"Don't let me be misunderstood": Critical AI literacy for the constructive use of AI technology

Strauß, Stefan

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Empfohlene Zitierung / Suggested Citation:
Strauß, S. (2021). "Don't let me be misunderstood": Critical AI literacy for the constructive use of AI technology. *TATuP - Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis / Journal for Technology Assessment in Theory and Practice*, *30*(3), 44-49. https://doi.org/10.14512/tatup.30.3.44

OPEN ACCESS oekom

RESEARCH ARTICLE

"Don't let me be misunderstood"

Critical AI literacy for the constructive use of AI technology

Stefan Strauß, *Institut für Technikfolgen-Abschätzung (ITA), Österreichische Akademie der Wissenschaften, Apostelgasse 23, 1030 Wien, AT (sstrauss@oeaw.ac.at)* *0000-0003-1877-2415*

Abstract • Research and development as well as societal debates on the risks of artificial intelligence (AI) often focus on crucial but impractical ethical issues or on technocratic approaches to managing societal and ethical risks with technology. To overcome this, more practical, problem-oriented analytical perspectives on the risks of AI are needed. This article proposes an approach that focuses on a meta-risk inherent in AI systems: deep automation bias. It is assumed that the mismatch between system behavior and user practice in specific application contexts due to AI-based automation is a key trigger for bias and other societal risks. The article presents the main factors of (deep) automation bias and outlines a framework providing indicators for the detection of deep automation bias ultimately triggered by such a mismatch. This approach intends to strengthen problem awareness and critical AI literacy and thereby create some practial use.

„Don't let me be misunderstood". *Kritische KI-Kompetenz für den konstruktiven Umgang mit KI-Technologie*

Zusammenfassung • *Gesellschaftlicher Diskurs sowie Forschung und Entwicklung zu Risiken künstlicher Intelligenz (KI) fokussieren oft einseitig entweder auf praxisferne ethische Aspekte oder auf technokratische Ansätze zur Bewältigung gesellschaftlicher Risiken allein durch Technologie. Es bedarf jedoch praktikabler, problemorientierter Perspektiven. Dieser Beitrag konzentriert sich daher auf ein zentrales Meta-Risiko von KI-Systemen: Deep Automation Bias. Es wird davon ausgegangen, dass Diskrepanzen zwischen Systemverhalten und Nutzungspraktiken in bestimmten Anwendungskontexten aufgrund KI-basierter Automatisierung zentrale Auslöser von Bias und gesellschaftlichen Risiken sind. Der Beitrag stellt zentrale Faktoren von (Deep) Automation Bias vor und entwickelt einen analytischen Rahmen mit Indikatoren zur Erkennung von Diskrepanzen in KI-Systemen. Dieser Ansatz will durch Stärkung von Problembewusstsein und kritischer KI-Kompetenz auch praktischen Nutzen erzielen.*

https://doi.org/10.14512/tatup.30.3.44
Received: Jun. 14, 2021; revised version accepted: Oct. 18, 2021; published online: Dec. 20, 2021 (peer review)

Keywords • *deep automation bias, AI assessment, machine learning, uncertainty, awareness*

Introduction

The hype around artificial intelligence (AI) is yet unbroken. Machine learning (ML) algorithms gain influence on economic, social and political decisions affecting individuals directly and indirectly. Accordingly, there is a scientific and political debate on how to tackle the various ethical risks of a broader use of AI. These discussions are, though, mostly dominated by either general ethical issues such as human versus machine autonomy, matters of trust, fairness, accountability and transparency (FAT) or on technical solutions to avoid algorithmic discrimination. Correspondingly, there is a number of guidelines for "ethical AI" or "trustworthy AI" issued by the EU Commission's high-level expert group on AI and others (Floridi et al. 2018; HLEG 2019; AlgorithmWatch 2019; Hallensleben 2020). And a growing community deals with developing technical solutions for de-biasing and FAT-ML, for example in the annual ACM-FAT conferences (Selbst et al. 2019; Wieringa 2020; Eid et al. 2021). Without doubt, this involves various relevant research and development activities.

