AI and Ethics
https://doi.org/10.1007/s43681-022-00162-8

ORIGINAL RESEARCH

A framework for assessing AI ethics with applications to cybersecurity

Danilo Bruschi[1] · **Nicla Diomede**[1]

Received: 14 October 2021 / Accepted: 13 April 2022

Abstract
In the last few years many scholars, public and private organizations have been involved in the definition of guidelines and frameworks for individuating the principles to adopt in the development and deployment of AI systems. Some authors, however, noted that the effectiveness of these guidelines or ethical codes on the developer's community is very marginal. One of the obstacles that opposes to the effective implementation of ethical principles is the lack of an approach for solving tensions which arise when principles are applied. A possible solution to such an issue could be the adoption of a risk-based approach which is also advocated by many sources. To our knowledge, no concrete proposals have been presented in literature on how to perform a risk-based ethical assessment. In this paper we contribute to close this gap by introducing a framework based on a qualitative risk analysis approach for assessing the ethical impact underneath the introduction of an innovation either technological or organizational in a system. We will also show how the framework can be used for individuating suitable safeguards to adopt for balancing potential ethical infringements that the innovation may entail once implemented. Some case studies in the cybersecurity context are also described for showing the effectiveness of our approach.

Keywords Artificial intelligence ethical aspects · Cybersecurity · Risk assessment · Ethical frameworks

1 Introduction

In the last few years many scholars, public and private organizations have been involved in the definition of guidelines and frameworks for individuating the principles to adopt in the development and deployment of AI systems, see for example [1]. To date, at least 84 AI ethics initiatives have published their own reports, each proposing its own set of ethical principles and values [14]. Fortunately, some authors [6, 8] noted that there is a substantial overlap between the different sets of principles proposed, and suggested to condense them in a set of 4, 5 principles that outline the fundamental traits of an AI application. These principles are [4]: respect of human autonomy, prevention of harm, fairness, and explicability.

In McNamara et al. [13] it has however been shown that so far, the effectiveness of all the above-mentioned efforts on the developer's community is almost zero. Starting from this observation, scholars started to investigate weaknesses and forthcomings of the AI ethics initiatives. Among them: [8, 19] have underlined the mere marketing aspect of many of them, [9, 14] the lack of enforcement mechanisms reaching beyond a voluntary and non-binding cooperation between ethicists and individuals, the same authors also questioned whether the principled approach, adopted in most of the above-mentioned studies, is the more adequate for AI ethics. This approach successfully applied in fields such as medicine and bioethics does not seem suitable to AI. A virtue ethic approach in building AI ethics frameworks instead of a principled one is advocated in Hagendorff [10].

In the analysis of the different ethical frameworks most of the authors individuated a common obstacle that opposes to the effective implementation of ethical principles: the solution of tensions[1] which arise when principles must be applied to AI systems. That is when AI applications are developed conflicting prescriptions of principles will emerge and no methods have been devised for solving them, leaving the solution of the trade-offs to the developers, who often do not have the knowledge for solving them and simply ignore

✉ Danilo Bruschi
danilo.bruschi@unimi.it

Nicla Diomede
nicla.diomede@unimi.it

[1] Dipartimento di Informatica, Università degli Studi di Milano, Milan, Italy

[1] Here by tension, we mean a conflict between two important values or principles where it appears necessary to give up one to realise the other.

Published online: 18 May 2022

them. Resolving these conflicts is where the real work starts for AI ethics [14].

When a trade-off between two values emerges, a choice must be made to prioritise one set of values over another and a useful method for supporting the choice would be to estimate the risk[2] associated to both values and decide based on the outputs. That is, an ethical risk assessment methodology must be adopted. Such a methodology should provide the means for determining the risk associated with the infringement of an ethical principle, given the likelihood that such an infringement occurs and the "losses" it can cause.

A risk-based approach for solving tensions between ethical principles is advocated in [6] and it is solicited by the recent field of algorithms auditing where it is required for estimating the ethical risks underneath the adoption of a given algorithm/solution, and consequently decide on its adoption and in a such a case the counterbalance measures to adopt it [11, 17]. To our knowledge, no concrete proposals have been presented in literature on how an ethical assessment could be performed.

