








Sustainability Effects of Robust and Resilient Artificial Intelligence

Torsten Priebe^(✉) , Peter Kieseberg , Alexander Adrowitzer ,
Oliver Eigner , and Fabian Kovac 

Institute of IT Security Research, St. Pölten University of Applied Sciences,
St. Pölten, Austria

{torsten.priebe,peter.kieseberg,alexander.adrowitzer,
oliver.eigner,fabian.kovac}@fhstp.ac.at

Abstract. It is commonly understood that the resilience of critical information technology (IT) systems based on artificial intelligence (AI) must be ensured. In this regard, we consider resilience both in terms of IT security threats, such as cyberattacks, as well as the ability to robustly persist under uncertain and changing environmental conditions, such as climate change or economic crises. This paper explores the relationship between resilience and sustainability with regard to AI systems, develops fields of action for resilient AI, and elaborates direct and indirect influences on the achievement of the United Nations Sustainable Development Goals. Indirect in this case means that a sustainability effect is reached by taking resilience measures when applying AI in a sustainability-relevant application area, for example precision agriculture or smart health.

Keywords: artificial intelligence · machine learning · resilience · security · sustainability

1 Introduction

Artificial intelligence (AI) can be usefully applied to build resilience in many areas. However, this can simultaneously open up new threats in the area of IT security, as the use of technology creates the risk of failure, vulnerability, or misuse. This paper discusses, how these threats can be countered and initially demonstrates how the creation of robust and resilient AI can also have sustainability effects in line with the United Nations (UN) Sustainable Development Goals [35].

To this end, we must first define what resilient AI means. We base this on multiple definitions by the National Institute of Standards and Technology (NIST), which we consolidate as follows: Resilience is the ability of an information system

This research was funded in part by the German Federal Ministry for the Environment, Nature Conservation, Nuclear Safety and Consumer Protection (BMUV) by resolution of the German Bundestag through the cooperative project CO:DINA as part of the AI Lighthouse initiative.

© The Author(s) 2023

A. Holzinger et al. (Eds.): CD-MAKE 2023, LNCS 14065, pp. 188–199, 2023.

https://doi.org/10.1007/978-3-031-40837-3_12

to mitigate the impact of known or unknown changes in the operating environment (including intentional attacks, accidents, and naturally occurring threats) by (a) anticipating and preparing for such events (e.g., through risk management, contingency, and business continuity planning), (b) the ability to withstand and adapt to attacks, adverse conditions, or other stresses and to continue operations (or rapidly regain the ability to do so), while maintaining essential and required operational capabilities, and (c) restoring full operational capability after such disruption in a time-frame consistent with mission requirements. Thus, it is also about “robustness” in the face of a changing environment. Großklaus [16] refers to this “ability to successfully drive the sustainable development of society under uncertain and changing conditions” as “transformative resilience”.

We would like to point out that AI systems can never be considered as isolated, abstracted entities, but must be seen in their social context as sociotechnical systems; it must be understood that algorithms and especially AI are not just technical artifacts – in the sense of “physical, human-designed objects that have both a function and a plan of use” as defined by Vermaas et al. [36] – but complex systems characterized by collective and distributive action [30]. In our case, this means that the concept of resilient AI has links to issues of acceptance and trust. No AI system can be called resilient if its use is fraught with fundamental mistrust, accountability gaps, accusations of unfairness, or criticism of its black-box nature. Likewise, the EU’s Joint Research Centre [19] distinguishes four dimensions of resilience: *societal*, *economic*, *organizational*, and *technological*. The focus of this work is on the *technological* side, touching *organizational* aspects where appropriate. The *societal* sense of resilience is in fact what we refer to as *sustainability*. This work is based on the assumption that the guiding principles of resilience (in a technical/organizational sense) and sustainability complement each other: In a crisis-ridden world, resilience becomes a basic requirement for the success of sustainability goals. To validate this assumption, this paper develops fields of action for resilient AI and examines their sustainability impacts – in general and in selected sustainability-relevant application areas. The underlying study is not yet complete; therefore, this paper represents initial considerations and results of an ongoing work-in-progress effort.

