big data and
cognitive computing

Article

Deep Automation Bias: How to Tackle a Wicked Problem of AI?

Stefan Strauß

Institute of Technology Assessment (ITA), Austrian Academy of Sciences, 1030 Vienna, Austria

Abstract: The increasing use of AI in different societal contexts intensified the debate on risks, ethical problems and bias. Accordingly, promising research activities focus on debiasing to strengthen fairness, accountability and transparency in machine learning. There is, though, a tendency to fix societal and ethical issues with technical solutions that may cause additional, wicked problems. Alternative analytical approaches are thus needed to avoid this and to comprehend how societal and ethical issues occur in AI systems. Despite various forms of bias, ultimately, risks result from eventual rule conflicts between the AI system behavior due to feature complexity and user practices with limited options for scrutiny. Hence, although different forms of bias can occur, automation is their common ground. The paper highlights the role of automation and explains why deep automation bias (DAB) is a metarisk of AI. Based on former work it elaborates the main influencing factors and develops a heuristic model for assessing DAB-related risks in AI systems. This model aims at raising problem awareness and training on the sociotechnical risks resulting from AI-based automation and contributes to improving the general explicability of AI systems beyond technical issues.

Keywords: artificial intelligence; machine learning; automation bias; fairness; transparency; accountability; explicability; uncertainty; human-in-the-loop; awareness raising

Citation: Strauß, S. Deep Automation Bias: How to Tackle a Wicked Problem of AI? *Big Data Cogn. Comput.* **2021**, *5*, 18. https://doi.org/10.3390/bdcc5020018

Academic Editors: Nigel Houlden and Vic Grout

Received: 12 March 2021
Accepted: 15 April 2021
Published: 20 April 2021

1. Introduction

Global society is subjected to a sociotechnical transformation writ large that is essentially driven by artificial intelligence (AI) and machine learning (ML). This development increasingly affects economic, social and political decision-making. The impact of AI is thus expected to be enormous and accordingly, there is also a growing debate on the societal risks and ethical problems related to the broader diffusion of AI systems in society. This paper explores these issues and proposes a conceptual approach in order to facilitate dealing with AI in a more constructive, risk-aware way. The main focus is on deep automation bias (DAB), which I identified in [1] as a crucial problem that is inherent to the use of AI in societal contexts. DAB represents a systemic metarisk that can be a useful reference point to assess the risks of AI systems. Based on this previous work, the paper further develops the approach. It discusses scope and different dimensions of DAB and how this metarisk could be conceptualized to be useful for detecting related risks in AI systems.

There is little doubt that AI has great innovation potential. However, there are also a number of serious risks that need to be tackled in order to avoid AI becoming one of the most severe technological threats to society ever. Even though the use of AI aims to benefit society, it is not sufficient to address risks with technical means only and assume that potential harm to society would be exaggerated. Apparently, it is always a matter of how technology is applied in particular application contexts and how its design relates to the corresponding individual, institutional, societal and ethical requirements. The crux is how to assess this and how to understand what kind of risks the use of AI bears. A main objective of this paper therefore is, to explore potential answers to this question by analyzing the role of automation in AI and how it affects the severity of sociotechnical risks. The focus is on how to improve the general understanding of the societal impact of AI with an analytical view on DAB, framed as a metarisk that can entail further risks. The paper is thus a contribution to strengthen a problem-oriented approach to deal with the societal challenges of AI technology. This is important because as of yet, discussions on the risks of AI seem to be dominated by a tendency to seek for technological fixes, e.g., to reduce bias in ML. This is a fallacy because societal problems cannot be fixed by technological means only. Several scholars thus argue that there is a need for approaches to better grasp the societal and ethical issues of AI ([1–5]). In line with this, the paper explores the multiple factors of DAB as sociotechnical metarisk and how this could support AI risk assessment.

The paper is structured as follows: after this introduction, Section 2 argues that the current emphasis on fairness, accountability and transparency in ML (FAT-ML) is not sufficient to tackle the risks of AI. Section 3 then discusses whether the use of AI leads to so-called wicked problems. Based on this, Section 4 then explains why DAB is a crucial issue, what main factors drive its occurrence and how they interrelate. Section 5 discusses the specific role of automation in AI systems and its societal issues based on particular examples. Section 6 then outlines a heuristic model to map and identity the occurrence of DAB in an AI system. Finally, Section 7 of the paper provides a brief summary and concluding remarks.

