GConsumers and Audiovisual Platforms: An Assessment of International Jurisdiction*

Andrea Guillen[1], Emma Teodoro[1]

[1]*Institute of Law and Technology, Autonomous University of Barcelona, Bellaterra, Spain*

Abstract

This paper explores how to design, develop, and deploy trustworthy AI tools in industrial manufacturing. After a brief overview of existing AI ethical frameworks, the paper focuses on actioning the four AI ethical principles identified by the AI HLEG. Given the context-dependency of AI tools, these AI ethical principles are framed within the manufacturing setting. This ethics-based approach requires the operationalization of such principles to truly design, develop, and deploy trustworthy AI systems. To this end, organizational and technical measures applicable to industrial manufacturing are suggested.

Keywords

AI ethics, industrial manufacturing, trustworthy AI

1. Introduction

The implementation of AI systems in industrial manufacturing brings about numerous benefits, from less machine downtime to less defects during the production process. Yet, there are considerable legal, ethical, and societal challenges to be addressed to fulfill their potential.

In recent years, the public sector, research institutions and private companies have issued various principles and guidelines for ethical trustworthy AI. However, as AI systems are context-dependent, these general AI ethical principles need to be adapted to the specific context of application to successfully imbue them into AI systems. Thus, not only technological aspects of industrial AI should be considered, but also other industrial requirements such as value creation, economic growth, human-machine interaction, and legal, ethical, and societal aspects.

This paper is conceived as a starting point to operationalize AI ethical principles in AI tools for manufacturing. The paper explores the most predominant AI ethical frameworks and examines four ethical AI principles contextualized within the industrial manufacturing context.

2. AI Ethical Frameworks

The public sector, research institutions, and private companies have issued various ethical frameworks to ensure trustworthy AI. A number of initiatives have aimed to capture this proliferation and map the landscape of such frameworks. For instance, the EU-funded project SHERPA

Proceedings of Artificial Intelligence Governance Ethics and Law (AIGEL), Reviewed, Selected Papers. November 02 - December 19, 2022, Barcelona, Spain

andrea.guillen@uab.cat (A. Guillen); emma.teodoro@uab.cat (E. Teodoro)

0000-0002-5093-1474 (A. Guillen); 0000-0001-9086-3616 (E. Teodoro)

CEUR Workshop Proceedings (CEUR-WS.org)

(Shaping the Ethical Dimensions of Smart Information Systems: A European Perspective) found over 70 relevant documents [1]. AlgorithmWatch AI Ethics Guidelines Global Inventory lists more than 160 documents, including industry related guidelines developed by Google, IBM and Microsoft [2].

The IEEE (Institute of Electrical and Electronics Engineers) Global Initiative on Ethics of Autonomous and Intelligent Systems, called 'Ethically Aligned Design' was officially launched in April 2016 as a collective program of the IEEE, the world's largest technical professional organization [3]. It identified over one hundred and twenty key issues and eight founding values and principles to be applied to all types of autonomous and intelligent systems which operate in real, virtual, contextual, and mixed-reality environments [3]. Namely, (i) human rights; (ii) well-being; (iii) data agency; (iv) effectiveness; (v) transparency; (vi) accountability; (vii) awareness of misuse; and (viii) competence.

The work published by the High-Level Expert Group on Artificial Intelligence of the European Commission (AI HLEG), "Ethics Guidelines for Trustworthy AI" [4] provides a set of ethical principles and requirements that should be embedded from the design into AI solutions to be deemed trustworthy. According to the AI HLEG there are four high-level ethical principles: i) human autonomy; ii) prevention of harms; iii) fairness; and v) explicability. These principles are turned into specific requirements for their practical implementation. These requirements are: i) human agency and oversight; ii) technical robustness and safety; iii) privacy and data governance; iv) transparency; v) diversity, non-discrimination, and fairness; vi) environmental and societal well-being; and vii) accountability.

This ethics-based approach can be used to operationalize AI ethical principles into a specific context of application—AI solutions for industrial manufacturing—which takes into account not only technological aspects of Industrial AI, but also other industrial requirements such as value creation, human-AI-interaction, ethical and regulatory aspects [5].