The rest of this paper is organized as follows: Sect. 2 outlines our research method and in particular the increasing focus in our stepwise approach. Section 3 introduces robust and resilient AI and develops a roadmap for fields of action. Section 4 discusses direct sustainability effects of resilient AI as well as indirect effects, if according measures are taken in sustainability-relevant application areas. Finally, Sect. 5 concludes the paper, given that we are presenting work-in-progress, with a focus on current and future work.

2 Research Method

As shown in Fig. 1, we proceed in three steps. In Eigner et al. [11], a survey of the scientific state-of-the-art as well as the identification of possible fields of action of robust and resilient AI was carried out; a summary can be found in

Sect. 3. This paper presents the continuation of this work. We have analyzed potential sustainability impacts of the identified action areas based on academic literature, project results and brainstorming with experts. In this second step we focused on sustainability-relevant application areas. The resulting overview of sustainability effects is provided in Sect. 4. Last not least, we are currently identifying recommended actions for public actors with a further increased focus on smart cities and regions, critical infrastructures and ecological sustainability goals using an exploratory scenario analysis [24].

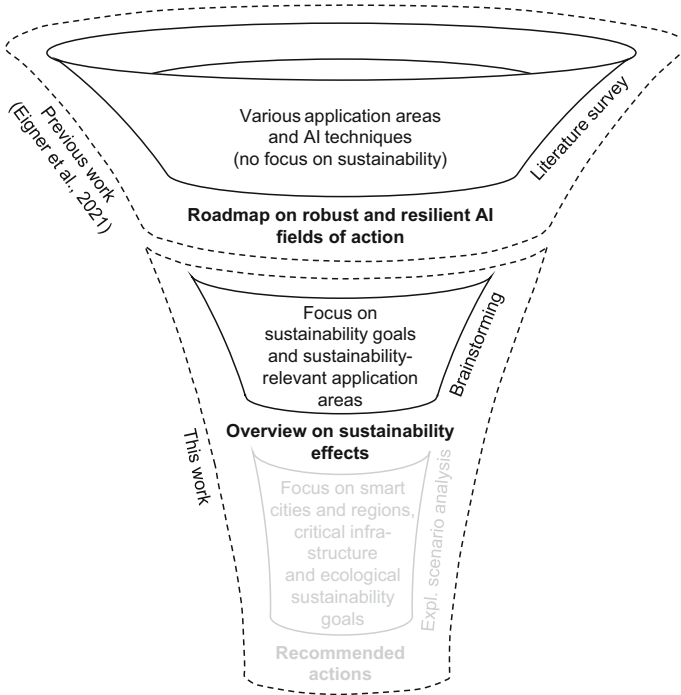


Fig. 1. Increasing focus and methods applied throughout our research

As shown in Fig. 2, different forms of dependencies between robust and resilient AI and sustainability emerge:

- Direct sustainability effects of a field of action, e.g., the reduction of bias in AI algorithms has a direct positive impact on gender equality.
- Indirect sustainability effects of robust and resilient AI in a specific application area, e.g., increasing the robustness of an AI in precision agriculture contributes to reducing hunger.
- We are developing concrete recommendations for resilient AI in selected areas, which may bring up new impacts. E.g., a recommendation to address drift by frequently retraining machine learning (ML) models may increase energy consumption and therefore have a negative sustainability effect.