But there is also a certain gap between important but impractical ethical concepts on the one side and technocratic approaches to fix societal problems with algorithms on the other. Not without irony, this situation could even reinforce the myriad of AI-related risks ranging from bias and discrimination, lacking transparency, erosion of privacy and security, loss of autonomy etc. There is thus need for a broader debate and problem-oriented approaches on how to effectively comprehend and conceptualize socio-technical risks related to AI.

A main argument of this paper[1] is that AI-based automation plays a particular role here. To explore the risks of AI thus requires a stronger analytical focus on automation. To facilitate

1 Parts of this paper represent a condensed and modified version of Strauß (2021).

this, I identified deep automation bias (DAB) as a meta-risk of the societal use of AI entailing further risks. DAB is a multidimensional, wicked problem inherent to AI technology alluding to progress in deep learning and self-optimizing algorithms (Strauß 2018, 2021). The aim is to develop this concept of DAB further and propose it as part of a problem-oriented assessment framework of AI. The premise here is that essentially, AI-based technology represents a socio-technical system that fosters automation at different levels. Bias can result from pre-existing prejudice during technical development, technical issues like poor data quality, insufficient models or inappropriate operation of ML-algorithms; but also from rule conflicts between AI design and AI application contexts due to complexity gaps between statistical assumptions in the system and user practices. In each case, the common denominator is automation, though, on different socio-technical levels.

To understand how these levels interact requires a multilayer view on the interplay between design and use of AI technology which together shape societal impacts. The main focus of the paper is thus on how to improve the analytical perspective on AI as a socio-technical issue to foster the basic understanding and awareness on the related societal challenges. This is a contribution towards what I call 'critical AI literacy' to avoid the fallacy of seeking for technological fixes for societal problems. The paper is structured as follows: after this introduction, section two briefly discusses why AI bears wicked problems which cannot be addressed with technical means only. Section three then sheds light on critical AI literacy and the role of automation. Based on main factors affecting DAB section four sketches a problem-oriented assessment framework. Section five presents a short summary and concluding remarks.

Wicked problems require more than fairness, accountability and transparency

As several scholars argue, there is need for alternative socio-technical approaches to better grasp the societal and ethical issues of AI (Edwards and Veale 2017; Selbst et al. 2019; Tsamados et al. 2020). This is particularly relevant as the use of AI systems can involve and reinforce so-called wicked problems (Strauß 2021). They are a "'class of social system problems which are ill-formulated, where the information is confusing, where there are many clients and decision makers with conflicting values, and where the ramifications in the whole system are thoroughly confusing'" (Churchman 1967 cited in Buchanan 1992, p. 15). Wicked problems bear tensions between the *artificial* and the *natural* (ibid.). This basic conflict can reinforce with the use of AI, particularly due to its high degree of automation: AI transforms decision-making and entails risks of reducing *natural* aspects of society to machine-readable data models that are interpretable by *artificial* algorithms.

Bias in ML is a wicked problem inherent to AI. However, unbiasing and fostering FAT is not sufficient to avoid the related risks of undetected failure, self-fulfilling prophecies and an incremental normalization of AI biases in society. Sheer techno-fixes could even intensify these risks. Research on FAT and bias in ML is dominated by debates on how different types occur, i. e., preexisting, technical or emergent bias and how to avoid that AI and algorithms lead to discrimination and injustice (Friedman and Nissenbaum 1996; Simon et al. 2020; Wieringa 2020). This is important work but there is a tendency to frame this socio-technical issue as a technological one or to get lost in general ethical debates on fairness, justice etc. and seeking technical solutions to ethical problems. This can be counterproductive. Unbiasing approaches, e. g., with adaptive algorithms, may increase complexity and opacity of AI which further reinforce societal risks.

