What Is an AI Vulnerability, and Why Should We Care? Unpacking the Relationship Between AI Security and AI Ethics

Lauri Tuovinen[1,*], Kimmo Halunen[1,2]

[1]Biomimetics and Intelligent Systems Group, P.O. Box 4500, FI-90014 University of Oulu, Finland
[2]Department of Military Technology, National Defence University, P.O. Box 7, FI-00861 Helsinki, Finland

Abstract

Artificial intelligence (AI) systems are vulnerable to new types of attack such as adversarial examples and prompt injection, which cause the system to behave in unintended ways and potentially lead to harm. Taking care of the security of AI systems is therefore viewed in AI ethics as an important part of ensuring that central values such as safety or personal privacy are not jeopardised by AI systems. However, this view of AI security oversimplifies its relationship with AI ethics, as there are also situations where security may need to be traded off against another ethics requirement or where the exploitation of an AI vulnerability can be argued to be ethically justified. To provide a more nuanced view, the paper reviews the conception of security as an ethics principle in a selection of AI ethics guides and examines some notable cases where a tension exists between security and some other AI ethics principle. Furthermore, the existence of vulnerabilities in AI systems does not directly translate into harm, so it is important to distinguish theoretical scenarios involving AI vulnerabilities from their actual real-world impact. To gauge the impact, a search targeting six different incident repositories was carried out; it was observed that such efforts are hindered by the concept of AI vulnerability being vaguely defined and by the lack of a good repository that would allow exploration of AI incidents specifically involving exploitation of vulnerabilities. The search yielded only a very small number of relevant incidents, which is taken to indicate that the scale of the problem is currently small. However, it is also recognised that there is likely to be some number of incidents that the search missed, either because they were not included in the databases searched or because the search method failed to find them.

Keywords

artificial intelligence, cybersecurity, adversarial attack, AI incident, AI ethics principle, non-maleficence

1. Introduction

One of the common safety concerns regarding artificial intelligence (AI) has to do with the robustness of AI systems against adversarial attacks. In addition to traditional types of cyber-attack, AI systems may be vulnerable to ones that specifically target the AI algorithms, such as the creation of adversarial examples to induce errors in the outputs of machine learning (ML) models. For example, in 2018 it was demonstrated how ML models trained for road-sign classification could be attacked by physically altering the signs in relatively subtle ways that leave the sign legible to a human viewer but confuse the classification model [1].

7th Conference on Technology Ethics (TETHICS2024), November 6–7, 2024, Tampere, Finland
*Corresponding author.
lauri.tuovinen@oulu.fi (L. Tuovinen); kimmo.halunen@oulu.fi (K. Halunen)
0000-0002-7916-0255 (L. Tuovinen); 0000-0003-1169-5920 (K. Halunen)

An obvious example of an application where algorithmic recognition of road signs is needed is autonomous vehicles. It is similarly obvious that if the ability of a self-driving car to perform such tasks reliably can be sabotaged by tampering with objects that a malicious actor can relatively easily gain physical access to, this represents a considerable safety hazard. However, there is a difference between the demonstration of a vulnerability in academic research or organisational red teaming and exploitation of the vulnerability by a malicious actor. Especially when the mainstream media reports on discovered vulnerabilities, there is a natural tendency to focus on worst-case scenarios, but it is important to separate these from the actual real-world impact of the vulnerabilities.

A problem that emerges here is that the lack of established reporting and cataloguing practices makes it difficult to accurately gauge that real-world impact. There are various online databases collecting and publishing reports of AI vulnerabilities and incidents, but these vary considerably in terms of inclusion criteria, information stored, coverage and general quality control. Non-AI-specific cyber incident databases, on the other hand, do not provide any easy way to discover reports pertaining to AI vulnerabilities specifically.

It is generally accepted that ensuring the security of AI systems is an important part of responsible AI development and use. However, the relationship between AI security and AI ethics is by no means straightforward. One aspect of this is that the connection between AI vulnerabilities and real-world AI harm remains difficult to define while the reporting of incidents remains inconsistent and while there is not even a consensus on what exactly constitutes an AI vulnerability. For example, the possibility of bypassing the built-in safeguards of general-purpose AI (GPAI) tools such as ChatGPT to generate unethical content may be classified as a vulnerability, but from another point of view, the possibility of malicious use is part of the fundamental nature of GPAI and any safeguards that may have been implemented are just arbitrary constraints on their functionality.