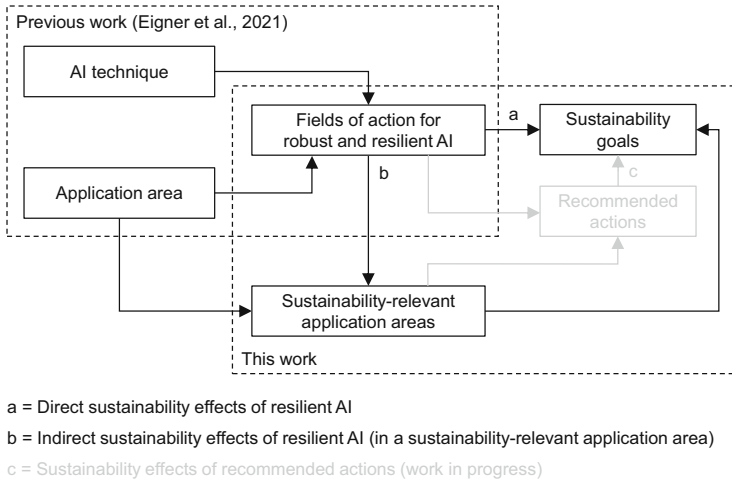


Fig. 2. Dependencies of the various concepts in our research

3 Robust and Resilient Artificial Intelligence

AI systems are becoming essential to our daily lives. Organizations should ensure their resilience as with any other critical asset. However, the black-box approach typically found in AI may make assessing and ensuring resilience different compared to traditional IT systems. In Eigner et al. [11], we provide an overview of the emerging field of resilient AI, both from the perspective of selected application areas and specific AI techniques. From this, we derive fields of action for robust and resilient AI. In Fig. 3 we structure these in the form of a roadmap, as some targets have already been or are being extensively addressed, while others will only reach practical applicability in the medium or long term. Following the European Union (EU) High-Level Expert Group on Artificial Intelligence [17], we divide the fields of action into *security*, *safety*, *accuracy* and *reliability*.

Security. Security incidents in AI can be distinguished by (a) the AI technology used, (b) the type of incident, or (c) the stage of the AI/ML pipeline in which the incident occurs. The type of disruption can be broadly divided into intentional disruptions, which include all types of hostile attacks on AI systems and unintentional disruptions, which can range from careless human interaction to rare special cases that systems may never have encountered before. With this in mind we see *plausibility tests*, i.e. rules that identify at least outliers or unexpected results of an AI algorithm, as a basic level of protection. More sophisticated methods of *mitigating hostile attacks* are known as *adversarial AI* [23].

Penetration testing (“pentesting” for short) is a key technique for assessing the security and resilience of IT services and products. Since there are myriads of possible threats in the field of AI that can affect the proper operation of AI

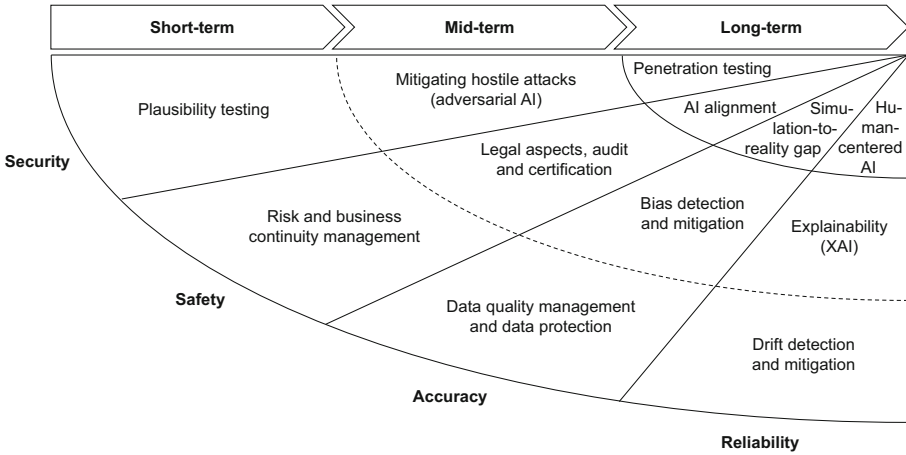


Fig. 3. Roadmap for robust and resilient AI fields of action

applications, pentesting of AI will certainly gain importance in the near future. Initial approaches have been proposed, e.g., by Das et al. [8] or Tjoa et al. [33], or are under development, still, currently no best practices methodologies exist.

Safety. Regarding the field of action *risk and continuity management*, KPMG highlights in its AI Risk and Controls Matrix various risks associated with AI [25]. Risks to be highlighted in this area include inadequate fallback solutions related to infrastructure, the AI solution itself, and business operations. In addition, the inability to restore service after an incident is highlighted as a specific AI risk, as the last good AI state may not be easily restored due to its complex and often black-box nature.

