AI and Ethics (2021) 1:49–53
https://doi.org/10.1007/s43681-020-00012-5

OPINION PAPER

A choices framework for the responsible use of AI

Richard Benjamins[1]

Received: 1 September 2020 / Accepted: 3 September 2020 / Published online: 6 October 2020

Abstract
Popular press and media often make us believe that artificial intelligence technology is ethical or unethical by itself. In this paper, we will argue that organizations that develop or apply AI have certain choices they can make that will lead to a more or less responsible use of AI. By approaching those choices in a methodological way, organizations can make better decisions toward the ethical use of this powerful technology.

1 Introduction

In spite of what popular press and media wants us to believe about the ethical and societal impact of artificial intelligence, until AI becomes more intelligent than recognizing patterns in large amounts of data, humans will remain responsible for preventing or mitigating the potential negative impacts of AI.

The use of artificial intelligence in all aspects of societies continues to grow, and for the moment, there are no signs that this will change in the near future. With this increasing uptake, there are also increasingly more examples of negative consequences of this powerful technology, and this is precisely one of the main motivations for the existence of Springer's AI Ethics Journal.

Like many scholars, I believe that there are many more positive opportunities thanks to AI, than the implied risks often associated with AI. But this does not mean that we should continue to develop and apply AI without critically thinking about its potential negative consequences or harm. The focus of this paper is on avoiding or mitigating the unintended, negative consequences of AI while the intention is a good use of AI, e.g., to solve difficult problems for health, society, business, climate, etc. For this, the paper proposes a framework that identifies relevant choices to be made when developing or using AI. The focus of this article is not on avoiding malicious uses of AI. We distinguish between two types of choices: choices related to what AI principle organizations adhere to as discussed in [1] and choices related to how to technically articulate those principles. The contribution of this paper is related to the second part (in the context of AI of today, that is, data-driven AI mostly based on machine learning and in particular on deep learning).

2 Choices of principles for the responsible use of artificial intelligence

In the past few years, a proliferation has occurred of AI ethical principles that are supposed to guide organizations in avoiding the negative consequences of AI. [2] analyses the AI principles of 36 organizations in 9 categories (human rights, human values, responsibility, human control, fairness and non-discrimination, transparency and explainability, safety and security, accountability, and privacy). The non-profit organization Algorithm Watch maintains an open inventory of AI guidelines with currently over 160 organizations [3]. It is, therefore, not easy for organizations to decide what AI principles to include. In [1], we provide three criteria that organizations can use for choosing the AI principles appropriate for their business and vision.

1. Distinguish between, on the one hand, principles relevant for governments, such as the future of work, lethal autonomous weapon systems, liability, concentration of power and wealth, and, on the other hand, principles that individual organizations can act on, such as privacy, security, fairness and transparency.
2. Distinguish between intended and unintended consequences. Many challenges of the use of AI are occurring as an unintended side effect of the technology (e.g. bias,

✉ Richard Benjamins
richard.benjamins@telefonica.com;
richard.benjamins@odiseia.org

[1] Telefonica and OdiseIA, Madrid, Spain

lack of explainability, future of work). Intended consequences are explicit decisions that can be controlled, such as using AI for good or for bad. Organizations better formulate their principles for the unintended consequences they can act upon.

3. Consider whether the AI principles should cover all aspects relevant for AI systems (e.g. safety, privacy, security, fairness, etc.) in an end-to-end manner, versus covering only AI-specific challenges (e.g. fairness, explainability, human agency).

Apart from those guiding criteria, the decisions organizations take should also take into account the sector they are operating in. For example, using AI in the aviation sector will put high value on safety, whereas the insurance sector will need to put high value on fairness, and the medical sector on explainability.

3 Technical choices for the responsible use of artificial intelligence

Apart from the choice of what AI principles to select to ensure positive social and ethical impact of the AI used by the organization, there are technical choices that determine or influence the social and ethical impact of the use of AI. As seen in the previous section, we can distinguish between AI-specific technical choices and generic digital technical choices. For instance, privacy is relevant for any digital system dealing with personal data. The eight technical choices are illustrated in Fig. 1.