This paper contains a contribution for addressing such a problem by introducing a framework for ethical risk assessment which can be concretely adopted. More precisely when considering an innovation which must be introduced in a system, our framework enables researchers to correlate the potential level of infringement of some ethical principle caused by the innovation with the risk that this innovation entails. For example, suppose we are considering enforcing an authentication system with an AI identity management system for reducing the risk of impersonation attacks, in this case we should consider that a reduction of the risk of our initial system, will imply an infringement of users' privacy. Our framework will enable decision makers to correlate such a risk reduction with the level of infringement of the privacy principle.

Once decision makers have decided to adopt an innovation despite the violation of some ethical principle, safeguards could restore the right balance. Our framework can also be used to individuate these safeguards.

To show the applicability of the proposed methodology we will show two case studies in which our framework is used for estimating the ethical impact of some AI innovations applied to cybersecurity systems.

The paper proceeds as follows. After a brief introduction in Sect. 2 on the state of the art of the AI ethics fields and an overview of its main strength and weaknesses, in Sect. 3 we underline the importance of a risk assessment approach in the AI ethics field. In Sect. 4 we introduce our risk assessment framework. In Sect. 5 we will apply our framework to a couple of case studies related to the introduction of AI innovations in cybersecurity systems. Section 6 is devoted to conclusions.

[2] We briefly recall that the risk is an unwanted outcome resulting from an incident, event, or occurrence, and it is given by the product of a loss by the probability of the incident occurring.

2 AI ethics

In the last few years scholars, government organizations and "big tech" all around the world developed a whole body of ethical guidelines or principles for driving the development and deployment of AI applications. The approach followed has been that of principalism. Even if some of these initiatives raised some criticisms [8, 20] since they have been mostly seen as marketing initiative, or an attempt by private sector institutions to avoid legislation or the creation of binding legal norms, they contributed to individuate and collect a very conspicuous set of requirements and principles. Cowls and Floridi [2] assessing five documents, found 44 AI principles addressed while [10] analyzing 22 guidelines extracted 22 different ethical requirements. Starting from Cowls and Floridi [2] some work has been done for exploiting the substantial overlap between the different sets of principles. The main result in such a direction has probably been the work of the European Commission's High Level Expert Group on artificial intelligence which proposed four principles [4, 7] which are briefly summarized in the following.

- *Respect of human autonomy*: AI systems should not unjustifiably subordinate, coerce, deceive, manipulate, condition or herd humans. Instead, they should be designed to augment, complement, and empower human cognitive, social, and cultural skills. The allocation of functions between humans and AI systems should follow human-centric design principles and leave meaningful opportunity for human choice.
- *Prevention of harm*: AI systems and the environments in which they operate must be safe and secure. They must be technically robust, and it should be ensured that they are not open to malicious use. Vulnerable persons should receive greater attention and be included in the development, deployment, and use of AI systems
- *Fairness*: considered in both a substantive and a procedural dimension. The substantive dimension implies a commitment to ensure equal and just distribution of both benefits and costs and ensure that individuals and groups are free from unfair bias, discrimination, and stigmatisation. The procedural dimension of fairness entails the ability to contest and seek effective redress against decisions made by AI systems and by the humans operating them.
- *Explicability*: the capabilities and purpose of AI systems need to be openly communicated, and decisions – to the

extent possible – explainable to those directly and indirectly affected.

2.1 On the difficulty of translating ethical principles in practice

The main objective of defining ethical frameworks or guideline for AI systems is that of affecting and influencing the development and deployment of AI products. The study reported in McNamara et al. [13] however showed that still there is a long way to go before to reach this objective. Based on a controlled study, it turned out that the effectiveness of guidelines or ethical codes on the developers' community is almost zero, and that they have no influence on the behavior of professionals from the tech community. This motivated some authors to investigate the weaknesses and shortcomings which characterize AI ethics programs and make it difficult to use them in the design and governance of AI products. In [9, 10, 12, 21] two main veins of criticism emerged. The first one is related to the adoption of a principalism approach for the definition of AI ethical frameworks. Following the approach successfully adopted by more traditional disciplines such as medicine and bioethics, most of the AI ethical guidelines have been compiled following a principalism approach. AI however lacks enforcement mechanisms, the tradition, the experiences and the commitment to public services that characterize these disciplines and could be less suitable to a principled approach (for a more detailed analysis see [14]), while an approach based on virtualism could be more fruitful [10].

