AI and Ethics (2022) 2:631–634
https://doi.org/10.1007/s43681-021-00113-9

ORIGINAL RESEARCH

The double-edged sword of AI: Ethical Adversarial Attacks to counter artificial intelligence for crime

Michał Choraś[1,2] · **Michał Woźniak**[3]

Received: 25 June 2021 / Accepted: 13 October 2021 / Published online: 26 October 2021

Abstract
Artificial intelligence (AI) has found a myriad of applications in many domains of technology, and more importantly, in improving people's lives. Sadly, AI solutions have already been utilized for various violations and theft, even receiving the name AI or Crime (AIC). This poses a challenge: are cybersecurity experts thus justified to attack malicious AI algorithms, methods and systems as well, to stop them? Would that be fair and ethical? Furthermore, AI and machine learning algorithms are prone to be fooled or misled by the so-called adversarial attacks. However, adversarial attacks could be used by cybersecurity experts to stop the criminals using AI, and tamper with their systems. The paper argues that this kind of attacks could be named Ethical Adversarial Attacks (EAA), and if used fairly, within the regulations and legal frameworks, they would prove to be a valuable aid in the fight against cybercrime.

Keywords Artificial intelligence · Cybersecurity · Ethical Adversarial Attacks

1 Introduction

Artificial intelligence has been replacing many human activities. It has brought about a major revolution in countless domains of people's lives, such as education, Industry 4.0, data science, transport, healthcare, etc. Usually, AI solutions outperform humans in solving complex tasks of prediction, handling incomplete data, and data mining [13]. Undoubtedly, automation has many advantages, but also poses a number of threats. They do not only result from unintentional errors made by machines, which are usually the effect of improperly planned learning, but can also be caused by an intentional action. This could be done, e.g., based on the input of incorrect data in teaching collections. This particular action is called an *adversarial attack*. In other words, it consists in cybercriminals disrupting the correct machine learning process so that the trained model could be used for criminal activities, as shown in Fig. 1. Therefore, as in the game of 'paper, rocks and scissors', the AI arms race continues, to create new and better tools and methods to stop AI for Crime (AIC), and be one step ahead of cybercriminals. One of viable cybersecurity solutions could be the application of the Ethical Adversarial Attacks (EAA), the concept of which is going to be introduced in this paper.

2 Good and bad scenarios of using AI

There are both optimistic and pessimistic possible scenarios of using artificial intelligence. Given the outcomes of its possible application, AI may be seen as a double-edged sword.

2.1 AI to do good things

As widely known, nowadays AI is increasingly used in many domains of our lives to help people (e.g., to make decisions, predict, solve complex problems, etc.). There are a myriad of such applications and deployment of AI solutions (discussed in [4, 5, 12]), to name just a few). Actually, due to the broad range of applications, as well as their complexity,

✉ Michał Choraś
chorasm@utp.edu.pl; mchoras@itti.com.pl

Michał Woźniak
michal.wozniak@pwr.edu.pl

[1] Bydgoszcz University of Science and Technology, Bydgoszcz, Poland

[2] ITTI Sp. z o.o., Poznan, Poland

[3] Wroclaw University of Science and Technology, Wrocław, Poland

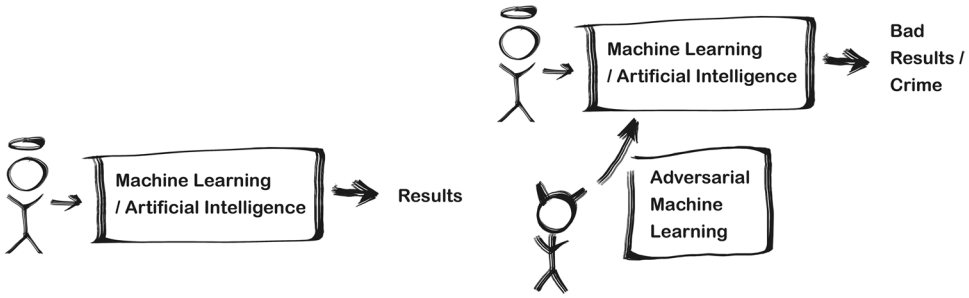

Fig. 1 The positive scenario of using AI (left) and the negative scenario of successful adversarial attacks on AI (right)

it would probably be impossible to mention all of them here. Nevertheless, AI technologies are commonly believed to be effective, reliable, created with best intentions and used to help and do good things within the framework of regulations and societal expectations.

2.2 AI designed to do bad things intentionally

Unfortunately, as with all the technologies, there is the possibility to misuse them for bad purposes. AI technologies may be utilized by criminals to enable fake news spreading, perform cyberattacks, commit computer crimes, launder money, steal data, etc. [2, 1]). The malicious use of AI has been so widespread, that the term AI for Crime (AIC) has been introduced [7].

