Policy Brief

2023

Author
Ashley Ferreira

Implications of Explainable Artificial Intelligence to the Ethics of Automated Warfare

ABOUT THE AUTHOR

Ashley Ferreira is a Physics and Astronomy undergraduate student at the University of Waterloo. Through the co-op program she has had the chance to work in research and data science at various national labs and government departments, most recently she worked in the Asia-Pacific Branch of Global Affairs Canada and in the Center for Operational Research and Analysis at Defence Research and Development Canada. She is now at the Canadian Space Agency working part-time as a Junior Data Scientist on the Data and Emerging Technologies Team. The views expressed in this policy brief are her own and not representative of the views of these organizations.

Funding Acknowledgment

This Policy Brief was funded by the Mobilizing Insights in Defence and Security (MINDS) Targeted Engagement Grant program designed to facilitate collaboration and mobilize knowledge between the Department of National Defence, the Canadian Armed Forces, and academia and other experts on defence and security issues. Through its Targeted Engagement Grants, collaborative networks, scholarships, and expert briefings, MINDS works and collaborates with key partners to strengthen the foundation of evidence-based defence policy making. These partnerships drive innovation by encouraging new analyses of emerging global events, opportunities, and crises, while supporting a stronger defence and security dialogue with Canadians.

Introduction

Explainable Artificial Intelligence (XAI) is a rapidly evolving field that aims to make the decision-making process of AI models more understandable, transparent, and accountable. As AI systems are increasingly adopted in critical defence and security applications, it is essential to ensure that these systems are explainable and justifiable to maintain trust, enable informed decision-making, and uphold ethical standards. However, as this brief will explore, although there are many promising recent advancements in the field of XAI that could be used in military domains, there are numerous downsides and ethical concerns that must be considered before deployment.

Background on XAI

Inherently interpretable AI models exist, which are designed to be transparent from the outset. They prioritize interpretability over pure performance. For example, decision trees that represent decisions through a series of if-else rules, or linear regression models that use weighted sums of input features for predictions [1]. These types of AI models are not considered to be black boxes in the first place, and are encouraged to be tried first when expandability matters, but they often cannot handle as complex tasks as their black box model counterparts.

Deep learning models, which are very commonly used in AI, are much more complex and often considered black boxes. Below, a selection of the main approaches used in XAI to peek into this black box are outlined:
Verbal descriptions: these attempt to explain, with natural language, the decision-making process of the model. For example, an output of a model with this sort of expandability applied could be "the system identified the object as an aircraft because of its colour, its radar cross section, and its altitude and speed" [2].

Visualizations: these methods highlight the most important features on the input data to making the models output decision [2]. A popular example of this are saliency maps, a kind of heat map where pixels are colored in relation to the magnitude of their contribution to the decision making of the model [3].
Counterfactual explanations: this approach involves using example data to probe the key decision-making factors of a model [2]. A range of changes in inputs is explored to see what results in a change in the AI prediction to understand what factors are most influential in the model's decision [1].
Approximate models: the goal of these models is to create a simple and understandable model that is able to mimic the decision making process of a more complex, black box model [2]. For example, LIME (Local Interpretable Model-Agnostic Explanations), which generates explanations by creating simplified models to explain local behavior [1].

For the most part these are all model-agnostic methods, meaning that they can be applied to any AI model, regardless of its underlying architecture [1].

One particularly interesting advancement introduced a method called MILAN [4], which helps make AI systems more understandable to non-subject experts using natural language descriptions for neurons in deep learning models used in computer vision applications. These explanations can be used for a range of purposes, such as analyzing and auditing individual neurons to understand and improve the performance of an AI model.

MILAN works by identifying the most relevant parts of an image that influence an AI model's decision-making process. It then uses a dataset called MILANNOTATIONS, which consists of detailed descriptions of various image features, to generate explanations in everyday language. MILAN, much like other methods mentioned, is an important advancement in the field of XAI, but it is not a perfect solution and should be used alongside other verification methods and expert human input.
Additionally, the rise of Large Language Models (LLMs) has been accompanied by research in prompting them to provide verbal explanations of their outputs. Approaches like "tree of thought reasoning" have the model output various approaches to a small step in the task it is given, then have it pick between the options and provide an explanation for this choice before moving on to the next step in achieving the task [5]. Methods like this provide more traceability to the factors considered by the model, and they have shown that requiring a model to explicitly justify its decisions improves performance on complex reasoning tasks [5, 6].

Relevancy to the Ethics of Automated Warfare

The essay «Introduction: The Ethics of Automated Warfare and AI» [7] provides an overview of the increasing integration of AI in defence and security applications, and the ethical concerns that accompany this. The role of XAI is not explicitly explored, but can serve as a reasonable response to some of the concerns mentioned. Many of these are explored in the following section.

In the mid-2010s, the United States Defense Advanced Research Projects Agency (DARPA), formulated an XAI program with the goals of creating "a suite of machine learning techniques that: produce more explainable models, while maintaining a high level of learning performance (prediction accuracy); and enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners" [8]. This was likely the most comprehensive military initiative in XAI [9], but definitely not the last.