Obviously, not just the technical design of AI is relevant but in particular, how (in-)compatible the technical system is with its socio-technical application contexts. This is crucial to tackle the risks of AI, which requires a broader, problem-oriented perspective that fosters analytical views on both, technical and societal issues of AI systems. To circumvent one-sided views, like ethical debates beyond practicability or misleading technocratic approaches and underestimation of risks, raising problem-awareness among decision-makers and persons interacting with AI systems is essential. However, as of yet, there is a lack of awareness and analytical perspectives in this regard. I thus suggest to focus more on the specific role of automation in AI and how to establish what I call here critical AI literacy. Critical AI literacy here means the ability to comprehend the core features of an AI system and its (in-)compatibility with its particular application contexts in a (necessarily) more complex sociotechnical reality.

Beyond practicability or misleading technocratic approaches and underestimation of risks, raising problem-awareness among decision-makers and persons interacting with AI systems is essential.

Critical AI literacy: understanding AI-based automation (bias)

A crucial question for the use of AI is whether it matches with the requirements of a particular application context. This implies that the contextual environment of an AI system affects the occurrence of bias. Tsamados et al. (2020) discuss context bias on the example of a healthcare system for resource management in hospitals. The system may function properly for one hospital that fits to the model the system uses but may cause problems in others, e. g., rural clinics with different contextual factors. But as argued, at the core, the various risks of AI ultimately derive from conflicts due to different forms of automation. Automation bias (AB) is the general risk of uncritically accepting the outcome of an automated system (Goddard et al. 2012, 2014). AI intensifies this risk and thus DAB represents a meta-risk of AI. The following examples highlight this:

Even very simple forms of automation can cause serious problems as the case of the automated renaming function in Excel tables shows: studies detected failure rates of 20 per cent implying that every third table containing genetic data presents false information as gene names are automatically renamed to dates (e. g., MARCH1 to 1-Mar). Abeysooriaya et al. (2021) show that this problem still exists and recommend human workarounds. Thus, even simple errors may create severe impact. Particularly, if these errors remain undetected and are processed further by AI systems.

Imagine an autopilot-system of an airplane, a classical form of automation. Basically, it is a rule-based system which functions with sensors and real-time data on geolocation, weather etc. Hence it needs a plausible data model of the plane's environment and reliable information on its behavior so that the human pilot can monitor if autopilot and plane operate as intended and can intervene immediately in case of problems. Any hidden error like a faulty label in a data table could threaten human lives. Recent cases of military drones autonomously attacking soldiers in Libya highlight that this is not a sheer theoretical risk (Hambling 2021). AB is a known risk of autopilots (Parasuraman et al. 2010; Goddard et al. 2014), mitigated with extensive training and technical features to improve controllability and avoid overreliance on the system. A precondition here is the basic predictability of system behavior and comprehensible rules determining its functionality. Hence system complexity must remain manageable. An autopilot that would permanently try to optimize a flight (e. g. with some predictability algorithm) without effective human intervention would be uncontrollable. The system would lever out human autonomy and agency and the conflict between system behavior and human intervention could escalate at any time. Tackling this risk requires more than transparency, accountability or explicability and is impossible without plausibility, reliability, predictability and effective intervenability to comprehend and correct the automated system.

Further examples are AI systems for job applications which evidently led to discrimination in various cases. As Harwell (2019) illustrates, the hiring platform HireVue calculated an "employability score" based on various data on job applicants including facial expressions and speech. Critics filed complaint and argued the system is biased, unfair and deceptive as it discriminates, e. g., due to different facial looks and spoken accents. Another system uses background images in applicants' portray photos to predict job qualification (Harlan and Schnuck 2021). For example, a person standing in front of a bookshelf then has higher chances to get a job offer for certain job sectors than a person submitting a photo with plain background. Obviously, skin color, ethnicity and background images have no relevance for a person's qualification. But people of color or persons with lower contrasting background images generally get lower scores. Hence the system reinforces racial and other forms of discrimination. This is an evident issue of various other AI systems, too. Various cases (O'Neil 2016; Borgesius 2018; Obermeyer et al. 2019; Köchling and Wehner 2020) demonstrate, how problematic it can be to automate social domains with AI. They underline the risk of DAB which is inevitable here if neither job applicants nor recruiters are unaware of the problem and no countermeasures to avoid discrimination are set. In any case, the system behaves unfair and unreliable.

Technical fixes like de-biasing to fix deficient image processing do not solve such problems as they are more than just technical issues. A typical technical solution to the above-mentioned bias would be to modify the algorithm so that it excludes image backgrounds when calculating a qualification score. This may ease bias resulting from images but any other problems with criteria the algorithm may process (e. g., ethnic facial features, residential district) would remain unsolved. Also, FAT is ineffective as transparency on the issue would not prevent from discrimination. Moreover, there are various cases of bias or stereotyping in data models and ML approaches with problematic effects on system behavior. Particularly sensitive is the use of AI in the health domain. Several studies reveal problems and unintended effects of decision-making systems here (Goddard et al. 2012; Cabitza et al. 2017; Gianfrancesco et al. 2018; Obermeyer et al. 2019). Gianfrancesco et al. (2018, p. 5) analyzed ML algorithms in clinical applications and found serious issues such as "overreliance on automation, algorithms based on biased data, and algorithms that do not provide information that is clinically meaningful". They conclude that easing these problems requires better understanding of AI and their ML approaches and corresponding measures to achieve this.

Towards a problem-oriented assessment framework

To raise problem-awareness, it is essential to understand how AI-based automation operates and how DAB occurs. The factors and framework presented here are meant as an approach to improve critical AI literacy. Basically, AI becomes problematic when there is a mismatch between system behavior and user

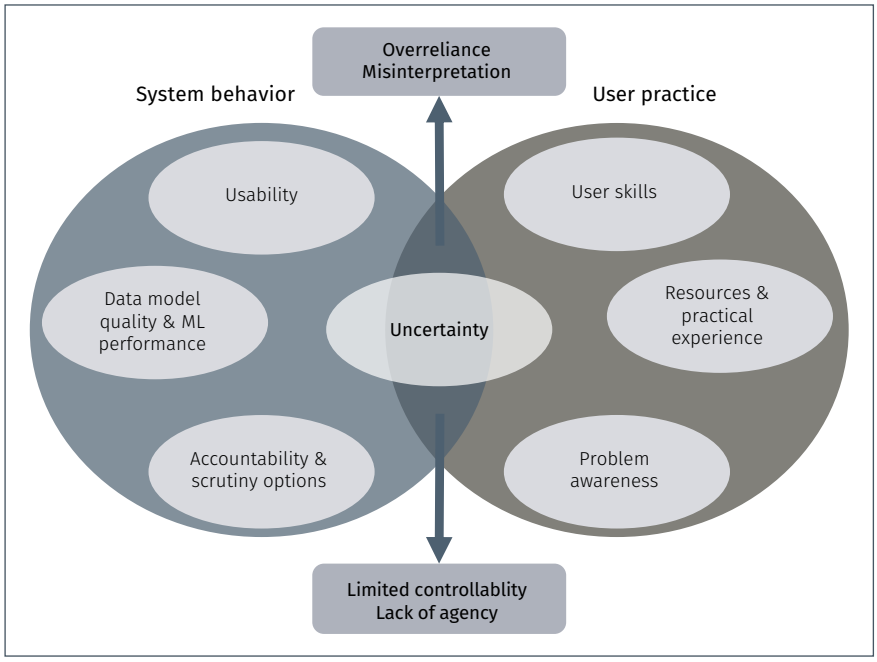

Fig. 1: Factors affecting DAB. *Source: Strauß 2021, p. 8*

practice in specific application contexts. To comprehend the meaning of mismatch is a precondition for the assessment of AI-related risks. This implies to understand the peculiarities of AI-based automation. Because irrespective of specific features, every AI system uses some form of automation and bears risks of (D)AB. But as Tsoukiàs (2020) reminds, automation is not inevitable with AI. It is a choice that needs to be legitimated and not an end in itself. It is crucial to scrutinize the automation approach of an AI system when assessing its impact. As a first step to sharpen the analytical lens, I suggest to conceptualize DAB as meta-risk of AI from which other risks derive from.

Main determinants of DAB

DAB addresses the problem of increasing complexity and opacity of AI technology that reinforce AB due to its dynamic, unpredictable and thus potentially uncontrollable behavior (Strauß 2018, 2021). Sophisticated ML approaches with features to self-optimization like deep learning or other forms of unsupervised learning aggravate this problem. It can extend the gap between human propensity to blindly trust AI technology and the limits of technology to match social complexity due to necessarily reductionistic models of society. The wider risk here is that dependence of society on AI systems reaches a stage where automated decisions – no matter if socially acceptable, ethical, correct or false – become uncontrollable. AI then constantly reshapes society and individual lives without effective alternatives. Figure 1 shows basic factors affecting DAB.

DAB bears at least two main risks: at the top is the "classical" risk of overreliance on the behavior of an AI system and misinterpretation; the bottom shows the additional risk of limited controllability and lack of agency. Both dimensions are interrelated and the severity of DAB depends on the interplay of different factors. The main connecting factor is uncertainty, shaped by technical as well as social issues that can reinforce mutually. System behavior strongly depends on the quality of data models and ML performance, usability, accountability and scrutiny options in the system. The social perspective involves user practices: various studies show that user skills, practical experience, resources (e. g. user knowledge to interpret a system, time and pressure to act), workload and effective options to scrutinize automated procedures affect AB (Goddard et al. 2012, 2014; Lyell and Coiera 2016). DAB further complicates these factors as AI increases system complexity, opacity and decreases options to scrutinize its functionality. Consequently, controllability, agency and options to intervene into automated decision-making can also decrease. The interplay of all these factors affect the severity of DAB and related risks. It is particularly higher, when the system lacks in options to scrutinize its behavior and/or problem awareness of the human user is low. For example, lacking accountability reduces agency and low problem awareness limits the user's ability to scrutinize, which again limits agency (Strauß 2021).

How to assess system behavior

Figure 2 sketches a three-level framework with indicators to identify DAB-related risks. The basic idea is to provide a simple checking tool to detect obscurities in system behavior. The focus is on the operational level as DAB risks become most apparent here.

The four main indicators (explicability, validity, plausibility and acceptability) represent a toolbox to check if the system operates properly. The related guiding questions apply to the whole operation process from input, output to action or decision. Any case of doubt or uncertainty triggers a more detailed review at all levels to uncover eventual mismatch between system behavior and user practice, technical flaws, or legal or ethical problems such as violation of human rights, discrimination etc. If all indicators point to normal system behavior, no DAB risk was detected. But regular detailed reviews of system behavior including all levels are advisable to avoid instability or any other issues. Obviously, all assessment levels are intertwined, but the schematic distinction supports comprehension of whether and how DAB occurs. If the system malfunctions because the data model is biased, then the system as a whole is biased. If the data model is OK but the system behavior is implausible there might be a

different reason; and if a system is basically explicable, its outcome is not necessarily ethical. For users interacting or testing AI, ethical reviews are impractical. It makes little sense to ask, for example, whether the system affects autonomy during operation. But it makes sense having some indicators to assess how exposed system behavior is to DAB-related risks. Acceptability is thus drawn at the intersection between operational and ethical level, because an unacceptable outcome during operation may indicate a severe legal or ethical issue which then needs to be analyzed further.

Briefly, applying the framework to the afore mentioned case of AI for applications underlines the necessity of different assessment levels. In a sheer technical sense, there might be no problem observable. Consider a typical ML framework with preexisting external data models embedded in the system from a trustful source. Without technical and operational assessment, revealing bias in the system, for example due to its mode of image processing, is impossible. To effectively assess how the system operates requires knowledge about technical features and how they affect the preselection of job applicants. The assessment would then show that the system is neither explicable nor valid, nor plausible, nor acceptable. Consequently, a deeper, ethical assessment would be triggered which then would reveal that the system approach in total is unethical as it discriminates persons based on irrelevant data. More detailed testing of the frameworks' practicability, as presented in this article, is subject to further research.

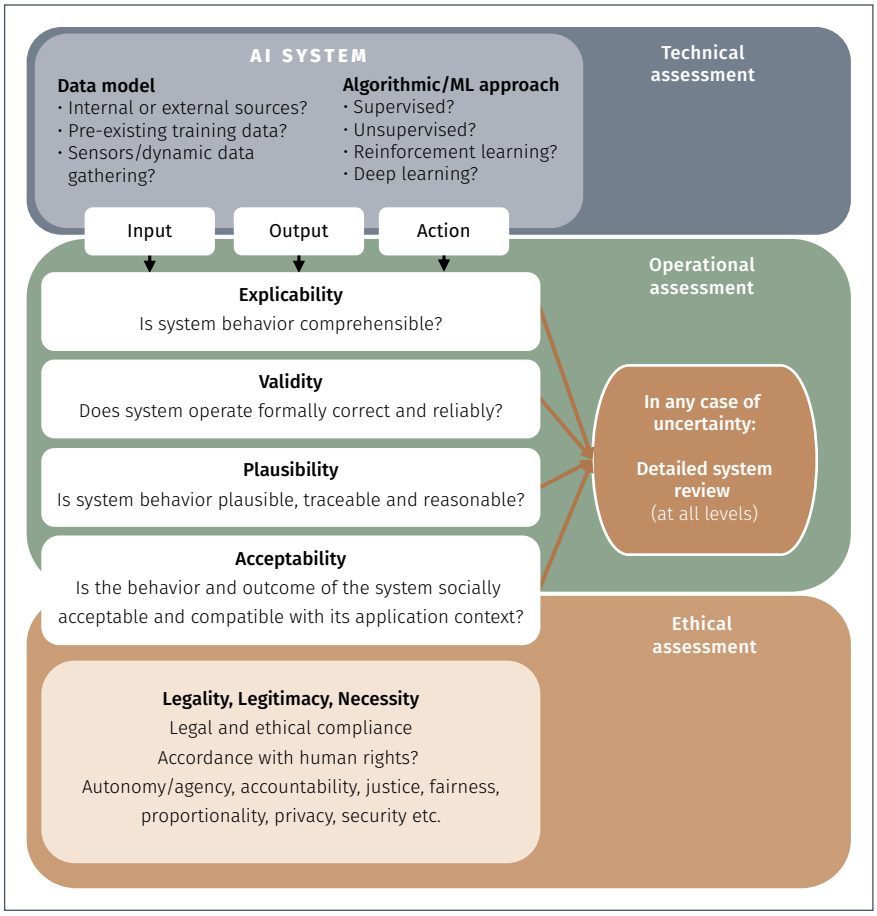

Fig. 2: A problem-oriented assessment framework. *Source: enhanced version of Strauß 2021, p. 9*

Concluding remarks

„I am just a soul whose intentions are good; oh lord please don't let me be misunderstood". This refrain of the famous song, first interpreted by Nina Simone in 1964, highlights the societal dilemma of the broader use of AI: (mis-)understanding is a key issue. AI bears high potential to transform society and there are many "good" intentions behind AI-based innovation. But good intentions, such as causing no harm and creating benefit, are not enough to tackle the myriad of risks and prevent AI from becoming a severe threat to society. The crux are misconceptions between technology design, specific application contexts and individual, institutional, societal and ethical requirements. Attempts to resolve them by integrating ethics into AI systems is like trying to explain to an automated vehicle why it should not cause accidents. This is doomed to fail because machines, irrespective of the degrees of their mechanical or digital automation processes, remain machines. Any such attempt aggravates ethical problems. To avoid that humans become "slave to the algorithm" (Edwards and Veale 2017, p. 1) we need more knowledge on the risks of AI and better strategies to cope with them. As a step in this direction, this paper suggests to foster critical AI literacy based on a problem-oriented approach with an explicit focus on DAB-related risks as trigger for further risks of AI. This approach is meant as awareness-raising tool which may also be of some practical use. The intention behind it is rather simple: to envision non-dystopian futures requires novel perspectives on AI to overcome technocratic approaches and revitalize humanistic perspectives on how to deal with AI in a constructive, socially acceptable manner. This can only work if all stakeholders, engineers, designers, policy makers, users and other persons concerned are aware of the factual risks and find ways to reduce them.