In terms of *legal aspects*, the United Kingdom Information Commissioner’s Office [21] identifies three key areas: the legal status of algorithms, sector-specific standards, and the interdependencies of privacy and AI. In addition, monitoring robustness to (even non-adverse) changes in the environment opens up another area of research requiring a holistic view of *audit and certification* of AI systems [37]. In related terms, the upcoming EU AI-Act [12] differentiates AI systems into four risk categories, ranging from “unacceptable” risks to “high”, “limited” and “minimal” risk levels. The categorization is not only depending on the AI technology in use, but also on the sensibility of data and application area.

AI alignment aims align AI systems with human preferences and ethical principles. AI systems, especially when using reinforcement learning, are based on specified objectives. As it can be difficult do define all desired and, in particular, undesired behaviors, AI systems may find loopholes to accomplish their objectives efficiently but in unintended, sometimes harmful ways [38].

Accuracy. Accuracy is one of the most important areas in the field of AI and has received a lot of attention in the last decades, both to enable new applications and to improve AI applications and make them market ready [1]. Data is the very source of good AI in this regard; if the *data quality* is poor or if there is distortion due to data pre-processing, many of these problems will be transferred to the result or even amplified [39]. This includes, in particular, *detection and avoidance of bias* [27]. Discrimination and bias often arise in the data collection and modeling process, for example, when target variables are incorrectly defined, questions are unclear, or historical data is used that was collected in a time when moral concepts are no longer in line with current ones [4].

A major challenge, especially in robotics, is also to bridge the *gap between simulation and reality* and to make “digital twins” robust to changing parameters of the environment. Here, domain randomization approaches [34] are promising. Furthermore, automation is especially problematic in so-called mixed environments, where robotic actors directly interact with human actors, which is a major problem in the area of autonomous driving [9].

Reliability. We consider the continuous reliability of AI, as well as aspects of trustworthiness and explainability. Machine learning (ML) models degrade over time. A major reason for this is that the world, and thus the data, are not static; therefore, the data to which the models are applied also change over time. This effect is referred to as “drift”. Methods for *drift detection and mitigation* have been discussed for some time [22], but adequate monitoring of AI systems has only recently been established by trends such as MLOps.

The *explainability* of AI often missing in many advanced methods refers to non-transparent (black-box-like) decision-making processes [15] for which, for example, testing for backdoors is practically impossible. *Human-centered AI* means to involve humans for labeling, improvement or correction. Especially the use of AI together with human expertise is promising here, e.g. by formalization with semantic technologies, resulting in the research area of semantic AI [5].

4 Sustainability Effects

In this section, we explore interdependencies between robust and resilient AI and the UN Sustainable Development Goals [35]. Are there synergies or do conflicting goals arise and how can these be negotiated? As outlined in Sect. 2, direct and indirect sustainability impacts can be identified. Here, indirect means an effect through a resilience field of action, provided the AI is used in a sustainability-relevant application area. In this paper, we consider *precision agriculture and forestry, smart health and precision medicine, smart cities and regions* (including energy and mobility transition), *industry and critical infrastructures*, and *police, justice, and military* as such sustainability-relevant application areas, informed by the project experience of the authors and the experts interviewed. Figure 4 illustrates the various dependencies broken down by resilience field of action and Sustainable Development Goal (SDG).

4.1 Direct Sustainability Effects

The AI Act of the European Union [12] includes, among others, a prohibition of discrimination against groups of persons on the basis of their sex or other characteristics. Dealing with *legal aspects and a corresponding certification* therefore leads directly to an improvement in relation to the sustainable development goals (SDGs) 5 (gender equality) and 10 (reduced inequalities). The same applies to the technical measures derived from this to *detect and mitigate bias*. *Explainability* of AI systems also leads to an improvement here, as inadequacies of the models used are at least made visible.