Therefore, researchers and societies, as well as law enforcement agencies, need to be prepared for those new, modern, and sometimes unprecedented AI-supported crimes, and most importantly should be aware that such crimes have become a part of current ecosystem, especially on the internet.

One of the interesting yet alarming examples of AIC is the situation when criminals or hackers attack (or fool) normally working, legal machine learning and artificial intelligence solutions; this in turn may result in their malfunctioning. Such practices are termed as *adversarial machine learning*; several classes of such attacks on AI systems have already been distinguished, such as evasion attacks, poisoning attacks, exploratory attacks, and many more. As a result, crucial AI systems, such as those used for medical images classification or the ones applied in intelligent transport and personal cars, while attacked, could generate mistakes, faults, could be simply fooled; all this might result in doing considerable harm.

So far, such attacks have not been common yet. However, there are some theoretical advances and considerations that foresee adversarial attacks as an emerging threat. For example, it has been shown that skillfully crafted inputs can affect artificial intelligence algorithms to sway the classification results in the fashion tailored to the adversary needs [3], and that successful adversarial attacks can change the results of medical images classification or healthcare systems [8], as well as other decision support systems.

3 Cybersecurity and ethics

Here, it should be clarified why one should be concerned about the countermeasures in cybersecurity being "ethical" at all. In substance, cybersecurity is the antithesis of cybercrime. It encompasses the concepts, technologies, tools, best practices, and all the other diverse elements of the complex ecosystem the objective of which is to mitigate cyberattacks, protect people's assets, rid of vulnerabilities in systems, and so on. Yet, despite the domain being wrongly perceived as purely technical, the results of the actions (or the lack thereof) are highly likely to influence various privileges of the individual, or even infringe basic human rights [10]. Thus, ethics and ethical behaviour ought to inescapably be taken into consideration in every cybersecurity-related planning, as a way of guaranteeing the protection of people's freedom and privacy [9].

4 Should Ethical Adversarial Attacks become a conventional cybersecurity tool?

In authors' opinion, one of the most crucial domains of the research in AI and security should be devoted to countering adversarial machine learning and proposing effective detectors [11]. Even though such attacks have not been carried out 'in the wild' yet, one can expect them to occur soon. The efforts must thus be made for the cybersecurity experts to be sufficiently prepared to tackling adversarial machine learning. One of the possible countermeasures and solutions to *AIC*, apart from detection mechanisms, could be attacking the AI and ML solutions used by criminals and wrongdoers, to stop them. An example of such an attack could consist in changing the labels of fraudulent transactions so that the type is not detected by the trained fraud detection system. It should also be noted that AI, like any new technology, may fall in the wrong hands and then be used as

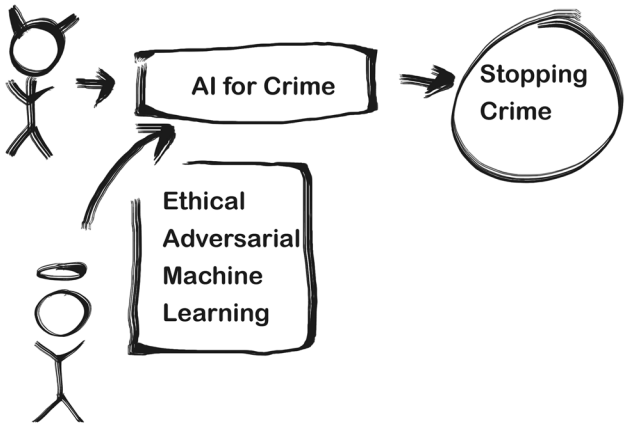

Fig. 2 The scenario of successful Ethical Adversarial Attacks (EAA) on AI for crime (AIC)

a powerful cybercrime tool. Criminals can also use AI to conceal malicious codes in benign applications or to create malware capable of mimicking trusted system components. Also, hackers can execute undetectable attacks as they blend with an organization's security environment, e.g., although TaskRabbit was hacked, compromising 3.75 million users, investigations could not trace the attack.[1] To combat hackers, AI is also used to improve computer systems security by continuous monitoring, network data analysis for intrusion detection and prevention, antivirus software, etc. Still, this approach is rather reactive, and mostly focuses on damage control.

Thus, it is worth considering whether cybersecurity experts should start resorting to an ethical method modelled on *Adversarial Attacks* to counteract the activity of criminals. Such an approach could be named Ethical Adversarial Attack, as depicted in Fig. 2.