3. Operationalizing AI Ethical Principles in Manufacturing

This section follows the AI ethical principles established by the High-Level Expert Group on Artificial Intelligence (AI HLEG) [4], which have been adapted to the context of industrial manufacturing from an action-guiding perspective. This approach allows us to glimpse which ethical challenges may be faced in this context and recommends organizational and technical measures.

3.1. Human Autonomy

The principle of human autonomy implies that AI-enabled technologies should be designed, developed, and deployed in a way that respects and protects fundamental rights and ensures human agency and oversight.

AI-enabled technologies must ensure human dignity. In the shopfloor, the objectification and dehumanisation of operators should be avoided. Workers should be treated as self-determined subjects whose physical and mental health must be protected. Worker's dignity might also be undermined by the consequences that the deployment of AI systems in the workplace may have on the de-skilling of the labour force and the meaning of work.

The use of AI systems in factories may also lead to an advanced system of surveillance and monitoring to which operators may be subject [6]. Surveillance may cause “chilling effects” on employees and may also negatively impact their freedom, autonomy, and privacy. Therefore, legal, ethical, and social impact assessments must be conducted to strike the right balance between the intended benefits of the deployment of technology in the shopfloor and the possible negative consequences for employees’ ethical values and fundamental rights [7].

To ensure human agency, operators should be able to make informed autonomous decisions regarding AI tools outcomes and have the skills to assess and challenge them. Therefore, training sessions are encouraged to ensure that operators have the knowledge to understand how the system works and how to interact with it [8].

The purpose of human oversight is to prevent or minimise the potential risks of AI-enabled technologies. Meaningful human control can only be achieved if human-centric design principles and appropriate human-machine interfaces are embedded into the technologies. Additional measures should be implemented to ensure that operators have the expertise, necessary competencies, and authority to exercise human control effectively, e.g., training sessions that enable the understanding of the capacity and limitations of the deployed technology, awareness of automation bias [9].

3.2. Prevention of Harms

The principle of prevention of harms means that AI-enabled technologies should not cause harm nor have detrimental consequences for individuals. This implies that operators’ dignity must be respected, and their mental and physical integrity protected. Particular emphasis must be placed on the potential harms that technology can cause or exacerbate to workers, who are considered by the AI HLEG vulnerable people given the power imbalance and information asymmetries with employers. To minimise the impact of AI-enabled technologies on operators, a participatory approach could be adopted where workers are involved in the development and deployment of the technology [10].

The potential harms that can be caused by AI-enabled technologies also require addressing: i) the technical robustness and safety of the technology; ii) privacy and data governance concerns; and iii) societal and environmental well-being.

Firstly, AI-enabled technologies must be robust, resilient, secure, safe, accurate, reliable, and reproducible. Technical robustness and resilience should be ensured to prevent the exploitation of vulnerabilities by third parties and misuse [11]. Therefore, the existence of potential security risks must be evaluated at the design, development and deployment phases, and mitigation measures must be implemented in accordance with the magnitude and likelihood of the risks. Security and safety measures should also be put in place to enhance operators’ safety and prevent detrimental consequences. To this end, a fallback plan can serve to ensure safety in case of a system failure. Likewise, AI-enabled technologies must be accurate. Accuracy rates should be particularly high when such systems can directly affect individuals, as is the case with operators whose integrity may be compromised. Accuracy must be monitored on an ongoing basis and procedures to mitigate and correct potential risks must be implemented. Additionally, operators need to trust the system to use it, therefore reliability and reproducibility are key aspects to ensure the adoption of the technology among them [12].

Secondly, the prevention of harms to privacy and data protection is paramount given the potential risks that AI-enabled technologies pose to these fundamental rights through the processing of massive amounts of personal data, including the unintended collection of personal data. These rights can also be at stake because personal information can be inferred from non-personal data [13].

Respect for workers' right to privacy and data protection must be ensured by complying with the GDPR and by aligning with existing standards or widely adopted protocols. Importantly, in IoT environments, it is particularly crucial to clarify data ownership, the roles of data controllers and processors and access to data [14]. Oversight mechanisms must also be put in place to ensure data quality (e.g., representativeness in the dataset) and integrity that minimises the risks of using biased, inaccurate, or compromised datasets. Therefore, processes and datasets must be scrutinised and documented throughout the AI system's lifecycle.