Furthermore, the translation of AI ethics principles into practical requirements is also not straightforward, as they must be interpreted within a specific real-world context [2]. Security is no exception here: while one might intuitively think that stronger security measures are always unambiguously better, circumstances may in fact arise where it is necessary to weigh security against another AI ethics principle (e.g. explainability) and seek an acceptable trade-off. There are also situations where an actor exploiting a vulnerability in an AI system, while malicious from the perspective of the operator of the system, arguably has a legitimate reason for their actions (e.g. defending their privacy against mass surveillance) and is not doing anything unethical or illegal.

In this paper we explore the cluster of vaguely defined concepts at the intersection of AI security and AI ethics, aiming to at least partially answer the following questions:

- What characterises security as an AI ethics principle?
- Based on incidents recorded in public databases, what is the real-world impact of known vulnerabilities in AI systems?
- Does this picture correspond to reality?
- Under what circumstances is security in conflict with other AI ethics principles?

The remainder of the paper is organised as follows: In Section 2, we review some well-known sets of AI ethics principles and examine the conception of security emerging from them. In

Section 3, we discuss the confusion regarding the definition of AI vulnerability and establish a set of criteria for incidents where the exploitation of an AI vulnerability resulted in demonstrable real-world harm. In Section 4, we present the results of a search for such incidents in six public databases. In Section 5, we review situations where security is misaligned with other important principles. In Section 6 we discuss our findings, and in Section 7 we present our conclusions.

2. Security as an AI Ethics Principle

In [3], 84 documents containing AI ethics principles or guidelines are reviewed and analysed. The authors of the paper synthesise their findings into a list of eleven values and principles. In order of the frequency at which they occur in the various documents studied, these are transparency, justice and fairness, non-maleficence, responsibility, privacy, beneficence, freedom and autonomy, trust, dignity, sustainability, and solidarity.

Notably, security is not identified as an ethics principle in its own right. Instead, security is mentioned in the more detailed discussion in the context of two of the principles: non-maleficence and privacy. This is not the whole picture, since the documents from which the principles have been synthesised represent a range of granularities: some feature security more explicitly, while others do not even include privacy as an independent principle but subsume it under the more general principle of avoiding harm. Overall, however, it would appear that security is usually not considered an ethics principle as such but rather an enabling requirement implied by some higher-level value or principle.

A look at some well-known sets of AI ethics principles will serve to illustrate this. In the Asilomar AI Principles [4], the safety principle simply states that AI systems should be "safe and secure throughout their operational lifetime", without discussing the meaning of the terms in any detail. UNESCO's Recommendation on the Ethics of Artificial Intelligence [5] is slightly more detailed in its safety and security principle, defining safety as the avoidance of "unwanted harms" and security as the avoidance of "vulnerabilities to attack". In the Montréal Declaration for a Responsible Development of Artificial Intelligence [6], the roughly equivalent principle is the one of prudence, which states (among other things) that AI systems must "meet strict reliability, security, and integrity requirements", "not put people's lives in danger, harm their quality of life, or negatively impact their reputation or psychological integrity", and "protect the integrity and confidentiality of personal data".

The Ethically Aligned Design document of the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems [7] mentions the requirement of safety and security under its human rights principle and discusses the risk of adversarial attacks in more detail under the awareness of misuse principle. In the Ethics Guidelines for Trustworthy AI proposed by the European Commission's High-Level Expert Group on Artificial Intelligence (AI HLEG) [8], the principle of prevention of harm states that AI systems "and the environments in which they operate must be safe and secure", and also that they must be "technically robust and it should be ensured that they are not open to malicious use". The AI4People framework [9], drawing on several of the above, similarly subsumes security under its non-maleficence principle.

It is not hard to come up with examples of how adversarial attacks against AI systems may compromise values other than safety and privacy. For instance, by poisoning the training data

of an AI system, an attacker could induce biases into its decision-making algorithms, causing it to violate the principle of fairness. It therefore makes sense to view security as a necessary-but-not-sufficient prerequisite to several – perhaps even all – ethics principles rather than an ethics principle per se. Some of the collections of ethics principles explicitly acknowledge the possibility of data poisoning or other AI-specific attacks such as "gaming", a term used in [7] to refer to exploiting the learned behaviour of an AI system to trigger responses not intended by the deployer.