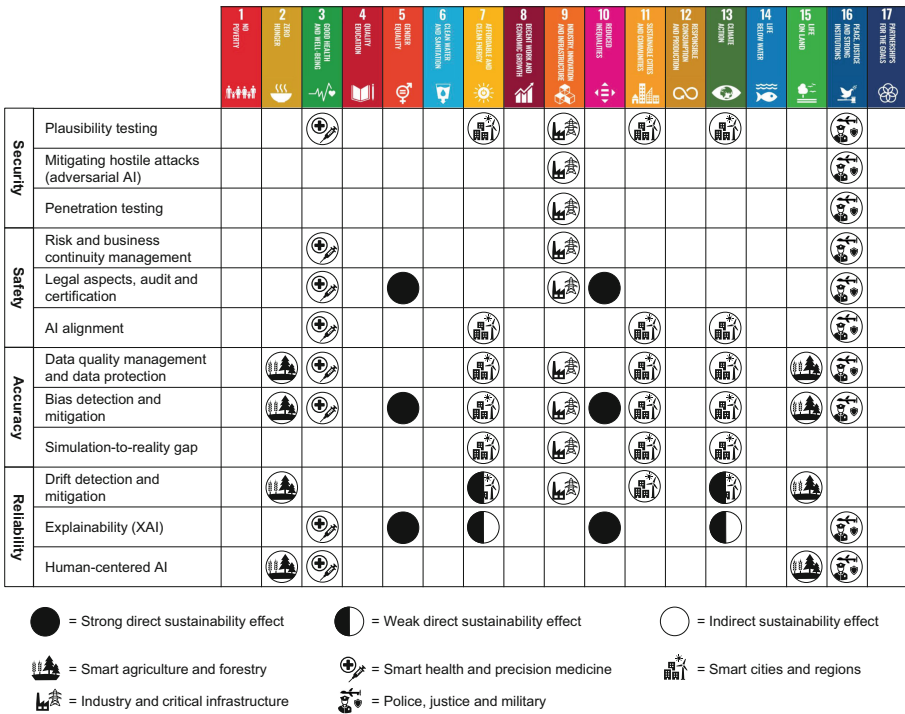


Fig. 4. Overview of sustainability impacts of robust and resilient AI fields of action

Explainable AI models also tend to use less computationally intensive algorithms, resulting in a sustainability impact with respect to SDGs 7 (affordable and clean energy) and 13 (climate action). On the contrary, *drift detection and mitigation* usually leads to regular (potentially frequent) retraining of AI models, causing an increase in energy consumption and therefore a negative effect on SDGs 7 and 13.

4.2 Sustainability Effects in Selected Application Areas

Precision Agriculture and Forestry. An important application area of AI in precision agriculture and forestry is plant pest and disease detection and prediction [20], which positively impacts SDG 2 (no hunger). Other goals include optimizing the use of scarce resources such as water, as well as fertilizers and pesticides, which also has a positive effect on SDG 15 (life on land). Relevant fields of action for resilience are *data quality* as well as *bias and drift detection and mitigation*. The latter is particularly important in a changing environment, e.g., due to climate change. Since labeled data is usually rare in such use cases [29], use of *human-centered AI* techniques (such as interactive learning) is also relevant.

Smart Health and Precision Medicine. Important application areas of AI in healthcare, and thus with an impact on SDG 3 (health and well-being) are, for example, individualized medications and semi-automated diagnosis. This is also referred to as P4 medicine (predictive, preventive, personalised and participatory) [13]. Here, AI always serves only as support; the final decision must rest with the physician. This is why *human-centered AI* approaches and *explainability* are so important [18].

Due to the (also legal) classification of medical products as “high-risk AI” [12], *risk and business continuity management* are also particularly important. Medical applications may not be common targets of cyberattacks (at least not like, e.g., critical infrastructure targets), however basic security measures such as *plausibility checks* should of course also be applied.

Medical AI systems have higher accuracy requirements than other applications. Therefore, consideration of *data quality* is particularly relevant, especially in the context of rare phenomena. This also applies to the avoidance of *bias*, since “biased” algorithms are more difficult to generalize. Given the nature of medicine directly affecting humans, *AI alignment* is of particular importance as well. Studying the effects of treatments, for example, on AI systems is not sufficient, basically like trying to study them only on lab mice.