Lastly, the use of AI-enabled technologies should aim at benefitting society and the environment. AI systems must be designed, developed, and deployed with sustainability and environmental friendliness in mind. Therefore, the ecological impact of the system should be evaluated throughout the system's lifecycle and measures to reduce such impact should be encouraged. The social impact of the system should be regularly assessed both at the individual and societal level. For instance, the evaluation of the impact of the technology on operators should cover physical and mental health issues, non-discrimination, de-skilling of the workforce, among others. As for the societal considerations, the impact on the job market and the societal consequences it may entail should be addressed [8].

3.3. Fairness

The principle of fairness entails equality, diversity and the prevention of discrimination and stigmatisation against individuals and groups. Equality requires that all persons by virtue of their humanity and regardless of age, gender, sexuality, disability, ethnicity or other group or relevant personal characteristic deserve equal regard and respect.

Fairness can be achieved by i) promoting diversity, inclusion and non-discrimination; ii) fostering societal and environmental well-being while reducing potential harms; and iii) adopting accountability measures.

Firstly, diversity and non-discrimination can be enhanced with oversight processes that identify, examine, address, and test biases in the datasets and at the design and development phases [15]. From a design perspective, technology should be understandable and accessible to all operators regardless of their age, abilities, or characteristics. In this regard, the participation of relevant stakeholders with diverse backgrounds and viewpoints at the different stages is highly encouraged to ensure that diversity is embedded into the system [16].

Secondly, AI-enabled technologies should be designed to strive for social and environmental well-being. Concerning the principle of fairness, the social impact of the system on operators should be evaluated in terms of causing or exacerbating discrimination, stigmatisation, or marginalisation.

Lastly, accountability requires the implementation of appropriate technical and organisational measures to report the system's performance and provide effective remedy and redress to the extent possible. Such measures include the assessment of design processes, the underlying technology, and the data sets used, which allows for the auditability of the system. Auditability involves reporting the negative impacts of the system, identifying appropriate mitigation measures, and feeding them into the system [7]. These negative impacts can be identified and assessed through comprehensive impact assessments that must be conducted on a regular basis [16]. Accountability also includes providing explanations of the system's outcomes and the ability to seek redress.

3.4. Explicability

The principle of explicability requires transparency of the AI system – including the datasets, the inner workings of the system and the business model –which ultimately enables human oversight [8]. For systems to be transparent, traceability measures must be implemented. This implies that datasets and the technology that underlies the system should be documented, e.g., the methods used for designing and developing the system, the methods used to test and validate it and the outcomes of the system. Given that traceability allows for the identification of the reasons behind systems' outcomes, it enables explainability.

Explainability means the ability to explain the outcomes made by the system intelligibly [11]. To this end, the rationale behind a system's outcome should be understood and traced by humans. Therefore, if a system's outcomes cause harm to operators, explanations of how the system arrived at it should be provided to the worker in plain language. In this regard, communication is crucial since operators must be aware that they are interacting with an AI system in the first place in order to be able to request an explanation. Consequently, operators must be informed in a clear and understandable manner about their interaction with an AI system, how the system works and its purpose, as well as its capabilities and limitations [8].

4. Conclusions

AI systems in industrial manufacturing do not only lead to positive impacts but have risks associated with it of negative effects on the workforce, on the environment, on the company's reputation and broader society. Turning AI ethics principles into practice requires the implementation of technical and organizational measures aimed at ensuring the design and development of trustworthy AI systems. Thereby, empowering operators, enhancing social inclusion, and informing up-skilling training programs. In sum, adopting AI technologies in the shopfloor that are beneficial to humans individually, organizationally and societally.

This paper is a first attempt to put AI ethical principles into practice in industrial AI. It provides a starting point for the discussion on how the effective operationalization of AI ethical principles can be achieved in manufacturing. Further research should focus on the implementation of the AI ethical requirements corresponding to these AI ethical principles. Namely, i) human agency and oversight; ii) technical robustness and safety; iii) privacy and data governance; iv) transparency; v) diversity, non-discrimination, and fairness; vi) environmental and societal well-being; and vii) accountability. This would provide a higher level of granularity that would allow more specific organizational and technical measures, thereby benefitting the operationalization of the high-level ethical principles and the design and development of trustworthy AI tools. Likewise, further research on the operationalization of AI ethical principles in manufacturing should be tailored to the multiple data-driven technologies used in the shopfloor— for instance, digital twins, cobots or VR /AR /XR— as they entail different legal, ethical and societal risks to be accounted for.