A second vein of criticism is well described in Whittlestone [21] where the tensions which can arise between different ethical issues when designing AI systems are described. Tensions usually emerge during the development of a system, and a solid ethical as well as political backgrounds are required to solve them. Developers usually do not have these competences, thus either a methodology is provided to them for solving this issue or simply they will ignore them.

2.2 AI and cybersecurity

Experts agree in considering Cybersecurity as one of the most promising field of application of AI/ML techniques. The main reason for such a general believe is that current cybersecurity solutions will soon be unable to effectively monitor all the internet traffic and timely detect attack vectors: IoT, cloud and 5G will generate levels of internet traffic in terms of data volume, velocity, and variety which are becoming increasingly difficult to analyze [22]. In such a scenario, organizations and society in general face an urgent need to improve their defense strategy with a major emphasis on key performance indicators such as the average time to detect/respond/resolve an incident. AI techniques are the best candidate for carrying out such a role. They are seen as the enabling element for implementing a proactive cybersecurity approach [3] supporting organizations in preventing computer attacks.

It is well known that many cybersecurity measures may contribute to infringe some civil rights for example monitoring devices such next generation firewall and intrusion detection systems which inspect internet traffic (not only metadata) to detect the presence of malware, attack patterns of malicious contents, can be easily transformed in surveillance tools for intercepting communication among peers. The dichotomy between cybersecurity and privacy is probable the most representative example of the conflict which characterize security with respect to other fundamental rights.

To harness the "disruptive" potentials of new AI applications researchers as well as government organizations all around the world developed a whole body of ethical guidelines or principles to which technology developers should adhere to as soon as possible. To apply these concepts to cybersecurity mechanisms is not easy, as often tensions showed up between the various principles which should be satisfied, and a risk-based approach must be undertaken as will be shown in the following sections.

3 On the importance of ethical assessment

A standard approach for dealing with tradeoffs in most scientific discipline is to recur to a risk assessment methodology, i.e., a process which enables to assign a value (either quantitative or qualitative) to a certain event by considering the combination of the likelihood of the event and its severity. In our specific case an ethical risk assessment methodology should provide the means for determining the risk associated with the infringement of an ethical principle, given the likelihood that such an infringement occurs and the "losses" which can be determined by it.

The necessity of a such a methodology is solicited by many sources: the European Commission's White Paper on Artificial Intelligence [5] calls for as risk-based approach to the adoption of AI systems, the recent field of Algorithms Auditing has underlined the need for a risk assessment process for determining whether to adopt a given solution, or the counterbalance measures in case of its adoption [11, 17]. Floridi [8] underlines the necessity that the debate on AI ethics evolves from the *what* to the *how*: not just what ethics is needed but also how ethics can be effectively and successfully applied and implemented. That is principles must be put into practice. In Hangerdorff [9] it is observed that ethics must partially transform to "microethics". This means that at certain points, a substantial change in the level of abstraction must happen insofar as ethics aims to have a certain impact

Table 1 A risk assessment matrix which reports on the rows the likelihood of an event and on the columns the estimated losses, any colored item represents the level of risk arising from the event for given likelihood and losses

	LOSS SEVERITY	Negligible	Minor	Moderate	Major	Catastrophic
EVENT LIKELYHOOD	Very likely to happen	MODERATE	HIGH	HIGH	VERY HIGH	VERY HIGH
	Likely to happen	LOW	MODERATE	HIGH	HIGH	VERY HIGH
	Possible to happen	LOW	LOW	MODERATE	HIGH	HIGH
	Unlikely to happen	VERY LOW	LOW	LOW	MODERATE	HIGH
	Very unlikely to happen	VERY LOW	VERY LOW	LOW	LOW	MODERATE

Table 2 Ethical risk assessment matrix, a three dimensions matrix which on the third dimension reports the level of infringement of some ethical principle