Something that is not necessarily mentioned explicitly but is at least implicit in discussions of security as an AI ethics principle is that they view security primarily as an obligation that the deployer of an AI system has to those affected by the system. The deployer, in contrast, would traditionally view security first and foremost through the lens of protecting the confidentiality, integrity and availability of its own assets. Often the interests of different stakeholders are mutually aligned or at least compatible, but not always; some examples of the latter are discussed in Section 5.

3. AI Vulnerabilities and Incidents

The current state of AI incident documentation is reviewed in [10], where four online repositories of AI incidents are examined: the AI, Algorithmic, and Automation Incidents and Controversies Repository (AIAAIC) [11], the AI Incident Database (AIID) [12], the AI Vulnerability Database (AVID) [13] and Where in the World is AI? [14]. Among these, AIAAIC, AIID and AVID present a traditional tabular view, while Where in the World presents a map view with the locations of incidents marked with different colours representing different application domains. The latter also differs from the others in that it collects both harmful and helpful cases, whereas the others focus on harmful ones.

The only one among the four repositories that explicitly aims to document AI vulnerabilities is AVID, which makes a distinction between *vulnerabilities* and *reports*. A vulnerability is defined as "a high-level evidence of an AI failure mode", whereas a report is "one example of a particular vulnerability occurring, supported by qualitative or quantitative evaluation". AVID reports are thus equivalent to what are termed *incidents* by AIAAIC and AIID and *cases* by Where in the World. On the face of it, AVID would seem to be the most promising option for finding cases of real-world harm resulting specifically from the exploitation of an AI vulnerability, but in practice this is not the case, for several reasons:

- The distinction between vulnerabilities and reports is not consistently observed; many of the entries listed as vulnerabilities describe individual examples and should therefore be classified as reports. In some cases there is one entry listed as a vulnerability and another, identical entry listed as a report (e.g. entries AVID-2023-V025 and AVID-2023-R0001, respectively).
- The quality of the entries is likewise inconsistent. In extreme cases, the details section of the entry contains only some placeholder text (AVID-2022-V003) or text referring to a completely different case, apparently as a result of copy-and-paste (AVID-2023-V022).
- The definition of vulnerability is very broad, resulting in the inclusion of failure modes whose relevance to security seems tenuous, such as ChatGPT failing to follow lexical

constraints in user prompts (AVID-2023-V025). On the other hand, the database also includes entries describing uses of AI that are ethically questionable but arguably do not represent a failure mode, such as the generation and distribution of a deepfaked video of the president of Ukraine (AVID-2022-V009).
- Despite the loose inclusion criteria, the coverage of the database is modest, with fewer than 50 entries (vulnerabilities and reports) in total at the time of writing. AIID covers some 650 incidents and AIAAIC, the largest one of the four, some 1400, although it is worth noting that the latter is not limited to incidents involving AI.

Regarding the third point, it should be noted that whether or not a failure mode can be adversarially exploited depends ultimately on the context of use. Given the vast range of potential uses of GPAI systems, we should be cautious about categorically declaring that a given failure mode does not have any security implications, even if the idea of it being exploited seems far-fetched. However, if any observed failure mode may or may not represent a vulnerability depending on context, this suggests that looking for AI vulnerabilities without considering the context of use is not very fruitful in the first place.

In addition to the four databases mentioned in [10], it is worth looking into generic databases of cybersecurity vulnerabilities and incidents. In the United States, the Cybersecurity and Infrastructure Security Agency (CISA) maintains the Known Exploited Vulnerabilities (KEV) catalogue [15], and in Europe, the European Repository of Cyber Incidents (EuRepoC) maintains the EuRepoC database [16]. These have the advantage of being more pertinent to vulnerabilities as the term is normally understood in the security domain, but also the major disadvantage that the vast majority of the database entries are not related to AI vulnerabilities, and there is no easy way to filter these out to reveal those entries that are relevant to the question at hand.

