ITSS
International Team
For the Study of Security
Verona

Finding balance: Ethics, AI, innovation and Autonomous Weapons

by Alessia Maira

ITSS Verona Magazine, Vol. 2, n. 2

Fall/Winter 2023

Finding balance: Ethics, AI, innovation and Autonomous Weapons

Alessia Maira

To cite this article: Alessia Maira, *Finding balance: Ethics, AI, innovation and Autonomous Weapons*, ITSS Verona Magazine, Vol. 2, no. 2, Fall/Winter 2023.

Keywords: autonomous weapons (AWs), ethics, warfare, AI, human oversight

ITSS Verona website: https://www.itssverona.it/itss-magazine

LinkedIn: https://www.linkedin.com/company/itss-verona/

Instagram: https://instagram.com/itss_verona?igshid=YmMyMTA2M2Y=

Published online: December 29th, 2023

Abstract: This article delves into the complex interplay between ethical concerns and innovation opportunities surrounding AI-based Autonomous Weapons (AWs). It emphasizes the critical importance of human judgement, drawing on historical examples that highlight the need for human oversight in critical scenarios in which AI is utilized. The goal of this article is to investigate the potential benefits of the use of AI with the case study of AWs and the ethical concerns that arise with their use. The limitations of AI in grasping contextual meaning and situational nuances are examined critically, raising questions about the deployment of AWs. The concept of "keeping a human in the loop" is explored as a mitigating strategy, balancing human oversight with the efficiency of AI in general. A comprehensive analysis of real-world examples and scholarly viewpoints is employed to explore the ethical dilemmas that surround AI-based AWs. The research methodology draws from the insights of philosophers, historians, scholars, and military strategists. It also involves a critical evaluation of AI capabilities which also considers the perspective of international bodies like the Red Cross or governmental entities like the U.S. Department of Defense. The article finally concludes by advocating for a careful balance between innovation, non-proliferation efforts, and ethical standards. While recognizing the potential of AWs, the article underlines the importance of prioritizing ethical principles. It asserts the necessity of continued international dialogue, urging discussions aimed at establishing a legally binding protocol for the usage of AWs.

On September 26, 1983, a Soviet Union lieutenant colonel for the Air Defense Forces named Stanislav Petrov was sitting in a bunker south-west of Moscow when alarms began to ring. According to the satellite detection system, five intercontinental missiles were on their way from the U.S. to the Soviet Union. However, Petrov felt something was wrong. "When people start a war, they don't start it with only five missiles" he reasoned. He followed his gut and phoned his commanders to notify them that the computer warnings were wrong.[1] Petrov's actions prevented a nuclear war. Were the early-warning missile detection system fully autonomous in response – to allow for faster reaction – the outcome scenario could have been totally different. This example highlights the critical role of human judgement and multi-faceted reasoning in critical scenarios. This paper argues that ethical concerns about AI-based autonomous weapons (AWs) are well-founded.

Ethical concerns

In the present world, with the progression of AI-based technologies, there are apprehensions with the idea of granting AI with the power to make life-or-death decisions. The debate over AWs is also gaining more prominence in public discourse as AI-based technologies become increasingly intertwined into everyday life. Many are raising concerns with regard to the advancement of AI, including author and historian Yuval Noah Harari who argues that AI has already: "hacked the operating system of human civilization."[2] Nick Bostrom, a Swedish philosopher, is going as far as proposing a scenario in which we prompt a computer with a simple task - such as counting paper clips - that would result in a no longer controllable or stoppable sentient AI that could surpass human intelligence and cause immense destruction in pursuit of its task.[3] Not as drastic, Bill Gates recognizes the risks of AI but deems them to be manageable.[4] Nevertheless, in such a complex and

[1] David Hoffman, "I Had A Funny Feeling in My Gut," *The Washington Post Foreign Service*, February 10, 1999,
https://www.washingtonpost.com/wp-srv/inatl/longterm/coldwar/soviet10.htm#TOP

[2] Yuval Noah Harari, "Yuval Noah Harari argues that AI has hacked the operating system of human civilization," *The Economist*, April 28, 2023,
https://www-economist-com.libproxy.kcl.ac.uk/by-invitation/2023/04/28/yuval-noah-harari-argues-that-ai-has-hacked-the-operating-system-of-human-civilisation

[3] Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: Oxford University Press, Incorporated, 2014), 116.

[4] Bill Gates, "The risks of AI are real but manageable," *Gates Notes*, July 11, 2023,

constantly developing landscape, it is arguably difficult to find a balance between innovation opportunities and ethical concerns that surround AI-based AWs.

One of the most relevant issues regarding current levels of AI applied to AWs is that they: "would select and attack targets without meaningful human control."[5] For instance, Kenneth Payne provides a clear example: an AI could be able to identify those holding a weapon, but not to understand *why* someone is holding one, for instance, in the air because they want to surrender. In this case, the AI could discriminate and unjustly decide to eliminate what it considers to be a threat due to its failure to understand the situational meaning and contextualize situations. Put simply, AI has an: "inability to grasp meaning with anywhere near the same facility as humans."[6] Furthermore, strange or unexpected behaviors might arise when AI-based systems are operating in a changing environment with fast-paced dynamics.[7]

One of the most relevant technologies under debate are groups of drones, known as "swarms". These can be defined as "cooperative, autonomous robots that react to the battlefield as one at machine speed", interacting with their environment and one another to coordinate action.[8] Swarms have the potential to bring revolution in how warfare is conducted, but not without raising ethical concerns. For example, slaughterbots are considered lethal autonomous weapon systems (LAWS) and their usage raises many ethical concerns. Conceived as a "machine to kill", slaughterbots are designed to operate as a single unit or in swarms by using discriminative

https://www.gatesnotes.com/The-risks-of-AI-are-real-but-manageable

[5] Human Rights Watch, "UN: Key Action on 'Killer Robots', International Move Toward Possible Ban," *Human Rights Watch*, December 16, 2016, https://www.hrw.org/news/2016/12/16/un-key-action-killer-robots

[6] Kenneth Payne, *Strategy, Evolution, and War: From Apes to Artificial Intelligence* (Washington: Georgetown University Press, 2018), 173.

[7] Irving Lachow, "The upside and downside of swarming drones," *Bulletin of the Atomic Scientist* 73, no. 2 (2017): 98, https://doi.org/10.1080/00963402.2017.1290879

[8] Paul Scharre, "How swarming will change warfare," *Bulletin of the Atomic Scientist* 74, no. 6 (2018): 385, https://doi.org/10.1080/00963402.2018.1533209

identification (e.g. based on physical markers) to kill their targets.[9] LAWS utilize AI to operate without human intervention, meaning that they independently identify and select targets to kill.[10]

Ethical issues are clearly identifiable, especially for the target identification standards based on factors like appearance and physical features. For instance, facial recognition technologies used in the U.S. have been criticized for their low accuracy rates and were: "banned for use by police and local agencies in several cities, including Boston and San Francisco."[11] In 2015, in response to rising concerns over AWs, the Future of Life institute published an open letter recognizing the potential of AI but also criticizing an eventual "military AI arms race", arguing that the latter should be avoided through: "a ban on offensive autonomous weapons [that operate] beyond meaningful human control."[12]

AI-based AWs also act faster than humans. Cognitively, the response time of AI is unmatched by that of humans, but this is not always for the best outcome. Research conducted by RAND on a wargame model of autonomous AI found that the speed at which machines operate led to an escalation of the wargame in several instances, both intentionally and inadvertently. They found that "manned systems may be better for deterrence than unmanned ones", and that "widespread AI and autonomous systems could lead to inadvertent escalation and crisis instability".[13] Furthermore, lower risks to military personnel might further increase the tendency to engage in armed conflict, as AWs are "more expendable" than human lives.[14]

[9] "Slaughterbots," YouTube video, 7:47, posted by "The Future of Life Institute," November 13, 2017, https://www.youtube.com/watch?v=HipTO_7mUOw&t=7s

[10] Taylor Jones, "An introduction to the issue of Lethal Autonomous Weapons," *Future of Life Institute*, November 30, 2021, https://futureoflife.org/aws/an-introduction-to-the-issue-of-lethal-autonomous-weapons/

[11] Alex Najibi, "Racial Discrimination in Face Recognition Technology," *Harvard University Blog for Science Policy and Social Justice*, October 24, 2020, https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/

[12] Future of Life Institute, "Autonomous Weapons Open Letter: AI & Robotics Researchers," *Future of Life Institute*, February 9, 2016, https://futureoflife.org/open-letter/open-letter-autonomous-weapons-ai-robotics/

[13] Yuna Huh Wong, John Yurchak, Robert W. Button, Aaron B. Frank, Burgess Laird, Osonde A. Osoba, Randall Steeb, Benjamin N. Harris, and Sebastian Joon Bae, *Deterrence in the Age of Thinking Machines* (Santa Monica, CA: RAND Corporation, 2020): x-xi. https://www.rand.org/pubs/research_reports/RR2797.html

[14] See note 8.

Finally, lack of human judgement, unpredictability and faster escalation are not the only ethical concerns. AWs, if used correctly, have the potential of discriminate targeting and quick, precise execution. In the wrong hands, however, they can cause serious issues, and risks of proliferation are particularly high in the case of drones.[15] The proliferation of AWs could be similar to that of cyberspace attacks, where only a computer, skills, and an internet connection are enough to cause disruption. Similarly, due to the low costs encountered: "a range of terrorist, insurgent, criminal, corporate and activist threat groups have already demonstrated the ability to use civilian drones for attacks."[16] Examples include Hezbollah, which has the longest history of drone use by a non-state group [...] in the border region between Syria and Lebanon, or, in the case of activists, a protester against Japan's nuclear energy policy once landed a drone loaded with radioactive sand on the rooftop of the Prime Minister's office in Tokyo.[17]

AI as an innovative force?

Despite ethical concerns that might arise, AWs could offer a range of innovation opportunities to military strategy, such as increased precision, less military personnel loss, fewer civilian casualties, and more rapid response times in critical situations, with the almost more than certain potential to cause a shift in war strategy and future warfare as a whole.

One of the first arguments in favour of AWs is their alleged potential to reduce civilian casualties. AI-based weapons are favoured not only for their increased precision, but also if the issue is approached by a different ethical perspective. The concept that Australian moral philosopher Peter Singer defines as an expanding circle of empathy in his book, i.e., that "ethics run deep in our species and is common to human beings everywhere" [18] can be interpreted with the liberal tradition in which the ethics of AI have been developed so far. The increased importance of

[15] Irving Lachow, "The upside and downside of swarming drones," *Bulletin of the Atomic Scientist* 73, no. 2 (2017): 99,
https://doi.org/10.1080/00963402.2017.1290879

[16] Chris Abbott, Mattew Clarke, Steve Hathorn, Scott Hickie, *Hostile drones: the hostile use of drones by non-state actors against British targets*, (London, UK: The Remote Control Project and Open Briefing, 2016), 1.
https://www.files.ethz.ch/isn/195685/Hostile%20use%20of%20drones%20report_open%20briefing_0.pdf

[17] See note above, 12-13.

[18] Peter Singer, *The Expanding Circle: Ethics, Evolution, and Moral Progress* (Princeton University Press, 1981), 27.

public opinion regarding acts of war is closely correlated with the expanding circle of moral concerns and ethics that characterises liberal societies. In parallel to ethical concerns, the increased precision of automatic weapons allows for the discriminate pursuing of targets with an unprecedented precision. Today, it would not be possible to justify an indiscriminate bombing campaign of any sort without public concerns and protests. On the other hand, drone strikes might be more easily accepted by the general population, if the latter is presented with ethical considerations which motivate AWs usage with the goal to cause less damage possible in the target country (fewer civilian casualties), and lower costs of loss of equipment and of human military personnel in case of failures or issues with the operation.

Furthermore, drone swarms can be used in different ways: attack, defence, and as support of military operations in surveillance and reconnaissance tasks. Payne argues that, for Clausewitzian principles, "the defence is stronger than the attack" due to knowledge of the terrain and the incentive of protecting one's homeland, but tactical AI systems will likely introduce some significant changes.[19] AWs will likely allow for easier attack campaigns even in remote locations. While, for now, the largest advantage of AI-based AWs lies in their attack advantage, swarms could also be of relevance in defensive settings, especially in response to another swarm attack, for example, according to Lachow, by "creating large numbers of decoys that sow confusion or actively disrupt an attacking force."[20] Arguably, under a swarm offensive, only a defensive swarm could be able to respond with the same rapidness and unpredictability.

How to mitigate ethical concerns?

Ethical concerns could be addressed through different measures and achievements. The Red Cross (ICRC) is not the only international body to advocate in favour of the maintenance of human control "over weapon systems and the use of force to ensure compliance with international law and to satisfy ethical concerns".[21] Concerns of this kind are well-founded. Due to the complexity of AI

[19] Kenneth Payne, *Strategy, Evolution, and War*, 178.

[20] Irving Lachow, "The upside and downside of swarming drones," 98.

[21] ICRC, "Ethics and autonomous weapon systems: An ethical basis for human control?," Working paper, Group of Governmental Experts of the High Contracting Parties to the Convention on Prohibitions or Restrictions on

systems, they sometimes exhibit strange or unexpected behaviours. In the case of AWs, and LAWS in particular, ethical issues are evident. A widespread approach to mitigate ethical concerns is that of always "keeping a human in the loop" to overlook the actions of AWs. For example, a U.S. Department of Defense (DoD) directive, "Autonomy in Weapon Systems", specifies that "autonomous and semi-autonomous weapon systems will be designed to allow commanders and operators to exercise appropriate levels of human judgement over the use of force" [22], a directive intended to diminish risks associated with failures and unintended engagement of AWs. However, others argue that while human control is valuable in certain contexts (e.g., to confirm whether to fire on a target or not) it might not always be the best solution. In fact, keeping a human in the loop might result in slower response times, at the expense of speed and efficiency.[23] The development of "explainable AI" would allow further mitigation of ethical concerns, but one could raise concerns over the feasibility of such technology, given that human choices are uncategorisable into a dichotomous definition of 'what is correct' and 'what is not'. Doing that with machines would entail, first, achieving a rigorous understanding of how the human mind works: an impossible task, given the existence of different cultures and myriads of moral perspectives.[24]

Conclusion

The focus of the debate surrounding AWs lies on ethical concerns that derive from the lack of human control over AWs. Stories like that of Petrov serve us as a reminder of the fundamental role of human judgement in critical scenarios. Concerns about the ethics of AWs are not unfounded, as these arguably new technologies might challenge the status quo of military strategy the same way that nuclear bombs did in the 19th century.

the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects, 2018: 1. https://documents-dds-ny.un.org/doc/UNDOC/GEN/G18/082/07/pdf/G1808207.pdf?OpenElement

[22] US Department of Defense, *DoD Directive 3000.09: Autonomy in Weapon Systems*, January 25, 2023 (Washington DC: Office of the Under Secretary of Defense for Policy): 3, https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf

[23] Jovana Davidovic, "What's wrong with wanting a 'human in the loop'?" *War on the Rocks*, June 23, 2022, https://warontherocks.com/2022/06/whats-wrong-with-wanting-a-human-in-the-loop/

[24] Keith Dear, "Artificial Intelligence and Decision-Making," *The RUSI Journal* 164, no. 5-6 (2019): 25, https://doi.org/10.1080/03071847.2019.1693801

“Is it more or less moral to employ autonomous rather than human-controlled weapons?” one might ask themselves.[25] Finding a balance between continued innovation, non-proliferation and ethical standards will be key for the correct continuous development of AI-based AWs and technologies. The concept of “keeping a human in the loop” is essential for a gradual transition towards more technologically oriented armed forces – while keeping ethical standards in check. Still, it is undeniable that there are arguments in support of the use of AWs, such as the innovation they could bring to armies, their increased precision, faster response time, the substitution of deployed personnel, and the potential to reduce civilian casualties. With continuous development, these technologies have the potential to continue to improve through the years and revolutionize warfare. However, for the foreseeable future, humans should maintain a place “in the loop” when possible, while placing efforts on resuming international talks under the United Nations Convention on Certain Conventional Weapons (UNCCW) to discuss a legally binding protocol for the usage of LAWS. While this might lead to slower decisions, safeguarding ethical principles should come first. For the time being, only humans should be able to make decisions over the lives of other humans.

[25] Irving Lachow, “The upside and downside of swarming drones,” 99.

Bibliography

Abbott, Chris, Mattew Clarke, Steve Hathorn, Scott Hickie. *Hostile drones: the hostile use of drones by non-state actors against British targets*. London, UK: The Remote Control Project and Open Briefing, 2016. https://www.files.ethz.ch/isn/195685/Hostile%20use%20of%20drones%20report_open%20briefing_0.pdf

Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press, Incorporated, 2014. https://ebookcentral.proquest.com/lib/kcl/detail.action?docID=5746177

Davidovic, Jovana. "What's wrong with wanting a 'human in the loop'?" *War on the Rocks*, June 23, 2022. https://warontherocks.com/2022/06/whats-wrong-with-wanting-a-human-in-the-loop/

Dear, Keith. "Artificial Intelligence and Decision-Making." *The RUSI Journal* 164, no. 5-6 (2019): 18-25. https://doi.org/10.1080/03071847.2019.1693801

Future of Life Institute. "Autonomous Weapons Open Letter: AI & Robotics Researchers." *Future of Life Institute*, February 9, 2016. https://futureoflife.org/open-letter/open-letter-autonomous-weapons-ai-robotics/

Gates, Bill. "The risks of AI are real but manageable." *Gates Notes*, July 11, 2023. https://www.gatesnotes.com/The-risks-of-AI-are-real-but-manageable

Harari, Yuval Noah. "Yuval Noah Harari argues that AI has hacked the operating system of human civilization." *The Economist*, April 28, 2023. https://www-economist-com.libproxy.kcl.ac.uk/by-invitation/2023/04/28/yuval-noah-harari-argues-that-ai-has-hacked-the-operating-system-of-human-civilisation

Hoffman, David. "I Had A Funny Feeling in My Gut." *The Washington Post Foreign Service*, February 10, 1999. https://www.washingtonpost.com/wp-srv/inatl/longterm/coldwar/soviet10.htm#TOP

Human Rights Watch. "UN: Key Action on 'Killer Robots', International Move Toward Possible Ban." *Human Rights Watch*, December 16, 2016. https://www.hrw.org/news/2016/12/16/un-key-action-killer-robots

International Committee of the Red Cross (ICRC). "Ethics and autonomous weapon systems: An ethical basis for human control?" Working paper, Group of Governmental Experts of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects, 2018. https://documents-dds-ny.un.org/doc/UNDOC/GEN/G18/082/07/pdf/G1808207.pdf?OpenElement

Jones. Taylor. "An introduction to the issue of Lethal Autonomous Weapons." *Future of Life Institute*, November 30, 2021. https://futureoflife.org/aws/an-introduction-to-the-issue-of-lethal-autonomous-weapons/

Kallenborn, Zachary. "Swarm talk: understanding drone typology." *Modern War Institute at West Point*, October 12, 2021. https://mwi.westpoint.edu/swarm-talk-understanding-drone-typology/

Lachow, Irving. "The upside and downside of swarming drones." *Bulletin of the Atomic Scientist*, 73, no. 2 (2017): 96-101. https://doi.org/10.1080/00963402.2017.1290879

Najibi, Alex. "Racial Discrimination in Face Recognition Technology." *Harvard University Blog for Science Policy and Social Justice*, October 24, 2020. https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/

Payne, Kenneth. *Strategy, Evolution, and War: From Apes to Artificial Intelligence*. Washington: Georgetown University Press, 2018. https://ebookcentral.proquest.com/lib/kcl/detail.action?docID=5394997

Scharre, Paul. "How swarming will change warfare." *Bulletin of the Atomic Scientist*, 74, no. 6 (2018): 385-389. https://doi.org/10.1080/00963402.2018.1533209

Singer, Peter. *The Expanding Circle: Ethics, Evolution, and Moral Progress*. Princeton: Princeton University Press, 2011. https://doi.org/10.1515/9781400838431

"Slaughterbots." YouTube video, 7:47. Posted by "The Future of Life Institute," November 13, 2017. https://www.youtube.com/watch?v=HipTO_7mUOw&t=7s

US Department of Defense. *DoD Directive 3000.09: Autonomy in Weapon Systems*, January 25, 2023. Washington DC: Office of the Under Secretary of Defense for Policy. https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf

Wong, Yuna Huh, John Yurchak, Robert W. Button, Aaron B. Frank, Burgess Laird, Osonde A. Osoba, Randall Steeb, Benjamin N. Harris, and Sebastian Joon Bae. *Deterrence in the Age of Thinking Machines*. Santa Monica, CA: RAND Corporation, 2020. https://www.rand.org/pubs/research_reports/RR2797.html