	LOSS SEVERITY	Negligible			Minor			Moderate			Major			Catastrophic		
	Infringement severity	N	S	VS	N	S	VS	N	S	VS	N	S	VS	N	S	VS
EVENT LIKELIHOOD	Very likely to happen	M	L	L	H	M	L	H	M	L	VH	H	M	VH	H	M
	Likely to happen	L	VL	VL	M	L	L	H	M	L	H	M	L	VH	H	M
	Possible to happen	L	VL	VL	L	VL	VL	M	L	L	H	M	L	H	M	L
	Unlikely to happen	VL	VL	VL	L	VL	VL	L	VL	VL	M	L	L	H	M	L
	Very unlikely to happen	VL	VL	VL	VL	VL	VL	L	VL	VL	L	VL	VL	M	L	L

VL stands for very low, *L* stands for low, *M* stands for moderate, *H* stands for high and *VH* stands for very high, *N* stands for none, *S* stands for significant, *VS* stands for very significant

and influence in the technical disciplines and the practice of research and development of artificial intelligence. As long as ethicists refrain from doing so, they will remain visible in a general public, but not in professional communities.

In the next section we introduce an ethical risk analysis methodology, which is also an example of micro ethical work which can be implemented easily and concretely in practice.

4 A framework for ethical assessment

In this section we introduce a framework for correlating the risk associated by the introduction of an innovation in a system, with the level of infringement of an ethical principle which can be caused by the innovation itself. For example, if an authentication system is under consideration, and we want to reduce its risk to be compromised from HIGH to LOW our framework will enable decision makers to ascertain that in this case the level of infringement of the privacy principle will be VERY SIGNIFICANT (see Sect. 5 for more details). On the other hand, if we want that the level of infringement of the ethical principle be at most SIGNIFICANT, then the risk of our authentication system cannot be lower than MODERATE.

Once such a choice has been done, the framework can also be adopted to individuate the safeguards to balance potential infringements. That is, once it has been chosen to deploy a solution with a SIGNIFICANT level of ethical infringement, the framework may support decision makers to individuate the most appropriate safeguards to balance such an infringement.

Table 3 A risk infringement matrix which reports on the rows the level of robustness of the safeguards which can be adopted, on the columns the severity of the infringement which must be managed

	Infringement severity	Negligible	Significant	Very Significant
Safeguard robustness	None	Negligible	Significant	Very Significant
	Medium	Negligible	Negligible	Significant
	High	Negligible	Negligible	Negligible

Any colored item represents the resulting level of infringement obtained when a given safeguard is applied to reduce an infringement of a certain severity

The main component of our framework is the ethical risk assessment matrix (see Table 2) which is a standard risk assessment matrix augmented with a further dimension for correlating the ethical risk underneath a system or a process.

We briefly recall that a risk assessment matrix (see Table 1) is a table representing the risk associated to a system. In our case, the rows represent the probabilities of occurrence of some event occurring and are represented on a scale of 5 values (VERY UNLIKELY TO HAPPEN, LIKELY TO HAPPEN, POSSIBLE, UNLIKELY TO HAPPEN, VERY UNLIKELY TO HAPPEN). The columns represent the qualitative values of estimated losses should the above-mentioned event happens. Even these values are represented in a scale of 5 (NEGLIGIBLE, MINOR, MODERATE, MAJOR, CATASTROPHIC).[3] Any element of the matrix represents the value of the risk (VERY LOW, LOW, MODERATE, HIGH, VERY HIGH) which is produced by the combination of the likelihood of an event with the losses caused by the event itself.

Using the risk assessment matrix, we can easily evaluate the overall risk of a system in the following way. Suppose the system we are considering is characterized by a probability of malfunctioning VERY LIKELY TO HAPPEN and that it has been estimated that in such a case the losses have been estimated MODERATE. This implies that the risk related to such a malfunctioning is HIGH.

Assigning the correct values to columns and rows is the most complex task which requires to undertake specific contextual considerations and it is usually performed by agents with deep knowledge and experience in the field under consideration. For example, for estimating the columns values i.e., for individuating the losses related to a malfunctioning we need to take care of the impact of the incident on the assets of the organization (people, information, infrastructures, services, reputation, etc.) as well as on the unquantifiable assets (quality of life, health, freedom, etc.).