There seems to be no single database that could give a straightforward and satisfactory answer to the question of the real-world impact of AI vulnerabilities. However, each of the six databases mentioned above could be reasonably expected to shed some light on the question, if searched using a suitable approach. To design such an approach, it is first necessary to specify precise criteria for entries considered relevant. These are as follows:

- **Incident:** A relevant entry must be based on a reliable report of a real-world event. Hypothetical scenarios and unconfirmed allegations are not considered relevant.
- **Intrusion:** A relevant entry must involve deliberate exploitation of a weakness in the system. Instances where the system fails without the involvement of an intruder are not considered relevant.
- **Impact:** A relevant entry must involve an incident where demonstrable real-world harm was inflicted. Theoretically possible consequences not proven in practice are not considered relevant.
- **Intelligence:** A relevant entry must involve a vulnerability specifically in an AI component of the system. Instances where a traditional vulnerability in an AI system is exploited are not considered relevant.

A search of the databases to identify entries that satisfy the criteria was carried out in November 2023. The search methodology and the results of the search are described in the next section.

4. Real-World Impact of AI Vulnerabilities

To narrow down the list of potentially relevant entries, different search approaches were applied to different databases. For AIAAIC and AIID, a list of search terms that could indicate malicious activity (e.g. “adversarial”, “attack”. “breach”, “exploit”, “hack”) was compiled and entries containing one or more of the terms on the list were flagged as candidates. For KEV and EuRepoC, a similar approach was used, but the search terms used were ones that could indicate that an AI/ML system was targeted, including general terms related to such systems (e.g. “model”) as well as more specific ones such as “poison” (for data poisoning, a type of attack against ML systems) and “prompt” (for prompt injection, a type of attack against systems based on language models).

For AVID, since the number of entries was manageable and since there was (again, on the face of it) reason to expect a high ratio of relevant to irrelevant entries, all entries in the database were treated as candidates. Where in the World does not provide a search function, so a filter was applied to hide the cases where AI was helpful and the remaining cases were treated as candidates. All of the candidates in all six databases were then scanned manually and evaluated against the criteria specified above to classify each candidate as fully relevant, partially relevant, possibly relevant or irrelevant. As a result of this process, six incidents were identified that were deemed to fully satisfy the relevance criteria. These are shown in Table 1; the columns of the table correspond to the criteria, providing a summary of what happened, what kind of intrusion was involved, what were the consequences and what was the role of AI in the incident.

In addition to the 6 fully relevant ones, there were 27 incidents classified as partially relevant, indicating that they were considered to satisfy three of the criteria fully and the fourth partially. The incidents in this category were mostly demonstrations of vulnerabilities by non-malicious actors, but there were also some cases where a vulnerability was exploited with malicious intent but either the harm inflicted was unclear or the role of AI was marginal. 8 incidents were classified as possibly relevant, indicating that they could be relevant but the source database did not provide sufficient details for a definitive verdict. For example, Where in the World would sometimes only provide a link to an article hidden behind a paywall, leaving the case ambiguous.

The number of fully relevant incidents is too small to permit any meaningful analysis, but if we include the partially relevant and possibly relevant incidents, a few notable themes emerge:

- **Deception of biometric identification systems:** Basic facial recognition models may be deceived by something as simple as a 2D photograph, whereas more advanced ones can be attacked by crafting a special mask. Identical twins have been reported to deceive both facial and voice recognition systems.
- **Evasion / manipulation of filtering algorithms:** Targeted algorithms include malware detection algorithms of cybersecurity software suites, spam filters and content moderation algorithms of social media platforms.
- **Manipulation of conversational AI:** In the archetypal case, the user circumvents the restrictions built into the system, causing it to generate unsafe content. In some incidents, prompt injection was used to achieved remote code execution on the host computer.

Table 1
Summary of Fully Relevant AI Incidents

Incident	Intrusion	Impact	Intelligence
Twitter bot Tay poisoned	Trolls deliberately sought to trigger inflammatory responses	Twitter users exposed to hate speech	Bot used AI to learn from interactions with human users
Facebook fact checks evaded during COVID pandemic	Deliberate small changes made to posts to deceive filtering algorithms	Facebook users exposed to COVID-19 disinformation	Facebook uses AI to detect posts that violate terms and conditions
Twitter bot remoteli.io tricked into carrying out arbitrary instructions	Prompt injection vulnerability exploited to hijack bot	Bot behaving in ways not intended by the owner	Bot used language model to generate responses to messages
TikTok content moderation evaded by suicide video	Massive coordinated uploading of different versions of the video	TikTok users exposed to traumatising content	TikTok uses AI to detect videos that violate terms and conditions
Chinese local government tax system compromised	Camera hijack attack used to gain access to system	$77 million acquired through fraudulent invoices	System used facial recognition for access control
PyTorch dependency chain compromised	Dependency replaced with binary containing malicious code	Malicious binary downloaded by unknown number of PyPI users	PyTorch is a widely used Python library for ML applications

- **Manipulation / hijacking of cyber-physical systems:** Included in this category are various attacks on autonomous vehicles, but also some that target customer service robots, allowing passive spying through the robot's sensors or taking active control of its functions.