Smart Cities and Regions. Sustainability in smart cities and regions is directly represented by SDG 11 (sustainable cities and communities). Application fields of AI here include the optimization of the use of renewable energy. Here, too, adaptability to a changing environment (e.g., due to climate change), i.e. *drift detection and mitigation*, is of central importance. An important aspect in many smart cities is the concept of the sharing economy, i.e. citizens make (private) resources available for use by others when they do not need them [14]. Studies have already been conducted on the impact of this sharing economy on the sustainability of such smart cities [3]. The use of AI in this area results in various trustworthiness and resilience requirements related to *bias and explainability*. Also, the fact that personal data is involved may require *data protection* measures such as anonymization, which is in fact a form of distortion in data preprocessing. The emerging use of digital twins in smart cities also requires consideration of the *simulation-to-reality gap*. Sharing itself might also increase the difficulty of attack attribution, which in itself is already a huge problem in IT Security [7].

Industry and Critical Infrastructure. Defending against *hostile attacks* and therefore *penetration testing* are particularly important in the area of industry and critical infrastructure (e.g., power plants and power grids), represented by SDG 9 (industry, innovation and infrastructure), as these are obvious targets for cyber warfare. Based on the attacks on the Ukrainian power grid in 2015, there was a strong increase in attention in this area [6] and the creation of corresponding technologies and organizational units like specialized CERTs for the energy sector. Applications of AI such as predictive maintenance or quality control require *data quality and distortion* handling and means of *drift detection and mitigation*. AI systems in critical infrastructure are legally classified as “high-risk AI” [12], hence requiring proper *risk and business continuity management*. *Standards and certifications* are also particularly important in this area [32].

Police, Justice and Military. Robust and resilient AI in the police, justice, and military sectors (e.g., through demonstrable avoidance of bias and discrimination) inherently impacts SDG 16 (peace, justice, and strong institutions) [2]. A prominent example here is the AI-based COMPAS database in the US, which was developed to predict the likelihood of recidivism among offenders [26]. However, predictive policing is rather controversial, both for ethical reasons (requiring resilience measures such as *bias mitigation* and *explainability*) and with regard to the actual verifiable benefit [28].

In the military field, the situation is still much more opaque, as many details are subject to secrecy. Nevertheless, some trends and frameworks can be identified, such as the ban on so-called Lethal Autonomous Weapon Systems (LAWS) [31] in the European Union [10] and specifically through the AI Act [12]. There are currently some well-known public programs, such as from the US, the focus in these publications is very much in the area of predictive maintenance, unmanned aerial vehicles (UAVs), training of personnel and augmentation.

5 Conclusion

In this paper, we have provided an initial overview of fields of action for robust and resilient AI and examined them for their sustainability effects. Direct effects were identified, e.g. through the reduction of bias. Indirect effects arise from the use of AI in sustainability-relevant application areas. We have selected precision agriculture and forestry, smart health and precision medicine, smart cities and regions, industry and critical infrastructures, as well as police, justice and military, based on the project experience of the authors and interview partners.

On the one hand, a further, broader survey would be useful to expand the analysis, which is certainly not complete to date. On the other hand, we also intend to go deeper and define concrete recommended actions to achieve resilience with a more narrow focus (public actors in critical infrastructures and smart cities and regions), which again need to be analyzed for their (in particular ecological) sustainability impacts. For example, a recommendation may be to regulate and therefore limit the use of large pre-trained AI models (due to

their intransparency and potential bias). However, given that these pre-trained models save training effort (and therefore resources) such a recommendation may have a negative sustainability effect. Both aspects are being addressed in the exploratory scenario analysis with our interviewees, which currently ongoing.