The rows of the matrix show the likelihood that some event happens. Even in this case specific context sensitive evaluation must be undertaken such as considering the "quality" of the information treated by the system, its application field, its impact on human behavior, etc.

Suppose we have estimated that the overall risk for an authentication system is HIGH. We may be interested in reducing such a risk from HIGH to MODERATE by introducing a more sophisticated tool for biometric authentication. Such a tool however will have an impact on privacy, and we are interested in estimating the "gains" in security against the "losses" in privacy. Such a correlation can be exploited by the ethical risk assessment matrix reported in Table 2 where a further dimension has been added to a standard risk assessment matrix. Such a dimension is used for representing the different levels of infringements to an ethical principle that the introduction of an innovation in a system may involve. In our specific case, we assume that a measure/event can have three different level of ethical impact on a system: NONE no impact, SIGNIFICANT the measure will have a significant impact and VERY SIGNIFICANT. The ethical risk assessment matrix can be used as follows, given a system characterized by a LIKELY TO HAPPEN event likelihood and a MAJOR loss, for reducing the overall risk of the system from HIGH to MODERATE we need to adopt a measure with a SIGNIFICANT IMPACT on some ethical principle.

Once an organization has decided to adopt a measure which compromises some principle, safeguards might be adopted to regain a balance. Even this process can be driven with the support of a matrix representation. To this aim we introduce the notion of a risk infringement matrix (see Table 3). Such a matrix has on the rows the robustness of safeguards which can be adopted for balancing an infringement, which for ease of exposition we represent on a scale of three values (namely None, Medium, High) and on the columns the severity of the infringement (None, Significant, Very Significant). Any element of the matrix represents the resulting level of infringement after the application of a given safeguard measure.

[3] The dimensions as well as the values contained in the risk assessment matrix are purely indicative and can be changed as needed.

Table 4 Example of ethical risk assessment matrix for analyzing the ethical impact of an AI identity management system, the abbreviations are the same as those adopted in Table 2

	Loss severity	Negligible			Minor			Moderate			Major			Catastrophic		
Incident likelihood	Infringement severity	N	S	VS	N	S	VS	N	S	VS	N	S	VS	N	S	VS
	Very likely to happen	M	L	L	H	M	L	H	M	L	VH	H	M	VH	H	M
	Likely to happen	L	VL	VL	M	L	L	H	M	L	H	M	L	VH	H	M
	Possible to happen	L	VL	VL	L	VL	VL	M	L	L	H	M	L	H	M	L
	Unlikely to happen	VL	VL	VL	L	VL	VL	L	VL	VL	M	L	L	H	M	L
	Very unlikely to happen	VL	VL	VL	VL	VL	VL	L	VL	VL	L	VL	VL	M	L	L

Table 5 Infringement matrix adopted for individuating the most suitable measures to adopt for counterbalancing the infringement of the privacy principle caused by the adoption of cybersecurity measures aimed at improving the quality of the authentication system

	Infringement severity	Negligible	Significant	Very significant
Safeguards robustness	None	Negligible	Significant	Very significant
	GDPR compliance with adequate technical and organization measures	Negligible	Negligible	Significant
	Further strong security mechanisms to protect Data and System Access; frequent audit; Virtue Ethics	Negligible	Negligible	Negligible

5 Ethical assessment of cybersecurity application

It is a general believe that current cybersecurity solutions will soon be unable to effectively monitor all the internet traffic and timely detect attack vectors: IoT, cloud and 5G will generate levels of internet traffic in terms of data volume, velocity, and variety which are becoming increasingly difficult to analyze [22]. In such a scenario, organizations face an urgent need to improve their defense strategy. AI techniques are the best candidate for such a role. On the other hand, it is well known that many cybersecurity measures may contribute to infringe some rights and that the unconstrained development of AI applications can lead to situations in conflict with fundamental human rights. Thus, the coupling between cybersecurity and AI needs to be scrutinized. In this section we show how this kind of analysis can be carried out using the framework previously introduced.

5.1 Identity and access management systems

AI systems can be successfully used to reduce errors during the authentication phase or equivalently to reduce the rate of impersonation attacks. In these cases, AI/ML algorithms needs a lot of data for working properly and this data is obtained by profiling human activities. As noted in [19] these systems may contribute to improve system resilience by tracking and collecting "sensor data and human-device interaction from your app/website. Every touch event, device motion, or mouse gesture is collected" thus leading to create a mass-surveillance effect.