Therefore, the authors would like to introduce the EAA concept, i.e., there is the postulate to discuss and acknowledge ethical adversarial machine learning, which would stop, fool or successfully attack AI/ML algorithms designed for malicious intentions and harming societies. Such tools and techniques should be created along with relevant legal and ethical frameworks. Even more importantly, the authors believe that the methods of this kind should be included in national and international research strategies and roadmaps. Naturally, although this might prove to be a very effective tool for fighting cybercrime, it is crucial for such AI solutions to be explainable and fair, following the xAI (explainable AI) paradigm [6]. This way, all the users and societies will be able to understand how and why EAA are applied, and that despite stemming from the tools utilized by criminals, the ethical attacks are in fact designed to do good and protect IT systems and citizens. Successful implementation of such a strategy would also mean a range of ethical issues would have to be considered. One of them would be, paraphrasing sentence from the Holy Bible *do not be overcome by evil, but overcome evil with good* (Romance 12:21), that one is not *overcome by evil, but overcomes evil with evil*. Another dilemma would concern the degree of confidentiality that would need to be preserved. On the one hand, making the results public helps other researchers in their fight against cybercrime; on the other hand, cybercriminals may use the very same results to dodge the cybersecurity measures. If the ethical questions of EAAs were properly addressed, they would also contribute to building greater trust in the solution among citizens as well as businesses and policy-makers.

[1] https://www.cisomag.com/hackers-using-ai/.

5 Conclusion

In the paper, the concept of Ethical Adversarial Attacks has been introduced. The authors have postulated to discuss EAA as the answer in the arms race against adversarial attacks or the misuse of AI systems (AI for Crime). The goal of this paper is to spark interdisciplinary discourse regarding the requirements and conditions for fair and ethical application of EAAs.

Funding This article was partially funded by Horizon 2020 Framework Programme (Grant No. 830892).

References

1. Aleksandra P, Michał C, Marek P, Rafał Kozik (2021) A $10 million question and other cybersecurity-related ethical dilemmas amid the COVID-19 pandemic. Bus Horiz 64(6):729-734 ISSN 0007-6813. https://doi.org/10.1016/j.bushor.2021.07.010 https://www.sciencedirect.com/science/article/pii/S0007681321001336
2. Caldwell, M., Andrews, J.T.A., Tanay, T., Griffin, L.D.: AI-enabled future crime. Crime Sci. **9**(1), 14 (2020). https://doi.org/10.1186/s40163-020-00123-8
3. Chakraborty A, Alam M, Dey V, Chattopadhyay A, Mukhopadhyay D (2018) Adversarial attacks and defences: a survey. arXiv:1810.00069

4. Choraś M, Pawlicki M, Kozik R (2019) The feasibility of deep learning use for adversarial model extraction in the cybersecurity domain. 353–360. https://doi.org/10.1007/978-3-030-33617-2_36
5. Earley, S.: Analytics, machine learning, and the internet of things. IT Prof. **17**(1), 10–13 (2015). https://doi.org/10.1109/MITP.2015.3
6. Gossen, F., Margaria, T., Steffen, B.: Towards explainability in machine learning: the formal methods way. IT Prof. **22**(4), 8–12 (2020). https://doi.org/10.1109/MITP.2020.3005640
7. King, T.C., Aggarwal, N., Taddeo, M., Floridi, L.: Artificial intelligence crime: an interdisciplinary analysis of foreseeable threats and solutions. Sci. Eng. Ethics **26**(1), 89–120 (2020). https://doi.org/10.1007/s11948-018-00081-0
8. Mozaffari-Kermani, M., Sur-Kolay, S., Raghunathan, A., Jha, N.K.: Systematic poisoning attacks on and defenses for machine learning in healthcare. IEEE J. Biomed. Health Inform. **19**(6), 1893–1905 (2015). https://doi.org/10.1109/JBHI.2014.2344095
9. Pawlicka, A., Choraś, M., Kozik, R., Pawlicki, M.: First broad and systematic horizon scanning campaign and study to detect societal and ethical dilemmas and emerging issues spanning over cybersecurity solutions. Personal Ubiquitous Comput. (2021). https://doi.org/10.1007/s00779-020-01510-3
10. Pawlicka, A., Choraś, M., Pawlicki, M., Kozik, R.: A $10 million question and other cybersecurity-related ethical dilemmas amid the COVID-19 pandemic. Bus Horiz. **64**(6), 729–734 (2021b). https://doi.org/10.1016/j.bushor.2021.07.010
11. Pawlicki, M., Choraś, M., Kozik, R.: Defending network intrusion detection systems against adversarial evasion attacks. Futur. Gener. Comput. Syst. **110**, 148–154 (2020). https://doi.org/10.1016/j.future.2020.04.013
12. Shekhar, H., Seal, S., Kedia, S., Guha, A.: Survey on applications of machine learning in the field of computer vision. In: Mandal, J.K., Bhattacharya, D. (eds.) Emerging Technology in Modelling and Graphics, pp. 667–678. Springer Singapore, Singapore (2020)
13. Taddeo, M., Floridi, L.: How AI can be a force for good. Science **361**(6404), 751–752 (2018). https://doi.org/10.1126/science.aat5991

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH (“Springer Nature”).
Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users (“Users”), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use (“Terms”). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.
These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.
We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.
While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval , sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.
These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.
To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.
Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.
If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com