References

1. Achour, Y., Pourghasemi, H.R.: How do machine learning techniques help in increasing accuracy of landslide susceptibility maps? *Geosci. Front.* **11**(3), 871–883 (2020). <https://doi.org/10.1016/j.gsf.2019.10.001>
2. Adensamer, A., Klausner, L.D.: Part man, part machine, all cop: automation in policing. *Front. Artif. Intell.* Forthcom. (2021)
3. Akande, A., Cabral, P., Casteleyn, S.: Understanding the sharing economy and its implication on sustainability in smart cities. *J. Clean. Prod.* **277**, 124077 (2020)
4. Barocas, S., Selbst, A.D.: Big data's disparate impact. *104 California Law Review*, p. 671 (2016). <https://doi.org/10.2139/ssrn.2477899>, <https://www.ssrn.com/abstract=2477899>
5. Breit, A., et al.: Combining machine learning and semantic web: a systematic mapping study. *ACM Comput. Surv.* (2023). <https://doi.org/10.1145/3586163>
6. Case, D.U.: Analysis of the cyber attack on the Ukrainian power grid. *Electr. Inf. Shar. Anal. Center (E-ISAC)* **388**, 1–29 (2016)
7. Clark, D.D., Landau, S.: Untangling attribution. *Harv. Nat'l Sec. J.* **2**, 323 (2011)
8. Das, N., et al.: MLsploit: a framework for interactive experimentation with adversarial machine learning research. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '19*, ACM, New York, NY (2019). https://www.kdd.org/kdd2019/docs/KDD2019_Showcase_2062.pdf
9. Di, X., Shi, R.: A survey on autonomous vehicle control in the era of mixed-autonomy: from physics-based to AI-guided driving policy learning. *Transp. Res. Part C Emerg. Technol.* **125**, 103008 (2021)
10. Doll, T.: Künstliche Intelligenz in den Landstreitkräften. *Amt für Heeresentwicklung* (2019). <https://www.bundeswehr.de/resource/blob/156024/d6ac452e72f77f3cc071184ae34dbf0e/download-positionspapier-deutsche-version-data.pdf>
11. Eigner, O., et al.: Towards resilient artificial intelligence: survey and research issues. In: *2021 IEEE International Conference on Cyber Security and Resilience*, pp. 536–542. CSR 2021, IEEE, Washington, DC (2021). <https://doi.org/10.1109/CSR51186.2021.9527986>
12. European Commission: *Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts* (2021). <https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex:52021PC0206>, proposal for a Regulation of the European Parliament and of the Council, No. COM/2021/206 final
13. Flores, M., Glusman, G., Brogaard, K., Price, N.D., Hood, L.: P4 medicine: how systems medicine will transform the healthcare sector and society. *Pers. Med.* **10**(6), 565–576 (2013)
14. Frenken, K., Schor, J.: Putting the sharing economy into perspective. In: *A Research Agenda for Sustainable Consumption Governance*, pp. 121–135. Edward Elgar Publishing (2019)

15. Goebel, R., et al.: Explainable AI: the new 42? In: Holzinger, A., Kieseberg, P., Tjoa, A.M., Weippl, E. (eds.) CD-MAKE 2018. LNCS, vol. 11015, pp. 295–303. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99740-7_21
16. Großklaus, M.: Vom Modewort zum transformativen Hebel: Wie die Konjunktur des Resilienzbegriffs für die digital-ökologische Transformation genutzt werden kann. IZT (2022). <https://codina-transformation.de/transformative-resilienz/>, CO:DINA position paper no. 11
17. High-Level Expert Group on Artificial Intelligence: Ethics Guidelines for Trustworthy AI. Publications Office of the European Union, Luxembourg (2019). <https://doi.org/10.2759/346720>
18. Holzinger, A., Weippl, E., Tjoa, A.M., Kieseberg, P.: Digital transformation for sustainable development goals (SDGs) - a security, safety and privacy perspective on AI. In: Holzinger, A., Kieseberg, P., Tjoa, A.M., Weippl, E. (eds.) CD-MAKE 2021. LNCS, vol. 12844, pp. 1–20. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-84060-0_1
19. Institute for the Protection and Security of the Citizen (Joint Research Centre): Towards Testing Critical Infrastructure Resilience. Publications Office of the European Union, Luxembourg (2014). <https://doi.org/10.2788/41633>
20. Java, O., Asprion, B., Priebe, T., Sarkozi, E., Neves Madeira, R.: Application of digital technology in agriculture: potential support for winegrowers. In: Proceeding of the 8th International Conference on Trends in Agricultural Engineering 2022. Prague (2022). <https://2022.tae-conference.cz/proceeding/TAE2022-32-Oskars-JAVA.pdf>
21. Kazim, E., Denny, D.M.T., Koshiyama, A.: AI auditing and impact assessment: according to the UK information commissioner’s office. *AI Ethics* **1**, 301–310 (2021). <https://doi.org/10.1007/s43681-021-00039-2>
22. Klinkenberg, R.: Learning drifting concepts: example selection vs. example weighting. *Intell. Data Anal.* **8**(3), 281–300 (2004). <https://dl.acm.org/doi/10.5555/1293831.1293836>
23. Kong, Z., Xue, J., Wang, Y., Huang, L., Niu, Z., Li, F.: A survey on adversarial attack in the age of artificial intelligence. *Wirel. Commun. Mob. Comput.* **2021**, TBD (2021). <https://doi.org/10.1155/2021/4907754>
24. Kosow, H., Gaßner, R.: Methods of future and scenario analysis: overview, assessment, and selection criteria, vol. 39. DEU (2008)
25. KPMG: AI Risk and Controls Matrix (2018). <https://assets.kpmg/content/dam/kpmg/uk/pdf/2018/09/ai-risk-and-controls-matrix.pdf>
26. Larson, J., Mattu, S., Kirchner, L., Angwin, J.: How We Analyzed the COMPAS Recidivism Algorithm. *ProPublica* (2016). <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>
27. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. *ACM Comput. Surv.* **55**(6) (2022). <https://doi.org/10.1145/3457607>
28. Meijer, A., Wessels, M.: Predictive policing: review of benefits and drawbacks. *Int. J. Public Adm.* **42**(12), 1031–1039 (2019)
29. Neves Madeira, R., et al.: Towards digital twins for multi-sensor land and plant monitoring. *Procedia Comput. Sci.* **210**, 45–52 (2022). <https://doi.org/10.1016/j.procs.2022.10.118>
30. Rammert, W.: Where the action is: distributed agency between humans, machines, and programs. In: Seifert, U., Kim, J.H., Moore, A. (eds.) *Paradoxes of Interactivity: Perspectives for Media Theory, Human-Computer Interaction, and Artistic*