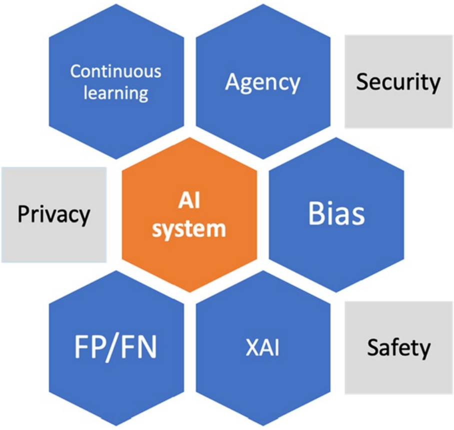


Fig. 1 Technical choices for the ethical and social impact of AI. AI-specific choices in blue. Generic digital choices in grey (colour figure online)

3.1 AI-specific technical choices

Continuous learning refers to the fact that the AI system's machine learning algorithm continues to learn autonomously from data that becomes available once the AI system is in production. This means that the performance of the AI system evolves over time without human intervention as new data becomes available. This is in contrast to AI systems whose production algorithms are updated periodically by AI engineers, with subsequent new versions and releases in the market. An example of possible continuous learning systems are self-driving cars that can learn from new data that comes available from other cars connected to the same back-end system. An example of a (non-autonomous) continuous learning system could be a churn prediction system that is monthly updated with new data from churners and loyal customers overseen by engineers. The main societal impact of this technical choice is related to liability and accountability. If autonomously continuous learning AI systems make errors that cause damage to citizens or organizations, the question arises of who is responsible: the creator of the AI system, the deployer, or the AI system itself? Currently there is an ongoing discussion about whether liability should be with the producer or the deployer [4]. It should, however, be clear that this decision has potentially large impact on the societal and ethical impacts of the AI system. As with any of the technical choices, the correct decision depends much on the type of application, whether it is a critical system in healthcare of transport or a more harmless systems such as movie recommendation.

Human **agency** refers to the degree that humans remain in control of the outcomes of the AI system (regardless of whether or not the system continuous to learn autonomously). The basic choices available are

- Human-in-the-loop (HITL), which means that the AI system may suggest decisions, but there is always a person taking the final decision.
- Human-on-the-loop (HOTL), which means that the AI system takes decisions by itself but there is always a person overseeing the results and intervening in case incorrect decisions are detected.
- Human-out-of-the-loop (HOOTL), which means that the system takes decisions by itself without any human intervention or oversight. Of course, in case of serious errors, the affected persons or organizations might have to possibility for redress, but they have to request it explicitly.

If not selected adequately, a wrong choice for a particular use case might have important negative consequences of the AI system. For instance, a killer drone should never use HOOTL, and always HITL. HOTL might be more acceptable for automated acceptance or rejection of financial loans.

Bias might lead to undesired or even illegal discrimination. An AI system might result in discrimination for a number of reasons, but it is always related to so-called sensitive variables which represent protected groups.

- If the training data set is not representative for the target audience, some protected groups might be discriminated against. For instance, an AI system trained on school performance data from rich neighbourhoods applied to a complete city (including poorer neighbourhoods) might result in discriminatory results for children of some ethnic origins.
- If the training data set contains sensitive variables as defined by law (e.g. Article 9 of the GDPR, [5], including: racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation), the AI system might learn to discriminate by one or several of those sensitive variables, which is against the law.
- Even if the training data set does not contain sensitive data, it might still contain so-called "proxy" variables that correlate highly with one of the sensitive variables. A typical example is the significant correlation between ethnic origin and postal code in some US cities.

Explainability refers to the possibility of understanding how an AI system comes to its conclusion [6]. There are black box machine learning algorithms such as deep learning, which are hard to understand for people, and there are white box algorithms such as decision trees that people are able to understand. Deep learning algorithms usually result in better performance. The choice for using a black box or white box algorithm, again, depends on the type of application. Critical systems in the health domain often require explainability, otherwise professionals are not comfortable with using the results. Entertainment applications usually require less explainable systems. The challenge here is to find the right balance between performance and explainability.

False positives and false negatives All data-driven AI systems have a certain performance but never reaches 100% accuracy; there is always an error rate. There is no universal answer to the question of what an acceptable error rate is. Indeed, it is up to domain experts to decide whether a 3% error rate is acceptable (e.g. in medical diagnosis) or a 25% (e.g. in movie recommendation). In this respect, a certain AI system might be acceptable or not in a particular domain. There is, however, an additional aspect to error rates of AI systems that may have an ethical or societal impact. Errors can be false positives (e.g., an AI system diagnoses a person as having a disease, but in reality, the person doesn't have that disease) or false negatives (e.g., an AI system diagnoses a person as not having the disease, but in reality the person has the disease). During their development process, AI systems can be optimized by reducing false positives, false negatives or both. For minimizing the negative social and ethical consequences of AI, it is important to consider the specifics of the domain when trying to reduce the error rate. In some domains, a false-negative causes much more harm than a false positive and vice versa. Therefore, the choice of how to optimize the error rate of an AI system potentially has important consequences depending on the domain.

3.2 Generic technical choices

Privacy, security and **safety** are all important aspects to consider when using AI systems, but they are not specific to AI. Any digital system that operates in the real-world needs to respect privacy in case it deals with personal data, be secure (difficult to hack) and be safe (not causing physical harm). For example, specific choices about how privacy is implemented and respected in AI systems (such as how to obtain explicit informed consent for using personal data) may have important societal and ethical impact. Indeed, the uncontrolled use of excessive personal data in the digital advertising industry has given rise to the GDPR and is still an area of significant debate.

4 Analysing specific cases

In this section, we look at specific cases of unintended negative consequences of AI systems and explain how they could be avoided using our "choices" framework.

4.1 COMPAS—assessment of recidivism

COMPAS is the, by now, well-known risk assessment system that helped US judges in their assessment of the likelihood of recidivism of a defendant. Pro-Publica investigated the COMPAS system and found that black defendants were far more likely than white defendants to be incorrectly judged of higher risk of recidivism [7]. This specific discriminatory aspect is a consequence of the (lack of) choice of the false-positive and false-negative rates for black and white defendants. It turned out that the false-positive rate was much higher for blacks than for whites, while the false-negative rate was higher for white defendants. The fact that race—a sensitive variable—is included in the data set implies that the machine learning algorithm can learn to discriminate on race, which is illegal.

There are other important choices for the COMPAS application. Deciding to keep a defendant in custody or releasing him/her until the date of court, based on the risk of recidivism has a large impact on people's life, and, therefore, explainability of the decision as well as limited autonomy of the system (agency) are important.

4.2 Amazon—hiring new employees

In 2014, Amazon started to use an AI system to support the hiring of new employees. But already in 2015, it found out that the system was not rating candidates for technical posts in a gender-neutral way [8]. The system discriminated against women, and the company decided to remove the system from production. In this case, the technical choice not given enough consideration was related to the bias in the data set used for training the machine learning algorithm. The data set consisted of CVs the company had received over a 10-year period and the algorithm learned patterns that led to successful hires. Given the fact that most technical jobs (especially years ago) are performed by men, the algorithm downgraded CVs containing words related to women. The problem here was that the training data set was not gender balanced, and this led the AI system to identify patterns that disfavoured women. Amazon stated that the system was never used to autonomously select candidates, which seems to be a correct decision given the impact on people's lives of hiring decisions.

4.3 Dutch court prohibits Government's use of AI software to detect welfare fraud

On February 5, 2020, the District Court of The Hague held that the System Risk Indication (SyRI) algorithm system, a legal instrument that the Dutch government uses to detect fraud in areas such as benefits, allowances, and taxes, violated article 8 of the European Convention on Human Rights (ECHR) (right to respect for private and family life) [9]. The system combined several governmental data bases to detect suspicious patterns without being transparent to the citizens involved and without asking them for consent. There are many choices that need to be reviewed before such a system can be put into operation, but the main reason why the Dutch court ruled it illegal was the lack of transparency of how the system reached its conclusion. In other words, explainability was the critical aspect in the court's view. Accusing someone of fraud is a major thing, and, therefore, it seems fair to require the system to be explainable. Agency is another important aspect. Indeed, the SyRI system "only" highlighted suspicious cases for further investigation, and never accused people directly of fraud. There is also an element of bias in this case. Social welfare is usually for the poorer parts of society, and, therefore, poorer citizens are object of being analysed by algorithms without their consent, while the richer citizens are not subject of such automation.

There are many other examples of AI systems that have been featured in the media because they did things in way that should have been avoided. For example, image recognition systems that produce better results for white people than coloured people, or gender bias in automatic translation, etc. All of those examples can be analysed in terms of the eight choices to be considered before launching an AI system. And if done properly, the risk of the AI system producing undesired behaviour can be significantly reduced.

Avoiding those unintended, negative consequences of AI is exactly the objective of methodologies for the responsible use of AI such as described in [10].

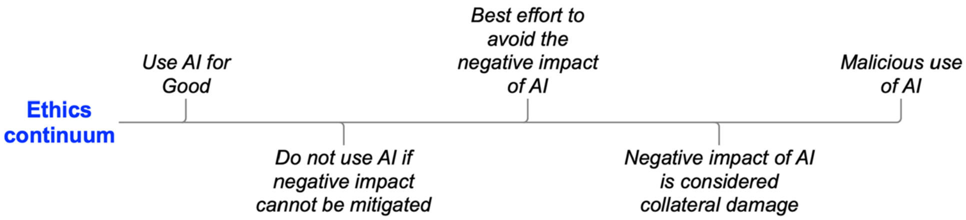
5 Summary and discussion

Artificial intelligence is "good", but organizations need to manage (minimize and mitigate) the unintended potential negative consequences. In this paper, we have presented a "choices" framework that allows to understand and act—through explicit choices—on the negative consequences of AI before they actually happen. AI-specific technical choices include bias, explainability, agency, errors (FP, FN) and continuous autonomous learning. Generic digital technical choices include privacy, security and safety. We also briefly discussed how to support organizations in choosing AI principles appropriate for their business. An important conclusion of this work is that AI by itself has no negative impact, but it is how it is configured and applied to solve specific (business) problems, and these are voluntary choices that organization have and should decide on.

However, not all unintended negative impacts of AI are due to the choices mentioned in this paper. Sometimes, it is the business model that leads to unintended negative consequences. For instance, the viral propagation of (fake) news through Facebook and its potential creation of tunnel vision, are consequences of Facebook's advertising business model which is driven by user engagement. The more users interact with the platform (click, like, share, comment, etc.), the more Facebook earns. Users tend to engage more with content of their preference, and this drives Facebook's AI algorithms to present that type of content to those users, increasing the risk of tunnel vision.

Even though organizations can be technically aware of all relevant choices to make to avoid and mitigate the negative implications of AI, it doesn't mean they will. But this has nothing to do with AI nor with technical solutions: it has to do with organizational values and norms. Generally speaking, we can think of a continuum from good to bad—an

Fig. 2 Ethics continuum of how AI can impact society


Ethics continuum
Use AI for Good
Do not use AI if negative impact cannot be mitigated
Best effort to avoid the negative impact of AI
Negative impact of AI is considered collateral damage
Malicious use of AI

ethics continuum (Fig. 2)—and it is up to each organization to recognize and/or decide where it fits or wants to be on this continuum.

We think it is a healthy exercise for organizations to understand and recognize where they sit on this continuum today and where they aspire to be.

References

1. Benjamins, R.: Towards organizational guidelines for the responsible use of AI. In 24th European Conference on Artificial Intelligence - ECAI 2020, https://ecai2020.eu/papers/1347_paper.pdf (2020). Accessed 20 Aug 2020
2. Fjeld, J., Achten, N., Hilligoss, H.. Nagy, A., Srikumar, M.: Principled artificial intelligence: mapping consensus in ethical and rights-based approaches to principles for AI. 2020. [Online]. Available at https://dash.harvard.edu/handle/1/42160420. Accessed 20 Aug 2020
3. Algorithm Watch, "AI Ethics Guidelines Global Inventory," 2020. [Online]. Available: https://inventory.algorithmwatch.org/. Accessed 20 Aug 2020
4. Committee on Legal Affairs, European Commission: DRAFT REPORT with recommendations to the Commission on a Civil liability regime for artificial intelligence. 2020. [Online]. Available at: https://www.europarl.europa.eu/doceo/document/JURI-PR-650556_EN.pdf. Accessed 20 Aug 2020
5. European Commission: Processing of special categories of personal data," intersoft consulting, [Online]. Available at https://gdpr-info.eu/art-9-gdpr/. Accessed 20 Aug 2020
6. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F.: Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inform Fus **58**, 82–115 (2020)
7. Larson, J., Mattu, S., Kirchner, L., Angwin, J.: How we analyzed the COMPAS recidivism algorithm," 2016. [Online]. Available at https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. Accessed 20 Aug 2020
8. Dastin, J.: Amazon scraps secret AI recruiting tool that showed bias against women," 2018. [Online]. Available at https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G. Accessed 20 Aug 2020
9. Gesley, J.: Netherlands: curt prohibits government's use of AI software to detect welfare fraud," 2020. [Online]. Available at https://www.loc.gov/law/foreign-news/article/netherlands-court-prohibits-governments-use-of-ai-software-to-detect-welfare-fraud/. Accessed 20 Aug 2020
10. Benjamins, R., Barbado, A., Sierra, D.: Responsible AI by design in practice. In Proceedings of the Human-Centered AI: Trustworthiness of AI Models and Data (HAI) track at AAAI Fall Symposium, Washington DC, 2019

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.