While these may give some indication of the types of AI vulnerabilities and attacks that are likely to result in real-world harm, the number of incidents is still small, even with the partially and possibly relevant ones included. In Section 6, we will discuss some possible reasons for the small number of relevant incidents found in the search, as well as some possible biases in the incidents that were found.

5. When Security and Ethics Clash

In ethics, it is not uncommon to face a situation where it is necessary to make a value-based trade-off between conflicting requirements, and AI ethics is no exception. For example, explainability of the decisions made by an AI system is widely considered an important ethics requirement, as demonstrated by transparency being the most frequently occurring AI ethics principle in the documents reviewed in [3]. However, as discussed in [17], a high level of explainability may be

difficult to achieve if a high level of accuracy is also required, although recent empirical work has challenged the traditional view that there is some kind of inverse relationship between the two [18, 19].

In comparison with e.g. explainability or privacy, security is not an aspect of AI ethics that figures very prominently in discussions of value trade-offs. However, as discussed in [20], the use of explainable AI (XAI) methods makes AI systems more susceptible to attacks that aim to degrade their performance or to infer sensitive information. Arguably this is not so much a trade-off as a reminder that with XAI, particular care should be taken to defend the system against such attacks; nevertheless, it is important to recognise that even though security is instrumental by nature and often not considered an ethics principle as such, it can be at odds with ethics principles and is not simply an enabler.

The fact that different stakeholders may have conflicting, yet legitimate – or at least not unambiguously illegitimate – interests regarding an AI system introduces further nuances into the relationship between security and ethics. Similarly to what the authors of [1] did to road sign recognition, facial-recognition systems can also be attacked through the use of physical objects. An obvious method is to simply hide one's face to such a degree that there is not enough information available for the system to reliably identify the person, but there are also methods that leave the face in view and target the specific weaknesses of facial-recognition models instead. Examples include 3D printing of spectacle frames [21] and placing of stickers on a hat [22] or directly on the face [23].

In the broader societal and ethical discourse, methods for thwarting facial-recognition systems are typically viewed in the context of surveillance systems in public places and the age-old debate where public safety is pitted against civil liberties. On the one hand, deliberately confusing a system adopted and operated by legitimate authorities under democratic supervision could be construed as a malicious act. On the other hand, if the purpose of the disguise is not to commit a criminal act, the individual can argue that they are enforcing their fundamental right to privacy – in a sense, opting out of data collection to which they do not consent.

A debate that has emerged much more recently has to do with the legal and ethical justification of training generative AI models with creative works scraped from the internet without licensing them from the copyright holders. Related to this, an interesting new development is Nightshade, a special type of training data poisoning attack targeting text-to-image models proposed in [24]. The authors of the paper explicitly discuss their method as a tool that content creators could use to protect their intellectual property and rectify the power asymmetry between themselves and the developers of AI models, who face no consequences from using scraping algorithms that do not comply with voluntary opt-out requests.

Legally, this appears to be a grey area at the moment, but it is at least possible that using copyright-protected works as training material for AI models is covered by exemptions such as the fair use doctrine in United States copyright law [25]. However, regardless of how courts of law rule on the issue or how lawmakers choose to approach it, creators may still see it as their right to opt out of their works being harvested and to use tools such as Nightshade to enforce that right if necessary. Certainly even if scraping is considered lawful, this does not imply that creators are under any obligation to cooperate with scrapers, so we can view this case as another example of how the relationship between AI security and AI ethics is not as straightforward as it might first seem.

6. Discussion

What explains the small number of AI incidents found that even partially or possibly satisfy the "4I" relevance criteria defined in Section 3? We have no evidence at the moment that would enable us to say anything conclusive, but we can identify a number of possible factors:

1. **Exploitation of AI vulnerabilities does not happen.** While it is unlikely that the discovered incidents represent the full picture, it seems plausible that compared to traditional vulnerabilities, the real-world impact of exploitation of AI vulnerabilities by malicious actors is still small. This does not mean, of course, that the impact will not grow, possibly even very rapidly.
2. **Exploitation does happen, but does not get reported.** There may be some number of incidents that would have satisfied the relevance criteria but could not be found because no report was made of them. This could be the case if, for example, the targeted organisation decided to cover up the incident to protect its reputation.
3. **Exploitation gets reported, but reports do not get included in the databases.** There may be some number of relevant incidents that were reported but could not be found because the reports are deposited somewhere other than the databases we searched. Such reports could reside, for example, in the internal databases of various security organisations or even on public websites.
4. **Reports get included, but the search method used fails to find them.** Finally, it is possible that some number of relevant incidents stored in the databases were not found because we failed to identify the right keywords or because the whole keyword-based search approach was flawed. Such an approach can only be successful if cyber incidents involving AI are consistently described using language that makes the nature of the incident clear, which is not necessarily the case.

For the time being, we are operating on the assumption that each of these contributed to some extent to the small number of search results. Further exploration of this question is a matter for future work; we will remark, however, that because of the variety of factors that may cause relevant AI incidents to be overlooked by the search, the search results are likely to be biased in various ways. For example, the relative prevalence of reports involving self-driving cars is arguably not so much an indicator of their particular vulnerability to adversarial attacks as an artefact of the considerable attention they attract in both academic research and the mainstream media. We should therefore be wary of drawing even very general conclusions about what the search results can tell us.

Cybercrime has become a well-organised industry with a range of business models [26], and its annual cost is estimated to be in the order of ten trillion USD globally [27]. Based on the available evidence, it would appear that exploitation of AI vulnerabilities does not yet account for any appreciable fraction of this, and it may well be that effective models for generating revenue through AI-specific attacks have yet to emerge, as opposed to more traditional approaches such as ransomware attacks. However, given the magnitude of the financial incentive, cybercriminals will undoubtedly seek to capitalise on the new opportunities created by the proliferation of AI systems, so it is important to monitor the situation, and to develop better instruments of monitoring if the currently existing ones prove inadequate.

Notably, neither of the two non-AI-specific databases, KEV and EuRepoC, yielded any incidents deemed (partially/possibly) relevant according to the 4I criteria. This may be at least partially because of item number 4 in the list above – i.e. there are some relevant reports in those databases but the search failed to find them – but it is also possible that (some) AI vulnerabilities are not even recognised as vulnerabilities in the traditional sense, and are therefore not being included in traditional cybersecurity repositories. Given this, combined with the elusive nature of the concept of AI vulnerability and the inconsistency of AI incident reporting practices, we argue that there is a need for more dialogue between the security and ethics communities on this topic. Without proper understanding and cataloguing of AI vulnerabilities, it will not be possible to understand the true scale and nature of their real-world impact.

7. Conclusion

In this paper we examined the security of AI systems from the perspective of AI ethics, aiming to clarify the relationship between the two and to highlight some of its nuances. By studying research literature and some well-known sets of proposed AI ethics principles, we found that while security is primarily viewed as an enabler for higher-level ethics requirements such as safety, in reality the situation is more complicated because of various trade-offs and value conflicts involving AI security. The security of AI systems therefore cannot be properly understood without considering the real-world sociotechnical context in which they are deployed.

We focused particularly on assessing the scale and nature of real-world harm resulting from the exploitation of AI security vulnerabilities, finding this difficult because of shortcomings in the way vulnerabilities and incidents involving AI are currently being recorded in public repositories. We defined a set of relevance criteria and carried out a search of four AI incident repositories and two cybersecurity incident repositories, resulting in only about 40 incidents that could be considered at least partially relevant to the research question. We concluded that while the real-world impact of AI vulnerabilities probably still is comparatively small, the set of incidents discovered by the search is likely to suffer from biases induced by a number of factors. Obtaining a more accurate picture requires further research as well as cross-community discourse between divergent perspectives on AI security.

Acknowledgments

The research reported in this paper was carried out with funding awarded by the Scientific Advisory Board for Defence (MATINE). We would also like to acknowledge the helpful suggestions of our colleague Arttu Pispa during the preparation of the manuscript.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, D. Song, Robust physical-world attacks on deep learning visual classification, in: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 1625–1634.

