Autonomous Weapon Systems

The Issues of Autonomy, Predictability, and Ethical Neutrality

AdalanAI

24.10.2022

Executive Summary

The development and proliferation of autonomous weapon systems (AWS) has been discussed from a variety of perspectives and disciplines. While literature exists on the ethical, legal, and security aspects of AWS deployment, the conceptual dimensions and implications of the issues of AWS autonomy and predictability have yet to be defined and analyzed. In addition, although there is a growing interest in the AI ethics, little or no attention is dedicated to the question of whether the principles that constitute the safe and fair use of AI inherently contain ethical preferences or whether they are completely value-free and neutral.

To eliminate the conceptual ambiguities surrounding the topics of AWS autonomy, predictability, and ethical neutrality, we first define what autonomy means and distinguish it from "automaticity." According to the definition, autonomy is the ability of an AI system to act independently and without human intervention from the beginning of deployment and activation.

With the clear view on the meaning of autonomy, It is easier to discuss how the principle of predictability works in the context of the AWS use. We propose that the degree of predictability and explainability depends on the three major variables- the system itself, the environment AWS operates in and the task AWS performs. Instead of outright prohibition of all forms of autonomous weapons, we advise policymakers and regulators to focus on the dynamics of these variables and then decide on reasonable restrictions.

Although it is relatively clear how the principle of predictability can be applied in practice, the essence of this principle is far from being explored. We believe that, unlike other principles such as fairness, accountability, or safety, the principle of predictability does not contain ethical preferences of its own. For this reason, the ethical neutrality inherent in the principle of predictability can be abused. Therefore, it is important that any principles governing the use of AI be based on universal ethical norms. We conclude that without predictability and explainability, it is extremely difficult to achieve safe, fair, and ethical use of AWS. And without a strong ethical framework and foundation, it is impossible for predictability and explainability to help create a safe, fair, and ethical AWS.

Introduction

The increasing development and deployment of autonomous weapon systems (AWS) is a much-debated topic among scholars, policymakers, and business leaders.The enormous and, to some, opaque potential of autonomous weapons raises urgent questions about the impact AWS is capable of generating in various fields and domains. Some emphasize the ethical dimensions of AWS use, pointing to the moral risks autonomous weapons pose when deployed in real-world, complex environments [1]. Others stress the importance of proactive and timely action by the international community and the need for International humanitarian law to grasp and take into account the realities of AWS [2]. In addition, discussions about the development and use of autonomous weapons are framed in terms of the AI arms race and international arms control. Some argue that we are already witnessing the AI arms race [3], others disagree and consider the "arms race" to be analytically useless in capturing the current reality of AWS Development [4]. To address the issues the weaponized AI raises and to formalize the related discussions, The Group of Governmental Experts (GGE) was established in 2016 under the UN Convention on Certain Conventional Weapons (CCW). Despite the relative formalization and expertization of the discussions, the concrete approaches and answers to the most problematic issues related to the use of autonomous weapons are still far from being universally shared and accepted. The purpose of this brief paper is to elucidate several pressing issues concerning autonomous weapons.

Autonomy

Many modern military technologies operate with some degree of autonomy. From modern air defense systems, remotely piloted drones, guided missiles to certain mines and booby traps, all of these technologies have features of autonomy [5]. The question arises, however, as to how we can distinguish between autonomous weapon systems and systems with certain levels of autonomy.

Scholars, policymakers, and civil society organizations have attempted to capture the meaning of autonomy in the realm of international security and military technology. According to the definition proposed by the International Committee of the Red Cross, autonomy refers to a system that selects and applies force to targets without human intervention [6]. This definition is clearly flexible and general, and this flexibility and generality is both an advantage and a disadvantage. In some cases, it is advantageous to use this definition because it has the potential to include many different models, but it is disadvantageous when one tries to demarcate the seemingly thin lines between autonomy and automaticity. "Selecting" and "engaging" targets "without human intervention" is the hallmark of many military technologies, such as landmines and aerial defense systems that detect and engage the target using various heating mechanisms and electromagnetic sensors. However, This capability does not make landmines an autonomous weapon system.

> ***"Selecting" and "engaging" targets "without human intervention" is the hallmark of many military technologies, such as landmines and aerial defense systems that detect and engage the target using various heating mechanisms and electromagnetic sensors. However, This capability does not make landmines an autonomous weapon system.***

In an effort to eliminate the ambiguity created by the conceptual fallacy of referring to "automatic" and "automated" systems as autonomous systems, research published by the Stockholm Peace Research Institute defines autonomy as "the ability of a machine to perform one or more tasks without human intervention by interacting with the environment through computer programming.”[7] This definition stands out from other definitional attempts in that it clearly emphasizes one of the essential domains of autonomy, i.e., the ability to function without human input or intervention. Others furtherly structure the framework of autonomy by proposing two dimensions of autonomy: planning autonomy (referring to the ability of AI systems to independently devise and construct plans to realize and fulfill operator’s orders), and learning autonomy (referring to the ability of AI systems to adapt to new environments) [8].

Considering the main advantages and disadvantages of the predominant definitions, we propose to define autonomy in military technology as follows: Autonomy is the ability of an AI system to act independently, without human intervention from the onset of deployment and activation. Unlike automated or automatic systems that perform their tasks with some degree of human control, autonomous systems are able to make a choice between different alternatives and act independently without human intervention from the beginning to the end of the task.

> ***We propose to define autonomy in military technology as follows:***
>
> ***Autonomy is the ability of an AI system to act independently, without human intervention from the onset of deployment and activation. Unlike automated or automatic systems that perform their tasks with some degree of human control, autonomous systems are able to make a choice between different alternatives and act independently without human intervention from the beginning to the end of the task.***

> ***Autonomy is not a problem per se. Risks and problems arise from the nature of the task and the environment.***

> ***Autonomy is not a monolithic condition. Rather, it varies according to the degree of autonomy, context, realities, and tasks.***

Autonomy is not a monolithic condition. Rather, it varies according to the degree of autonomy, context, realities, and tasks. An autonomous system that performs complex tasks in a complicated context and reality with a high degree of autonomy is different from an autonomous system that performs relatively less complex and simpler tasks in conditions that are normal and much less complicated. Accordingly, the risks posed by AWS vary depending on the tasks that AWS has to perform and the environment in which it is deployed. Using elements of autonomy in the navigation system is not risky, but autonomous targeting poses a great risk to humans [9]. The debate about the autonomy of AWS should focus more on the tasks that AWS has to perform and the complexity of the environment it faces in accomplishing those tasks. That being said, autonomy is not a problem per se. Risks and problems arise from the nature of the task and the environment. The international community should address the problem of using military autonomous systems in dangerous situations while human control is minimal or nonexistent.

Predictability and Explainability

Predictability and explainability are essential concepts and features when it comes to safe and fair use of AWS. Predictability is the ability to predict what the AI system will do. AI systems are predictable if their outputs are possible to predict. Explainability is the condition when the explanation for algorithmic and data-driven decisions are easy to understand. Predictability and explainability are different determinants, but they are closely related. If the system is predictable, there is a high probability that it is explainable at the same time, and vice versa. While predictability and explainability constitute the essential features for safe AWS, It is high time to ask -why is it important for AWS to be predictable and explainable? And what does it mean for AWS to be predictable and explainable?

The requirement for predictability and explainability of AI systems in general and AWS in particular derives from the right to explanation, i.e., the legal right of individuals to a clear and understandable explanation of the overall performance of an AI system. This right provides the legal basis for implementing the principles of predictability and explainability. Predictability is the tool to determine the possible actions of the AI and to guarantee obtaining the intended outcome. On the other hand, explainability provides information as to why an AI system works the way it does. An appropriate interplay between predictability and explainability creates a solid foundation for safe AI.

It is obvious that achieving a high level of predictability and explainability is important to create trustworthy and safe AI, but there are considerable trade-offs. The more complex and productive an AI system is, the less predictable and explainable its actions tend to be (predictability/explainability- performance dilemma) [10]. Sophisticated AI models with the ability to perform difficult and complex tasks are usually black-box models. When the AI model functions as a black box, it is extremely difficult to identify, predict, and communicate the logic of AI behavior. If the functional logic of the model is unclear, it means that it cannot be changed significantly.

The degree of predictability and explainability depends on several variables. The first variable is the **AI system itself**, i.e., how sophisticated and complex the system is. The characteristics of the system largely determine the level of predictability and explainability. When the AWS consists of complex and dynamic mechanisms and functions, the probability of correct prediction and clear explanation decreases. Again, we face predictability/explainability-performance dilemma that dictates the logic behind the contradictory interaction between high performance and high predictability/explainability. Rather than outlining general and abstract approaches to regulating AWS, it could be more useful to focus more on patterns of concrete manifestations of AWS. If AWS is a black-box model, it would be appropriate to either ban it altogether or restrict its use in hazardous situations.

Here we come to the second variable on which the degree of predictability/explainability depends. This variable is the environment and context. The probability of high predictability decreases when the AWS operates in a complex environment, and vice versa. For example, if the AWS is deployed in a small building where there are only 5-10 enemy targets, it is quite easy to predict how the AWS will engage the targets (here we should also consider the first variable - the nature of the AI system itself). However, if the operator activates the AWS in a complex environment, e.g. in an urban area where it is extremely difficult to distinguish enemy targets from non-combatant civilians, the probability of low predictability increases.

The third variable that should be emphasized is the nature of tasks AWS has to perform. If the task is multifaceted and sophisticated, it is highly likely that the level of predictability/explainability will be low. The nature of the task has the potential to determine how predictable and explainable the AWS will be. To sum up, rather than totally banning every manifestation of autonomous weapons, the policymakers and regulators should focus on the AI system, the environment it operates in and the task it performs.