Acknowledgement

This research article has not received any external funding.

References

Abeysooriya, Mandhri; Soria, Megan; Kasu, Mary; Ziemann, Mark (2021): Gene name errors. Lessons not learned. In: PLoS Computational Biology 17 (7), p. e1008984. https://doi.org/10.1371/journal.pcbi.1008984

AlgorithmWatch (2019): Automating society. Taking stock of automated decision-making in the EU. Available online at https://www.algorithmwatch.org/automating-society, last accessed on 12. 10. 2021.

Borgesius, Frederik (2018): Discrimination, artificial intelligence, and algorithmic decision-making. Study for the Council of Europe. Strasbourg: DG of Democracy.

Buchanan, Richard (1992): Wicked problems in design thinking. In: Design Issues 8 (2), pp. 5–21. https://doi.org/10.2307/1511637

Cabitza, Federico; Rasoini, Raffaele; Gensini, Gian (2017): Unintended consequences of machine learning in medicine. In: JAMA 318 (6), pp. 517–518. https://doi.org/10.1001/jama.2017.7797

Edwards, Lilian; Veale, Michael (2017): Slave to the algorithm? Why a 'right to explanation' is probably not the remedy you are looking for. In: Duke Law & Technology Review 16 (1), pp. 18–84. https://doi.org/10.2139/ssrn.2972855

Eid, Fatma-Elzahraa et al. (2021): Systematic auditing is essential to debiasing machine learning in biology. In: Communications Biology 4 (183), p. 1–9. https://doi.org/10.1038/s42003-021-01674-5

Floridi, Luciano et al. (2018): AI4People – an ethical framework for a good AI society. Opportunities, risks, principles, and recommendations. In: Minds & Machines 28, pp. 689–707. https://doi.org/10.1007/s11023-018-9482-5

Friedman, Bataya; Nissenbaum, Helen (1996): Bias in computer systems. In: ACM Transactions on Information Systems, 14 (3), pp. 330–347. https://doi.org/10.1145/230538.230561

Gianfrancesco, Milena et al. (2018): Potential biases in machine learning algorithms using electronic health record data. In: JAMA Internal Medicine 178 (11), pp. 1544–1547. https://doi.org/10.1001/jamainternmed.2018.3763

Goddard, Kate; Roudsari, Abdul; Wyatt, Jeremy (2012): Automation bias. A systematic review of frequency, effect mediators, and mitigators. In: Journal of the American Medical Informatics Association 19 (1), pp. 121–127. https://doi.org/10.1136/amiajnl-2011-000089

Goddard, Kate; Roudsari, Abdul; Wyatt, Jeremy (2014): Automation bias. Empirical results assessing influencing factors. In: International Journal of Medical Informatics 83 (5), pp. 368–375. https://doi.org/10.1016/j.ijmedinf.2014.01.001

Hallensleben, Sebastian et al. (2020): From principles to practice. An interdisciplinary framework to operationalise AI ethics. Gütersloh: Bertelsmann Stiftung. Available online at https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/WKIO_2020_final.pdf, last accessed on 12. 10. 2021.

Hambling, David (2021): Drones may have attacked humans fully autonomously for the first time. In: New Scientist, 27. 05. 2021. Available online at https://www.newscientist.com/article/2278852-drones-may-have-attacked-humans-fully-autonomously-for-the-first-time/, last accessed on 12. 10. 2021.