[2] B. Mittelstadt, Principles alone cannot guarantee ethical AI, Nature Machine Intelligence 1 (2019) 501–507.

[3] A. Jobin, M. Ienca, E. Vayena, The global landscape of AI ethics guidelines, Nature Machine Intelligence 1 (2019) 389–399.

[4] Future of Life Institute, Asilomar AI principles, 2017. URL: https://futureoflife.org/open-letter/ai-principles/, accessed April 3, 2024.

[5] UNESCO, Recommendation on the ethics of artificial intelligence, 2021. URL: https://unesdoc.unesco.org/ark:/48223/pf0000381137, accessed April 3, 2024.

[6] Université de Montréal, Montréal declaration for a responsible development of artificial intelligence, 2018. URL: https://montrealdeclaration-responsibleai.com/the-declaration/, accessed April 3, 2024.

[7] IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems, version 2, 2017. URL: https://standards.ieee.org/industry-connections/ec/ead-v1/, accessed April 3, 2024.

[8] High-Level Expert Group on Artificial Intelligence, Ethics guidelines for trustworthy AI, 2019. URL: https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1, accessed April 3, 2024.

[9] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, E. Vayena, AI4People—an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations, Minds and Machines 28 (2018) 689–707.

[10] V. Turri, R. Dzombak, Why we need to know more: Exploring the state of AI incident documentation practices, in: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, 2023, pp. 576–583.

[11] AIAAIC repository, 2024. URL: https://www.aiaaic.org/home, accessed March 20, 2024.

[12] AI incident database, 2024. URL: https://incidentdatabase.ai/, accessed March 20, 2024.

[13] AI vulnerability database, 2024. URL: https://avidml.org/, accessed March 20, 2024.

[14] Where in the world is AI?, 2024. URL: https://map.ai-global.org/, accessed March 20, 2024.

[15] Known exploited vulnerabilities catalog, 2024. URL: https://www.cisa.gov/known-exploited-vulnerabilities-catalog, accessed March 20, 2024.

[16] EuRepoC database, 2024. URL: https://eurepoc.eu/database/, accessed March 20, 2024.

[17] P. P. Angelov, E. A. Soares, R. Jiang, N. I. Arnold, P. M. Atkinson, Explainable artificial intelligence: an analytical review, WIREs Data Mining and Knowledge Discovery 11 (2021) e1424.

[18] A. Bell, I. Solano-Kamaiko, O. Nov, J. Stoyanovich, It's just not that simple: An empirical study of the accuracy-explainability trade-off in machine learning for public policy, in: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency,

2022, pp. 248––266.
[19] L.-V. Herm, K. Heinrich, J. Wanner, C. Janiesch, Stop ordering machine learning algorithms by their explainability! a user-centered investigation of performance and explainability, International Journal of Information Management 69 (2023) 102538.
[20] C. N. Spartalis, T. Semertzidis, P. Daras, Balancing XAI with privacy and security considerations, in: Computer Security. ESORICS 2023 International Workshops, 2024, pp. 111–124.
[21] M. Sharif, S. Bhagavatula, L. Bauer, M. K. Reiter, Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition, in: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 2016, pp. 1528––1540.
[22] S. Komkov, A. Petiushko, AdvHat: Real-world adversarial attack on ArcFace face id system, in: 2020 25th International Conference on Pattern Recognition (ICPR), 2021, pp. 819–826.
[23] M. Shen, H. Yu, L. Zhu, K. Xu, Q. Li, J. Hu, Effective and robust physical-world attacks on deep learning face recognition systems, IEEE Transactions on Information Forensics and Security 16 (2021) 4063–4077.
[24] S. Shan, W. Ding, J. Passananti, S. Wu, H. Zheng, B. Y. Zhao, Prompt-specific poisoning attacks on text-to-image generative models, 2024. arXiv:2310.13828.
[25] P. Samuelson, Generative AI meets copyright, Science 381 (2023) 158–161.
[26] C. Griffy-Brown, D. Lazarikos, M. Chun, Cybercrime business models: Developing an approach for effective security against better organized criminals, Journal of Applied Business and Economics 19 (2017).
[27] S. Morgan, Cybercrime to cost the world $9.5 trillion USD annually in 2024, 2023. URL: https://cybersecurityventures.com/cybercrime-to-cost-the-world-9-trillion-annually-in-2024/, accessed April 11, 2024.