Recently, more regulation has been released by states regarding the use of algorithmic decision making. The European Union's new General Data Protection Regulation provides individuals with the "right to explanation" for decisions that significantly affect them [10] and it is reasonable to expect similar regulations to be adopted by various additional states. While military actions do not always fall under the same purview as normal civilian regulation, the adoption of automated warfare technologies is bound to affect many civilians, and thus the development of XAI techniques to accompany this technology is imperative.

Advantages to Use of XAI in Automated Warfare

In the context of automated warfare, XAI can provide positive contributions such as:
Transparency for legal and ethical compliance: AI systems used in military contexts must adhere to international humanitarian law and the law of armed conflict. XAI helps determine if these systems meet legal requirements and ethical standards by providing insights into the decision-making processes of AI-driven technologies.

Reducing unintended civilian harm: AI-driven military technologies have the potential to cause unintended harm to civilians. XAI can help identify and mitigate risks by revealing the factors that influence AI systems' decisions, ultimately contributing to better protection of civilians in conflict zones.
Enhancing trust and confidence: as AI applications in automated warfare become more sophisticated and integrated into military technologies, trust in these systems is vital. XAI can help build trust by offering explanations that are accessible to non-experts, allowing stakeholders to better understand AI-driven technologies and their potential impact on warfare.

Flagging points of failure and bias: the use of XAI methods can aid in flagging potential points of failure before deployment. This can also include identifying bias within AI models, which may be of special interest in cases of non-armed conflict such as the collection of intelligence information and identification of national security threats. Additionally, XAI can help mitigate against adversarial attacks, which are malicious attempts to mislead an AI model by corrupting its input data and these can easily be used by adversaries if the model is not robust against small changes to the input data [11].

An overarching point for consideration is that the decision making of humans, especially under such high pressure and fast-moving scenarios as warfare, is known to be flawed and at times only reasoned out after the fact to justify the instinctual decisions that were made [12]. The larger argument for the use of AI for these decisions, that is accompanied by well-vetted XAI techniques, are that they are more reliable, consistent, transparent, and thus significantly more correctable than human operators alone.

Disadvantages to Use of XAI in Automated Warfare

However, there are numerous downsides to the use of XAI in this context, they include:
Bad explanations: the field of XAI is still under active development, in large part because these methods need improvement before they can be confidently deployed [9]. XAI tools can be challenging to thoroughly test [9] and there are numerous cases of popular XAI tools producing incorrect or misleading explanations. Even explanations that are considered correct are often extreme simplifications of the decision-making process of the opaque AI system, and therefore not always insightful [2]. One paper in this area of XAI warns: "We believe that it is fundamentally unethical to present a simplified description of a complex system in order to increase trust if the limitations of the simplified description cannot be understood by users, and worse if the explanation is optimized to hide undesirable attributes of the system. Such explanations are inherently misleading, and may result in the user justifiably making dangerous or unfounded conclusions." [12, 13]

Overreliance on explanations: users of XAI systems may become overly reliant on the explanations provided. This overreliance could lead to a reduced sense of personal responsibility or accountability for the actions taken by the AI system, potentially blurring the lines of ethical and legal responsibility in the context of automated warfare and allow for more trust to be given to the AI systems than deserved. Increased complexity and computational overhead: implementing XAI techniques in AI systems adds complexity and computational overhead. In high-stakes applications like automated warfare, the need for real-time decision-making may conflict with the additional computational resources required to generate and process explanations. This trade-off could hinder the system's performance and response time. Interpretability-performance trade-off: inherently interpretable models used in XAI may sacrifice some performance or accuracy compared to more complex models. This limitation on a system's capabilities in situations where performance and accuracy are crucial can be difficult to justify [2]. Limited adaptability and unforeseen circumstances: XAI methods often rely on extracting explanations from past data or relying on predefined rules. In rapidly evolving and unpredictable scenarios of warfare, XAI systems may struggle to adapt to new situations or unforeseen circumstances that deviate from the training data, potentially leading to suboptimal responses. This is part of a larger issue where AI systems struggle more with out of distribution data than their classical counterparts.

Additionally, there are countless arguments against the development of technologies that advance the effectiveness of warfare, and the ability to cause harm. Even if we believe that our governments would only use acts of war in a responsible way, there is no assurance that these AI-driven technologies, which may be more readily adopted due to XAI, will not fall into the hands of governments and organizations that do not share our same values.

To leave this section on an optimistic note, there has been increasingly more exploration into developing XAI techniques that can generate provably correct explanations, which have yielded promising results with increasingly lower computational demands [14].

Recommendations

Given the advantages and disadvantages mentioned in the preceding sections, policymakers should consider the following recommendations:

Invest in XAI research: there is still a significant need for development of XAI techniques before they can reliably be deployed in high-stakes scenarios such as warfare. There is no assurance that the technology will ever mature to this point but, if there is interest in using these technologies in such a way, more research breakthroughs are needed. Funding general XAI research is an effective solution as it can benefit many public interest initiatives in addition to military applications.