Acknowledgments

This research was funded by the OPTIMAI project as part of the European Union's Horizon 2020 research and innovation programme, under grant agreement No. 958264.

References

[1] P. Brey, B. Lundgren, K. Macnish, M. Ryan, A. Andreou, L. Brooks, Tilimbe Jiya, R. Klar, D. Lanzareth, J. Maas, I. Oluoch, B. Stahl, D3.2 Guidelines for the development and the use of SIS, SHERPA (2021). doi: 10.21253/DMU.11316833.V3.

[2] AlgorithmWatch, AI Ethics Guidelines Global Inventory, (2020). URL: https://algorithmwatch.org/en/ai-ethics-guidelines-global-inventory.

[3] IEEE, IEEE Ethics In Action in Autonomous and Intelligent Systems" (n.d.). URL: https://ethicsinaction.ieee.org/.

[4] AI HLEG, Ethics guidelines for trustworthy AI, (2019). URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[5] M.W. Hoffmann, R. Drath, C. Ganz, Proposal for requirements on industrial AI solutions, in: J. Beyerer, A. Maier, O. Niggemann, Eds., Machine Learning for Cyber Physical Systems, Springer, Berlin, Heidelberg, 2021, pp. 63–72. doi:10.1007/978-3-662-62746-4_7.

[6] I. Ajunwa, K. Crawford, J. Schultz, Limitless Worker Surveillance, California Law Review, 105 (2017) 735-776. doi:10.15779/Z38BR8MF94.

[7] J. Fjeld, N. Achten, H. Hilligoss, A. Nagy, M. Srikumar, Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI, (2020). URL:http://nrs.harvard.edu/urn-3:HUL.InstRepos:42160420.

[8] Eurofound, Game-changing technologies: Transforming production and employment in Europe, (2020). URL: https://www.eurofound.europa.eu/publications/report/2020/game-changing-technologies-transforming-production-and-employment-in-europe

[9] Article 29 Data Protection Working Party, Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, (2017). URL: https://ec.europa.eu/newsroom/article29/items/612053.

[10] B. Törpel, A. Voss, M. Hartswood, R. Procter, Participatory Design: Issues and Approaches in Dynamic Constellations of Use, Design, and Research, in: M. Büscher, R. Slack, M. Rouncefield, R. Procter, M. Hartswood, A. Voss, Eds., Configuring User-Designer Relations. Springer London, London, 2009: pp. 13–29. doi:10.1007/978-1-84628-925-5_2.

[11] M.W. Hoffmann, R. Drath, C. Ganz, Proposal for requirements on industrial AI solutions, in: J. Beyerer, A. Maier, O. Niggemann, Eds., Machine Learning for Cyber Physical Systems. Springer Berlin Heidelberg, Berlin, Heidelberg, 2021: pp. 63–72. doi:10.1007/978-3-662-62746-4_7

[12] P. Jansen, P. Brey, D4.4: Ethical Analysis of AI and Robotics Technologies, SIENNA, n.d. https://www.sienna-project.eu/digitalAssets/884/c_884668-l_1-k_d4.4_ethical-analysis--aiand-r--with-acknowledgements.pdf

[13] S. Wachter, B. Mittelstadt, A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI, Columbia Business Law Review, 2 2019. doi:10.31228/osf.io/mu2kf.

[14] S. Wachter, Normative Challenges of Identification in the Internet of Things: Privacy, Profiling, Discrimination, and the GDPR, Computer Law & Security Review, 34 (2018) 436–449. doi:10.1016/j.clsr.2018.02.002.

[15] J. Beyerer, A. Maier, O. Niggemann, Eds., Machine Learning for Cyber Physical Systems: Selected papers from the International Conference ML4CPS 2020. Springer Vieweg, 2021. doi:10.1007/978-3-662-62746-4.

[16] Access Now, Human Rights in the Age of Artificial Intelligence, (2018). URL: https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf.