Risk Analysis Technique for the Evaluation of AI Technologies with Respect to Directly and Indirectly Affected Entities

Joachim Iden
TÜV Rheinland Japan Ltd., Osaka, Japan

Felix Zwarg[1]
TÜV Rheinland Industrie Service GmbH, Köln, Germany

Bouthaina Abdou
TÜV Rheinland Industrie Service GmbH, Köln, Germany

Abstract

AI technologies are often described as being transformative to society. In fact, their impact is multifaceted, with both local and global effects which may be of a direct or indirect nature. Effects can stem from both the intended use of the technology and its unintentional side effects. Potentially affected entities include natural or juridical persons, groups of persons, as well as society as a whole, the economy and the natural environment. There are a number of different roles which characterise the relationship with a specific AI technology, including manufacturer, provider, voluntary user, involuntarily affected person, government, regulatory authority, and certification body. For each role, specific properties must be identified and evaluated for relevance, including ethics-related properties like privacy, fairness, human rights and human autonomy as well as engineering-related properties such as performance, reliability, safety and security. As for any other technology, there are identifiable lifecycle phases of the deployment of an AI technology, including specification, design, implementation, operation, maintenance and decommissioning. In this paper we will argue that all of these phases must be considered systematically in order to reveal both direct and indirect costs and effects to allow an objective judgment of a specific AI technology. In the past, costs caused by one party but incurred by another (so-called ‚externalities') have often been overlooked or deliberately obscured. Our approach is intended to help remedy this. We therefore discuss possible impact mechanisms represented by keywords such as resources, materials, energy, data, communication, transportation, employment and social interaction in order to identify possible causal paths. For the purpose of the analysis, we distinguish degrees of stakeholder involvement in order to support the identification of those causal paths which are not immediately obvious.

2012 ACM Subject Classification Social and professional topics → Computing / technology policy; Social and professional topics → Computing and business; Software and its engineering → Risk management

Keywords and phrases AI, Risk Analysis, Risk Management, AI assessment

Digital Object Identifier 10.4230/OASIcs.SAIA.2024.5

Category Practitioner Track

1 Introduction

AI systems impact humans, societies and the environment through various casual pathways. These effects arise not only through their intended functionalities but also through the limitations of those functionalities (e.g. biases) and the prerequisites for their operation in terms of computing facilities, their construction and material supply needs like electrical

[1] corresponding author

power and water. The term "risk" in the context of this article refers to the possibility of detrimentally affecting entities in various relationships to a planned or deployed AI system. The goal is to present a systematic approach to evaluate possible impacts, which could lead to violations of essential properties for sustainably deploying ethically sound and trustworthy AI systems, without hidden costs or side effects. The Artificial Intelligence Act (AI Act) [2] of the European Union, refers to the following seven principles identified by the AI High Level Expert Group (AI HLEG) as relevant for trustworthy AI: human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity, non-discrimination and fairness, societal well-being, and accountability. In addition to these principles, the AI Act notably also recognizes energy and environmental sustainability, biodiversity and the rights of the Charter of Fundamental Rights of the European Union as relevant in this context [8, 7].

2 Lifecycle Phases and Corresponding Impacts

Machine learning-based AI systems rely on the collection and processing of large amounts of data, requiring computing equipment often organized in the form of data centers housing great numbers of computing devices with corresponding needs for energy supply and cooling facilities [1, 6]. The lifecycle therefore must take into account the construction of such data centers and the manufacture of their main systems including the computing devices and their supply infrastructure. It is possible to perform an analysis on different levels of granularity. For the current purpose of describing the general approach we distinguish the following phases.

- Data center site construction (physical building, utilities and means of access)
- Data center installation (incl. installation of computing and supply facilities)
- Data center operation (incl. data collection, preparation, model training)
- Data center decommissioning (incl. service discontinuation, building demolition)

Each phase encompasses sub-phases which include dependencies in terms of work and services performed by other businesses, organizations and, ultimately, the human agents involved. Some of this work is directly contracted by the data center operators, while significant parts are hidden in a complex supply and service chain. Indirectly involved work and activities are the manufacture of components for the data center infrastructure and the extraction and processing of resource materials for that purpose. Mining operations for materials like copper, cobalt, lithium and rare earths often have significant environmental impacts, affect human health and agriculture [9]. The demand for specific resource materials may also spawn illegal mining operations where work is performed for minimal income but at high risk to human health and life. Business interests related to mining may further result in direct human rights abuses [4].

Hidden and mostly ignored work involved in data processing includes labelling of data for creating the "ground truth" input data for model training, which are often delivered by an anonymous work force of "volunteer" internet users for minimal pay [3].

Data center operation will provide the basis for companies who deliver derived services utilizing computing facilities and pre-trained models. Such services must be evaluated on their own for their implications. A very successful AI-driven marketing strategy may lead to substantially increased demand for production and shipment of physical items with all pertaining aspects of fair and safe working conditions and environmental impact. Data center decommissioning will involve the aspects of service discontinuation, removal and disposal of infrastructure equipment, and the removal or possible repurposing of the constructed

buildings. Regarding service discontinuation, a relevant question may be whether the service itself is entirely discontinued or whether other facilities are available to provide it instead. Complete service discontinuation needs to be considered with respect to its ripple effects on dependent businesses and general users.

3 Affected Roles / Stakeholders

To address and analyse the risks of AI systems, it is important to understand their impacts related to different domains and costs/benefits for specific stakeholders. In the context of a regulatory technical domain, one can consider the following roles/stakeholders:

- manufacturer
- provider
- user
- workers & employees
- regulatory authority
- certification body

Each of these roles/stakeholders is affected by or is affecting the functionalities of an AI system differently. Therefore, it is important to differentiate the risk analysis based on each role.

4 Degrees of Involvement

Both the intended functionality and any malfunction of an AI system can exert impact on the various stakeholders. There are different ways to relate to the use of a specific technology. For this reason, and in order to be able to discuss less obvious relationships, we differentiate between several degrees of involvement (refer to Table 1).

The recognition of the 0th degree of involvement allows us to also analyse requirements for sustainability as the environment or society are providing pre-conditions and resources for developing, operating and using any technological system and are in turn affected by such systems in very complex ways.

There is no general or specific relationship between lifecycle phases, stakeholders/roles and degrees of involvement. Instead, the approach is to systematically query a lifecycle process with respect to these aspects and investigate at each phase what entities are instrumental in realizing that specific phase, by what means they contribute to that phase and what are the intended and unintended effects of their contribution.

Table 1 Degrees of Involvement.

Degree	Explanation	Classification
1^{st}	the main agent or instigator of an activity	intentional and directly affected
2^{nd}	party intentionally participating in the agent's activity	intentional and in-/directly affected
3^{rd}	party randomly encountered and not intentionally involved	unintentional and in-/directly affected
n^{th}	party in other location, usually not encountered and whose existence may even be unknown to the agent	unintentional and indirectly affected
0^{th}	natural environment, society, economy	unintentional and indirectly affected

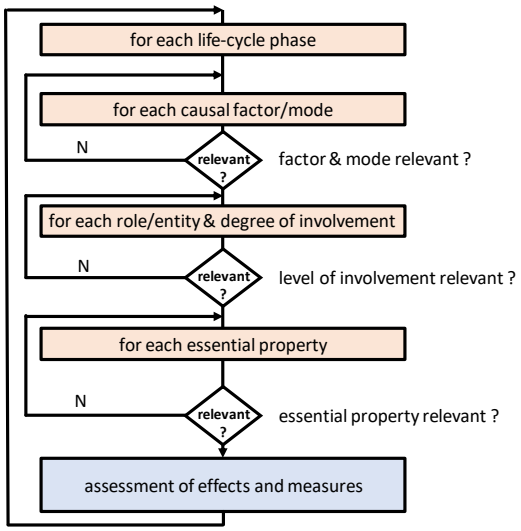

Figure 1 Risk analysis procedure.

5 Risk Analysis Procedure

The proposed approach to risk analysis has similarities to the concept of causal paths as described in [5] but differs in the detailed application. In causal modeling a causal path is a sequence of events which connects a cause to an effect. This connection can either be direct or involve several intermediate effects. In our approach we aim to construct the cause-effect relations iteratively, for each event repeating the search for modes of causation which will identify further connected events. We consider causal factors or causative media indicated by the following keywords:

- resources,
- materials,
- energy,
- data,
- communication,
- transportation,
- employment,
- social interaction

Performing certain process steps in the overall lifecycle may reveal the need to subcontract services or to procure equipment. These services and the implications regarding the acquired equipment from obtaining the necessary materials, through manufacture to delivery themselves need to be analysed in a similar way, as they contribute to the overall "impact footprint" of the operation. In addition to the causal factors, we also consider the modes of impact exertion, which are expressed through contrastive pairs of terms, for example: consumption - release, gathering - dissemination, demand - supply, improvement - impairment, addition - removal. These causal factors, considered in accordance with their associated modes are evaluated for their impact on the relevant AI properties for affected entities. The examples in Table 2 only show the application of the method in principle. Essential to note is the capacity of the approach to reveal effects, which are often not explicitly stated in the discussion of AI technologies, but which are part of their overall impact and require consideration.

Table 2 Example applications.

causal factor	mode	effect	entity	property	description
physical object (building)	addition	obstruction	environment	environmental integrity	change of microclimate
resource extraction, mining	demand	release of harmful byproducts and substances, dust, smoke, gas	local population	human health	effects on human health due to both direct effects of harmful substances and indirect effects due to diminished possibility for agriculture, forestry, fisheries and tourism
employment	demand	increased employment opportunities	local and global population	fair and safe working conditions	"microwork", low remuneration

6 Conclusion

We have outlined an approach for systematically investigating the possible impacts of AI technologies at each stage of their deployment lifecycle. These impacts can occur both directly through their intended functionalities and indirectly through the prerequisites for their deployment. Figure 1 outlines the key steps in the analysis. Each traversal of the diagram encounters the step labelled "assessment of effects and measures" and this is where the possible impacts are to be described and documented. In the next steps, effective countermeasures are to be planned to mitigate the detrimental effects which were identified during the analysis.

References

1 Luiz André Barroso, Urs Hölzle, and Parthasarathy Ranganathan. *The Datacenter as a Computer: Designing Warehouse-Scale Machines*. Springer International Publishing, 2019. doi:10.1007/978-3-031-01761-2.

2 Council of European Union. Council regulation (EU) no 2024/1689, 2024. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L:2024:1689.

3 P. Hitlin. Research in the crowdsourcing age: A case study, 2016. URL: http://www.pewinternet.org/2016/07/11/research-in-the-crowdsourcing-age-a-case-study.

4 Amnesty International. Democratic republic of the congo: Industrial mining of cobalt and copper for rechargeable batteries is leading to grievous human rights abuses. accessed on 2024-09-16. URL: https://www.amnesty.org/en/latest/news/2023/09/drc-cobalt-and-copper-mining-for-batteries-leading-to-human-rights-abuses/.

5 Chris Leong, Tim Kelly, and Robert Alexander. Incorporating epistemic uncertainty into the safety assurance of socio-technical systems. In Alex Groce and Stefan Leue, editors, *Proceedings 2nd International Workshop on Causal Reasoning for Embedded and safety-critical Systems Technologies, CREST@ETAPS 2017, Uppsala, Sweden, 29th April 2017*, volume 259 of *EPTCS*, pages 56–71, 2017. doi:10.4204/EPTCS.259.7.

6 Peng Li, Jianyi Yang, Mohammad Atiqul Islam, and Shaolei Ren. Making ai less "thirsty": Uncovering and addressing the secret water footprint of ai models. *ArXiv*, abs/2304.03271, 2023. URL: https://api.semanticscholar.org/CorpusID:257985349, doi:10.48550/arXiv.2304.03271.

7 Publications Office of the European Union. Charter of fundamental rights of the european union. Technical Report 12012P/TXT, European Union, Brussels, Belgium, October 2012. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:12012P/TXT.

8 European Commission Publications. Ethics guidelines for trustworthy ai. Technical report, European Commission, Brussels, Belgium, April 2019. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

9 Nicolás C. Zanetta-Colombo, Tobias Scharnweber, Duncan A. Christie, Carlos A. Manzano, Mario Blersch, Eugenia M. Gayo, Ariel A. Muñoz, Zoë L. Fleming, and Marcus Nüsser. When another one bites the dust: Environmental impact of global copper demand on local communities in the atacama mining hotspot as registered by tree rings. *Science of The Total Environment*, 920:170954, 2024. doi:10.1016/j.scitotenv.2024.170954.