Encourage the cautious deployment of military AI-systems: we still do not have a clear way of truly understanding how deep learning algorithms make their decisions and, as this brief has discussed, there are significant issues with unpredictable performance on out-of-distribution data. Although it is tempting to deploy new AI-powered innovation for defence and security applications, it would be wise to do so slowly and cautiously, especially for a country like Canada that is not actively threatened by an ongoing war. This could mean first deploying AI-systems and XAI methods on lower stakes military tasks like intelligence gathering, for example, using it to help flag anomalous information that may have been missed by an analyst. Additionally, it is important to always follow best-practices, such as leaving the final decision on any action items to a human that can be held accountable.

Establish a comprehensive governance framework: encourage international collaboration to develop a unified governance framework for XAI in defence and security applications. This framework should promote transparency, accountability, and ethical use of AI-driven technologies, while also considering mandating a certain level of explainability for any military AI-systems that could be deployed, which can help mitigate against the preemptive deployment of these systems. These discussions should also work towards defining when the burden of accountability for error falls on the operator or manufacturer of the XAI tool.

Foster interdisciplinary research and collaboration: support research initiatives that bring together AI experts, policymakers, armed forces personnel, as well as legal and military experts to develop XAI solutions and policy that adhere to international laws and ethical standards. It is important to have AI experts involved in these discussions, as they can provide an invaluable understanding of the state of the field and convey more technical aspects to the others.

Host subject matter experts: the pace of progress of XAI and AI more broadly is quite quick. If organizations hope to stay aware of developments in these areas that are important to their mandate, there is a need for personnel to monitor recent developments, evaluate AI and XAI solutions for the organization, provide education to other personnel, answer technical questions and provide relevant resources.

To conclude, in defence and security sectors, AI models have the potential to influence life-altering decisions and therefore it is crucial that these models can provide clear explanations for their predictions, to maintain effective human oversight, comply with legal and ethical requirements, and ensure accountability for the actions taken. The recommendations made in this report aim to help mitigate against the downsides of using XAI in warfare applications while benefiting from the advantages it provides.

References

1 R. Dwivedi et al., “Explainable AI (XAI): Core ideas, techniques, and solutions,” ACM Computing Surveys, vol. 55, no. 9, pp. 1–33, 2023. doi:10.1145/3561048

2 A. Holland Michel, “The Black Box, Unlocked: Predictability and Understandability in Military AI,” United Nations Institute for Disarmament Research, 2020. doi: 10.37559/SecTec/20/AI1

3 B. Subhas, “Explainable AI: Saliency Maps,” Medium, 2022. https://medium.com/@bijil.subhash/explainable-ai-saliency-maps-89098e230100

4 E. Hernandez et al., “Natural Language Descriptions of Deep Visual Features,” arXiv preprint, 2022. arXiv:2201.11114v2

5 J. Wei et al., “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models,” arXiv preprint, 2023. arXiv:2201.11903v6

6 S. Yao et al., “ReAct: Synergizing Reasoning and Acting in Language Models,” arXiv preprint, 2023. arXiv:2210.03629v3

7 B. Momani et al., “Introduction: The Ethics of Automated Warfare and AI,” Centre for International Governance Innovation, 2022. https://www.cigionline.org/articles/introduction-the-ethics-of-automated-warfare-and-ai/

8 “Explainable Artificial Intelligence (XAI) (Archived),” Defense Advanced Research Projects Agency. https://www.darpa.mil/program/explainable-artificial-intelligence

9 L. J. Luotsinen et al., “Explainable Artificial Intelligence: Exploring XAI Techniques in Military Deep Learning Applications,”Totalförsvarets forskningsinstitut, 2019. https://www.foi.se/rest-api/report/FOI-R--4849--SE%20doi

10 B. Goodman et al., “European Union regulations on algorithmic decision-making and a «right to explana-tion»,” arXiv preprint, 2016. arXiv:1606.08813v3

11 M. Frąckiewicz, “The Role of Explainable AI in Detecting and Mitigating Adversarial Attacks,” TS2 Space, 2023. https://ts2.space/en/the-role-of-explainable-ai-in-detecting-and-mitigating-adversarial-attacks/

12 M. Kovite, “I, BLACK BOX: EXPLAINABLE ARTIFICIAL INTELLIGENCE AND THE LIMITS OF HUMAN DELIBERA-TIVE PROCESSES,”War on the Rocks, 2019. https://warontherocks.com/2019/07/i-black-box-explainable-ar-tificial-intelligence-and-the-limits-of-human-deliberative-processes/

13 L. H. Gilpin et al.,“Explaining Explanations: An Overview of Interpretability of Machine Learning,”arXiv preprint, 2019. arXiv:1806.00069v3

14 S. Bassan et al.,“Towards Formal XAI: Formally Approximate Minimal Explanations of Neural Networks,” arXiv preprint, 2023. arXiv:2210.13915v2