- Investigations, pp. 62–91. transcript, Bielefeld (2008). <https://doi.org/10.14361/9783839408421-004>
31. Righetti, L., Pham, Q.C., Madhavan, R., Chatila, R.: Lethal autonomous weapon systems [ethical, legal, and societal issues]. *IEEE Robot. Autom. Mag.* **25**(1), 123–126 (2018)
 32. Tao, F., Qi, Q., Wang, L., Nee, A.Y.C.: Digital twins and cyber-physical systems toward smart manufacturing and industry 4.0: correlation and comparison. *Engineering (Beijing)* **5**(4), 653–661 (2019). <https://doi.org/10.1016/j.eng.2019.01.014>
 33. Tjoa, S., Buttinger, C., Holzinger, K., Kieseberg, P.: Penetration testing artificial intelligence. *ERCIM News* **123**, 36–37 (2020). <https://ercim-news.ercim.eu/en123/r-i/penetration-testing-artificial-intelligence>
 34. Tobin, J., et al.: Domain randomization for transferring deep neural networks from simulation to the real world. In: 2017 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, pp. 23–30. IROS 2017, IEEE, Washington, DC (2017). <https://doi.org/10.1109/IROS.2017.8202133>
 35. United Nations: Transforming our World: The 2030 Agenda for Sustainable Development (2015). <https://sdgs.un.org/2030agenda>, resolution No. A/RES/70/1
 36. Vermaas, P., Kroes, P., van de Poel, I., Franssen, M., Houkes, W.: A Philosophy of Technology: From Technical Artefacts to Sociotechnical Systems, Synthesis Lectures on Engineers, Technology, and Society, vol. 17. Morgan & Claypool, San Rafael, CA (2011). <https://doi.org/10.2200/S00321ED1V01Y201012ETS014>
 37. Winter, P.M., et al.: Trusted Artificial Intelligence: Towards Certification of Machine Learning Applications. TÜV Austria, Brunn am Gebirge (2021). <https://www.tuv.at/loesungen/digital-services/trusted-ai>
 38. Yudkowsky, E.: The ai alignment problem: why it is hard, and where to start. Symbolic Systems Distinguished Speaker (2016). <https://intelligence.org/stanford-talk/>
 39. Zelaya, C.V.G.: Towards explaining the effects of data preprocessing on machine learning. In: Proceedings of the 35th International Conference on Data Engineering. pp. 2086–2090. ICDE '19, IEEE Computer Society, Washington, DC (2019). <https://doi.org/10.1109/ICDE.2019.00245>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