> ***Rather than totally banning every manifestation of autonomous weapons, the policymakers and regulators should focus on the AI system, the environment it operates in and the task it performs.***

Predictability/Explainability as value free features

Discussions concerning the importance of predictability and explainability in AWS tend to avoid and exclude the fact that predictability and explainability are ethically neutral features that are obviously fundamental to the safe use of AWS, but are nonetheless value-free by their very nature. Unlike the principles of fairness, safety, and accountability, the concepts of predictability and explainability contain no intrinsic ethical preferences. Although there are many different metrics of fairness and even more theories about fair action, the basic tenet of this principle is still universal and states that equal individuals/groups should be treated equally. The principle of safety also has a universal and relatively clear application, which is the basis for the safe use of AI. In addition, even though the application of the principle of accountability depends on the type of legal system, the principle has the ultimate and intrinsic ethical preference of regulating who will be responsible for the actions of AWS. The principles of predictability and explainability remain neutral in this regard.

> ***Unlike the principles of fairness, safety, and accountability, the concepts of predictability and explainability contain no intrinsic ethical preferences.***

Predictability and explainability are essential requirements for ensuring safe application of AWS, but they are not sufficient to achieve this goal. For example, highly predictable and clearly explainable AWS can still be used for immoral, evil purposes if, say, that weapon is in the hands of a brutal dictator, who programs AWS to commit another predictable crime against humanity. Predictability and explainability should be based on the universal ethical system that dictates the moral part of AWS behavior. Without predictability and explainability, it is extremely difficult to achieve safe, fair, and ethical use of AWS. And without a strong ethical framework and foundation, it is impossible for predictability and explainability to contribute to the creation of safe, fair, and ethical AWS.

> Without predictability and explainability, it is extremely difficult to achieve safe, fair, and ethical use of AWS. And without a strong ethical framework and foundation, it is impossible for predictability and explainability to contribute to the creation of safe, fair, and ethical AWS.

Conclusion

One of the first steps to ensure the safe and ethical use of AWS is to clearly define what the autonomy means. In addition, it is important to elucidate that there are different degrees and operational variables of predictability. And finally, for the principle of predictability to be applied ethically, it is essential to be aware of its value-free nature.

References

1. Bhuta, Nehal, and Et Al. 2016. *Autonomous Weapons Systems : Law, Ethics, Policy*. Cambridge: Cambridge University Press.
2. Davison, Neil. 2018. “A Legal Perspective: Autonomous Weapon Systems under International Humanitarian Law.” *UNODA Occasional Papers No. 30, November 2017*, January, 5–18. https://doi.org/10.18356/29a571ba-en.
3. Geist, Edward Moore. 2016. “It’s Already Too Late to Stop the AI Arms Race—We Must Manage It Instead.” *Bulletin of the Atomic Scientists* 72 (5): 318–21. https://doi.org/10.1080/00963402.2016.1216672.
4. “Debunking the AI Arms Race Theory.” 2021. Texas National Security Review. June 28, 2021. https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/.
5. Asaro, Peter. 2020. “Autonomous Weapons and the Ethics of Artificial Intelligence.” In *Ethics of Artificial Intelligence*, edited by S. Matthew Liao2. Oxford University Press.
6. “International Committee of the Red Cross (ICRC) Position on Autonomous Weapon Systems: ICRC Position and Background Paper.” n.d. International Review of the Red Cross. Accessed September 28, 2022. https://international-review.icrc.org/articles/icrc-position-on-autonomous-weapon-systems-icrc-position-and-background-paper-915#footnoteref6_u7pw2m9.
7. Boulanin, Vincent. 2016. “Mapping the Development of Autonomy in Weapon Systems: A Primer on Autonomy.” SIPRI. December 1, 2016. https://www.sipri.org/publications/2016/other-publications/mapping-development-autonomy-weapon-systems-primer-autonomy.
8. Roff, Heather M., and David Danks. 2018. “‘Trust but Verify’: The Difficulty of Trusting Autonomous Weapons Systems.” *Journal of Military Ethics* 17 (1): 2–20. https://doi.org/10.1080/15027570.2018.1481907.
9. “The Black Box, Unlocked | UNIDIR.” 2020. Unidir.org. September 22, 2020. https://unidir.org/publication/black-box-unlocked.
10. “The Black Box, Unlocked | UNIDIR.” 2020. Unidir.org. September 22, 2020. https://unidir.org/publication/black-box-unlocked.

AdalanAI

Questions?

Contact us

Adalan AI is management consulting & SaaS platform for AI Governance, Policy and Ethics. AdalanAI is about driving operational excellence, achieving greater customer satisfaction and building trust in AI product. We help corporates, VCs, investors and policy-makers in Artificial Intelligence product risk assessment and management, impact assessment, policy-tracking and policy formulation by improving/building internal governance strategies, structures, processes and people skills across organizational functions.

We are global

www.adalanai.com

ana.chubinidze@adalanai.com