomous systems are "inevitable" given pervasive data vulnerability. "We know that such problems exist either now or will emerge in the future, but we cannot characterize or specifically anticipate them. One might call such data issues 'known unknowns'."[6]

Finally, not all militaries can be counted on to deploy weapons (including autonomous weapons) under an interpretation of the law that gives sufficient weight to civilian protection, and in some cases may give little concern for the law at all. In so far as the proponents of autonomous weapons present them as superior to people they miss that they will be programmed and reprogrammed by people, and used by people in specific contexts. The history of weapons technology demonstrates that weapons spread out of the hands of those who initially wish to monopolize them and into the hands of their enemies. Because some weapon technologies in this space are likely to be comparatively inexpensive and easy to scale up compared to other complex military systems their spread to nonstate armed groups would be very likely. The amplifying capability of autonomous weapons that is now an aspiration could come to be lamented.

Again, we should note at this point that the arguments set out in this section are not aimed at asserting the need for specific prohibitions and other regulations in this space (though that is our wider contention). Rather the arguments here are a rejection of the claims made by proponents of autonomous weapons that these systems will spare us from the perceived shortcomings of human decision makers.

The architecture of autonomous systems reproduces the dehumanization of others which is a precursor to atrocity

Understanding this claim requires understanding three key features of the process of dehumanization: distance, the causal chain, and "moral slide."

Researchers note that atrocity "up close" requires passing several behavioral thresholds that inhibit violence.[7] When we engage another person in direct conflict, the pull of their humanity is urgent; our responsibility for harming them is immediate; and their specificity as individuals forces us to make spur-of-the-moment contextual judgments and choices (do they seem young or old? like or unlike me? frightened? wounded?). Harming a person directly, philosopher Thomas Nagel writes, "puts you in a special relationship to [them]," which means, from a moral perspective, actions you take "may have to be defended."[8]

6 https://unidir.org/known-unknowns

7 Jonathan Glover, *Humanity: a Moral History of the Twentieth Century* (New Haven: Yale University Press, 2000), pp. 113-116.

8 Thomas Nagel, *Mortal Questions* (Cambridge: Cambridge University Press, 1979), pp. 68-69.

Violence is easier to inflict (and, for observers, to allow) when the target of violence is distant because none of these inhibiting features are present. Targets are not specific people facing you. They are, with drones for instance, images on a screen being targeted by an organizational and technological nexus of which you are a part. At this sort of remove, Nagel writes, we are not "facing or addressing the victim at all, but operating on [them]" – "a purely bureaucratic operation."[9] This layer of distance leads inevitably to a deterioration of respect for the individual humanity of others, as evidenced by the name given to computer software developed by the Pentagon during the war in Iraq, "Bugsplat," which was used to calculate collateral damage. With autonomous weapons, the dangers of distance and dehumanization are necessarily amplified.[10] Targets are not "bugsplat" images on the screen. They are at an even further remove from face-to-face humans: *they remain hypothetical*.

The causal chain of violence is another key aspect of war crimes. As moral philosophers frequently note, we feel a deep difference in moral responsibility between doing something ourselves and allowing something to be done. The famous "trolley" problem of moral philosophy hinges entirely upon the difference between what if feels like to cause a train to run somebody over and to *not stop* a train from running somebody over. With autonomous weapons, we allow violence rather than commit it. Our felt responsibility as actors is radically diminished, further reducing the behavioral inhibitions that we see most clearly in "up close" violence.

Dehumanization combined with indirect causal chains makes the choice to commit violence easier, promoting the "moral slide"[11] that is an essential feature of atrocity. Philosopher Jonathan Glover argues that the bombing of Hiroshima and Nagasaki was made possible only because of the earlier, seemingly more tolerable steps that preceded it: the blockading of cities to coerce the civilian population enabled a shift from (comparatively) targeted aerial bombing to the carpet-bombing cities, which in turn enabled nuclear destruction of whole cities to be countenanced. In my own work with war criminals, I have described how even the very worst war criminals from World War II began as ordinary people with normal moral identities. They were trained, through a process of "incremental escalation," to overcome their moral inhibitions and resort to violence with increasing ease.[12] The end result was not perfected violence that was quick and efficient but rather violence that was increasingly indiscriminate and disproportionate to military aims.

With autonomous weapons development there is also a direct, market-driven analog to 'moral slide': a "race to the bottom on safety," which Paul Scharre has argued is a likely outcome as nations "cut corners" to deploy systems faster than their competitors, fielding weapons before they are ready.[13] Without clear normative and legal constraints, it is easy imagine an initial expansion of autonomous weapon system

9 Nagel, p. 67.

10 https://www.washingtonpost.com/archive/politics/2003/02/21/military-turns-to-software-to-cut-civilian-casualties/af3e06a3-e2b2-4258-b511-31a3425bde31/

11 Glover, p. 115.

12 James Dawes, *Evil Men* (Cambridge: Harvard University Press, 2013), p. 68.

13 https://stanleycenter.org/wp-content/uploads/2020/05/MilitaryApplicationsofArtificialIntelligence-US.pdf

use, within a loose consensus around how they should be designed, where and how often they should be deployed, and what means of violence they may use, only for this to change over time. Subsequent stages – defined by dehumanization, diffusion of responsibility, and moral slide – would likely see designs and deployments becoming less constrained: less limited in terms of target types and subject to less meaningful human control. Such a slide, in turn, could change how conventional warfare is perceived, with increasing tolerance of previously unacceptable violence.

In sum, war defined by hypothetical and distant targets is war that begins by failing to positively acknowledge the humanity of the enemy, by reducing people to categories, to simple, interpretable packets of information. Even if we indulge in the fantasy of AI that makes those sample humans so fine-grained that atrocity-by-error is impossible (the weapon can, in complex and dynamic environments, make all the necessary distinctions between individuals that a face-to-face human can make), we are nonetheless left in a world where the key features of atrocity are now absorbed into the larger institutional structures that embed autonomous weapons.

WWW.ARTICLE36.ORG
INFO@ARTICLE36.ORG
@ARTICLE36

Article 36 is a specialist non-profit organisation, focused on reducing harm from weapons.