It turns out that a tension arises between the privacy and prevention of harm principles. Our framework can support decision makers in solving such a tension in the following way. Using specific contextual considerations, the ethical risk assessment matrix presented in Table 3 is built considering the ethical impacts that an AI identity management system may have on the privacy principle as a function of the features enabled. NONE: personal profiling data is not collected. SIGNIFICANT: biometric data is collected. VERY SIGNIFICANT: both biometric and behavioral data is collected (Table 4).

Looking at this matrix, if at the end of the risk analysis phase we estimated that our system is characterized by a MODERATE loss, but the incident likelihood is VERY LIKELY TO HAPPEN, we know that the risk is HIGH and it can be reduced to MODERATE by tuning our identity management system to a SIGNIFICANT infringement mode of use. On the other hand, if our system is characterized by a MINOR loss with an incident likelihood which is LIKELY TO HAPPEN, it is not necessary to tune our mechanism with a very aggressive modality as the same performance can be obtained with a "milder" option.

Once we decided the "tuning" of our identity management system, the infringement matrix reported in Table 5 tells us that once we have chosen a SIGNIFICANT level of infringement of our ethical principle if we want to get

Table 6 Infringement matrix adopted for individuating the most suitable measures to adopt for counterbalancing the infringement of the human autonomy principle caused by the adoption of cybersecurity measures aimed at improving the efficiency of a security operation center

	Infringement severity	Negligible	Significant	Very significant
Safeguards robustness	None	Negligible	Significant	Very significant
	Measures are in place for monitoring and accountability of the behavior of the AI system	Negligible	Negligible	Significant
	Stricter governance, extensive testing and measures are in place for the human control when it is necessary	Negligible	Negligible	Negligible

a NEGLIGIBLE one, we need to take into consideration the introduction of cybersecurity countermeasures as those defined by the GDPR.

5.2 Cognitive SOC

Security operation centers (SOC) are infrastructures aimed at detecting/responding in real-time to computer attacks. Currently, in SOCs the experts are struggling in trying to take pace with the rate of alerts they receive daily. On average a SOC receive 11.000 alerts per day, nearly 20% of them are manually inspected/triaged, 28% are ignored and they can give rise to attack that go unnoticed for weeks or even months, compromising the organization infrastructure [16]. Given the scale and speed of threats in a near future, the only viable solution for containing the number of computer attacks is the adoption of threat intelligence automation. That is delegate to AI algorithms the management of most alerts, leaving to the human experts the most critical cases. This however raises a tension between the principles of respect of human autonomy and prevention of harm.

We can use our framework to assess the impact of threat intelligence automation. For brevity reason, we assume that the risk assessment matrix we obtain by our analysis is the same we get in the previous example and which is reported in Table 4. It turns out that if the system that the SOC must protect is characterized by a MAJOR loss with an incident probability which is LIKELY TO HAPPEN, the only way we have for reducing the risk of the system, if this is our risk management decision, is to tune the SOC to work with a VERY SIGNIFICANT impact on the respect of human autonomy principle. If we are interested in counter balancing with suitable safeguards our decision, the infringement matrix reported in Table 6 tells us the adequate measures to put in place. Where, by monitoring measures we mean that a process must be in place to verify the correct behavior of the AI System with ability to override a decision when it is necessary.

6 Conclusions

When applying AI ethics in the development of systems, tensions among principles inevitably arise. The solution of these tensions requires political choices which consider both the material and immaterial values involved. A risk–benefit analysis can be very useful for supporting the decision makers to undertake the right decision and is advocated by many sources. The adoption of a risk-based approach to ethical assessment can be greatly simplified by the adoption of a methodology or framework which will drive the researcher in such a process. The framework presented in this paper represents a first step in such a direction [1, 7, 12, 15, 19].

Funding Open access funding provided by Università degli Studi di Milano within the CRUI-CARE Agreement. No funding was received for conducting this study.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interest to disclose.