Harlan, Elisa; Schnuck, Oliver (2021): Objective of biased? On the questionable use of artificial intelligence for job applications. Available online at https://web.br.de/interaktiv/ki-bewerbung/en/, last accessed on 12. 10. 2021.

Harwell, Drew (2019): A face-scanning algorithm increasingly decides whether you deserve the job. In: Washington Post, 06. 11. 2019. Available online at www.washingtonpost.com/technology/2019/10/22/ai-hiring-face-scanning-algorithm-increasingly-decides-whether-you-deserve-job/, last accessed on 12. 10. 2021.

HLEG – High-Level Expert Group on Artificial Intelligence (2019): Ethics guidelines for trustworthy AI. Available online at https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419, last accessed on 28. 09. 2021.

Köchling, Alina; Wehner, Marius (2020): Discriminated by an algorithm. A systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. In: Business Research 13 (3), pp. 795–848. https://doi.org/10.1007/s40685-020-00134-w

Lyell, David; Coiera, Enrico (2016): Automation bias and verification complexity. A systematic review. In: Journal of the American Medical Informatics Association 24 (2), pp. 424–431. https://doi.org/10.1093/jamia/ocw105

O'Neil, Cathy (2016): Weapons of math destruction. How big data increases inequality and threatens democracy. New York, NY: Crown.

Obermeyer, Ziad; Powers, Brian; Vogeli, Christine; Mullainathan, Sendhil (2019): Dissecting racial bias in an algorithm used to manage the health of populations. In: Science (336), pp. 447–453. https://doi.org/10.1126/science.aax2342

Parasuraman, Raja; Manzey, Dietrich (2010): Complacency and bias in human use of automation. An attentional integration. In: The Journal of the Human Factors and Ergonomics Society 52 (3), pp. 381–410. https://doi.org/10.1177/0018720810376055

Selbst, Andrew; Boyd, Danah; Friedler, Sorelle; Venkatasubramanian, Suresh; Vertesi, Janet (2019): Fairness and abstraction in sociotechnical systems. In: Association for Computing Machinery New York, NY (ed.): FAT* '19, Proceedings of the conference on fairness, accountability, and transparency, pp. 59–68. https://doi.org/10.1145/3287560.3287598

Simon, Judith; Wong, Pak-Hang; Rieder, Gernot (2020): Algorithmic bias and the value sensitive design approach. In: Internet Policy Review (9) 4, p. 1–16. https://doi.org/10.14763/2020.4.1534

Strauß, Stefan (2018): From big data to deep learning. A leap towards strong AI or 'intelligentia obscura'? In: Big Data and Cognitive Computing 2 (3), pp. 1–19. https://doi.org/10.3390/bdcc2030016

Strauß, Stefan (2021): Deep automation bias. How to tackle a wicked problem of AI? In: Big Data and Cognitive Computing 5 (2), pp. 1–14. https://doi.org/10.3390/bdcc5020018

Tsamados, Andreas et al. (2020): The ethics of algorithms. Key problems and solutions. Available online at SSRN's eLibrary. https://doi.org/10.2139/ssrn.3662302

Tsoukiàs, Alexis (2020): Social responsibility of algorithms. An overview. Available online at https://arxiv.org/pdf/2012.03319.pdf, last accessed on 12. 10. 2021.

Wieringa, Maranke (2020): What to account for when accounting for algorithms. A systematic literature review on algorithmic accountability. In: Association for Computing Machinery New York, NY (ed.): FAT* '19, Proceedings of the conference on fairness, accountability, and transparency, pp. 1–18. https://doi.org/10.1145/3351095.3372833

DR. STEFAN STRAUSS

is senior scientist at the Institute of Technology Assessment (ITA) at the Austrian Academy of Sciences. His main research focus is on the interplay between technology and society, governance of socio-technical systems and the question how digitization affects social practices, human rights, policy and value systems.