References

1. AI Now: AI Now Report. AI Now Institute (2019). https://ainowinstitute.org/AI_Now_2019_Report.pdf. Retrieved June 2020

2. Cowls, J., Floridi, L.: Prolegomena to a white paper on an ethical framework for a good AI society. SSRN J (2018). https://doi.org/10.2139/ssrn.3198732
3. CRAE: Cybersecurity resource allocation and efficacy index (2020). https://www.cyberriskalliance.com/wp-content/uploads/2020/08/CRAE-Index.pdf. Retrieved July 2020
4. EU: High level expert group on artificial intelligence. Ethics guidelines for trustworthy AI (2019). https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf
5. EU: White paper: on artificial intelligence—a European approach to excellence and trust (2020). https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf. Retrieved June 2020
6. EU1: European parliament resolution of 20 October 2020 with recommendations to the commission on a framework of ethical aspects of artificial intelligence, robotics and related technologies (2020/2012(INL)) (2020). https://www.europarl.europa.eu/doceo/document/TA-9-2020-0275_EN.pdf. Retrieved October 2020
7. EU: European Commission, Proposal for a Regulation laying down harmonised rules on artificial intelligence, 2021/0106 (COD) and Annexes (2021). https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence
8. Floridi, L.: Translating principles into practices of digital ethics: five risks of being unethical. Philos. Technol. **32**(2), 185–193 (2019)
9. Hagendorff, T.: The ethics of AI ethics: an evaluation of guidelines. Mind. Mach. (2020). https://doi.org/10.1007/s11023-020-09517-8
10. Hagendorff, T.: AI virtues—the missing link in putting AI ethics into practice (2021). https://arxiv.org/abs/2011.12750
11. Koshiyama, A., Kazim, E., Treleaven, P., Rai, P., Szpruch, L., Pavey, G., Ahamat, G., Leutner, F., Goebel, R., Knight, A., Adams, J., Hitrova, C., Barnett, J., Nachev, P., Barber, D., Chamorro-Premuzic, T., Klemmer, K., Gregorovic, M., Khan, S., Lomas, E.: Towards algorithm auditing: a survey on managing legal, ethical and technological risks of AI, ML and associated algorithms (2021). https://doi.org/10.2139/ssrn.3778998
12. Lo Piano, S.: Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward (2020). https://www.nature.com/articles/s41599-020-0501-9. Retrieved July 2020
13. McNamara, A., Smith, J., Murphy-Hill, E.: Does ACM's code of ethics change ethical decision making in software development. In: Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering - ESEC/FSE 2018. New York, ACM Press, pp. 1–7 (2018)
14. Mittelstadt, B. Principles alone cannot guarantee ethical AI. Nat Mach Intell 1, 501–507 (2019). https://doi.org/10.1038/s42256-019-0114-4
15. Nadeem Javaid N., Sher A., Nasir H., Guizani N.: (2018) Intelligence in IoT-based 5G networks: opportunities and challenges. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8493126
16. Palo A.: The 2020 state of security operations (2020). https://start.paloaltonetworks.com/forrester-2020-state-of-secops.html. Retrieved June 2020
17. Raji, I., Smart, A., White, R., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., Barnes, P.: Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing, pp 33–44 (2020). https://doi.org/10.1145/3351095.3372873.
18. Simon Y., Zhibin H., Chun Y., Donghwan L., Myung K., Dong S.: HARMer: cyber-attacks automation and evaluation (2020). https://arxiv.org/abs/2006.14352, Retrieved July 2020
19. Taddeo, M.: Three ethical challenges of applications of artificial intelligence in cybersecurity. Minds Mach. (2019)
20. Wagner, B.: Ethics as an escape from regulation: from ethics-washing to ethics-shopping? In: Mireille Hildebrandt (Ed.): Bein Profiled. Cogitas ergo sum. Amsterdam University Press, Amsterdam, pp. 84–89 (2018).
21. Whittlestone, J., Nyrup, R., Alexandrova, A., Cave, S.: The Role and limits of principles in AI ethics: towards a focus on tensions (2019). https://doi.org/10.1145/3306618.3314289.
22. Zhang, J., Huang, T., Wuang, S., Liu, T.: Future internet: trends and challenges. Front. Inf. Technol. Electron. Eng. (2019). https://doi.org/10.1631/FITEE.1800445