2. Why Fairness, Accountability and Transparency Are Not Enough

With AI becoming more widespread, there is an intensified debate on issues regarding algorithmic bias, discrimination and debiasing by fostering FAT in ML. Many scholars from various disciplines have been pointing out that algorithms and particularly ML are biased, unfair and lack in transparency and accountability ([1–12]). To deal with these serious issues, there is important research proceeding in the AI/ML community, particularly in computer science, data science and related disciplines aiming at developing approaches for debiasing and improving FAT ([3,4,12–15]). Correspondingly, there is a growing research community dealing with these issues. This is observable, e.g., in annual conferences on FAT/ML, algorithmic transparency etc. (e.g., the annual ACM FaccT conferences formerly known as FAT*ML, www.fatml.org, accessed on 12 March 2021, or specific workshops like Fair ML for Health). However, these research activities are largely dominated by a quest for technical solutions to reduce bias in data sets, e.g., by integrating automated debiasing functionality into AI systems (e.g., such as IBM's OpenScale approach). While it is obviously relevant to tackle these issues of data quality and technical bias, there is also a need for approaches to address the societal and ethical issues of AI. Several scholars thus argue that bias is not merely a technical issue ([1,3–6,11–17]). The complexity of the problem is already observable in the various different, and partially contradictory notions and definitions of fairness ([5,11]). Similar is the case for the other concepts of FAT. There is no general definition for concepts such as fairness, unbiased, neutral, etc. [11]. Transparency, which is basically essential, and which is not just an ethical principle but an enabling condition for ethical practices ([5,17]), can also have ambiguous meanings. From a technical design perspective, transparency can even mean hiding information from users, with the aim to reduce complexity and ease understandability by avoiding information overload. This can be in opposition to common understandings of transparency as comprehensibility (cf. [18], (p. 67ff.). Hence, there is a paradox situation as technical approaches to tackle bias might create additional bias with further societal impact.

In spite of the currently strong research focus on bias in ML, the problem as such, bias in digital technology, is nothing new. In their work, Friedman and Nissenbaum (1996) argued for a better understanding of bias in computer systems, which can lead to discrimination of individuals or groups. Biased systems that affect individuals bear the risk of becoming "instruments of injustice" [6], (p. 345f.). They differ in three types of bias: (1) preexisting, (2) technical and (3) emergent bias. (1) Preexisting bias exists in society, in social institutions, cultural practices or attitudes of individual persons. It occurs independently from a technological system but it can flow into the system in different ways, e.g., as implicit prejudice of an individual developer or a common cultural practice, or also explicitly as intended form of discrimination. However, in most cases, it probably results from entrenched, nonreflective preconceptions. (2) Technical bias results from constraints due to the design or functionality of a system. This can include hard- and software issues like rulesets programmed into algorithms that unreasonably formalize social values or similar. (3) Emergent bias occurs when a system is in use that does not correspond with the peculiarities of application contexts; or when there is some kind of mismatch between what users or a usage context requires, and what the system features [6,16]. The context of deployment of an AI system, thus, can make a significant difference to the proper functioning of this system. Accordingly, the question is crucial if the AI or ML approach fits the particular context it is applied to. In [5], they discuss this issue as context bias that emerges when a system "is used in a new environment". They refer to the example of a healthcare algorithm developed for the context of a hospital to manage resource allocation. Even though the system may function appropriately for the hospital applying it, e.g., for a smaller rural clinic, the system may be flawed and produce inappropriate or misleading results. This is because the application context of the system does not correspond with the initial context the system was designed for. Amongst others, this example reminds us that it is crucial to bear in mind that the context in which an AI system, ML algorithm, etc., is applied has an effect on its functioning which consequently has an impact on the system and its sociotechnical environment as a whole.

Against this background, it is thus generally questionable whether technical fixes can effectively contribute to ease problems that are at their core sociotechnical issues with serious ethical and societal implications. From the perspective of technology assessment and science and technology studies (STS), a techno-centered understanding of the problem itself can even aggravate the problem because the functioning of technology does not always correspond with the societal contexts it is applied to. An essential precondition for accountability and transparency is that AI needs to function in a predictable and controllable way. A system, though, that is biased and automatically attempts to debias may become even more opaque and thus uncontrollable for humans. Because then neither the occurrence of bias nor the technical approach of debiasing might be comprehensible. This could lead to a complex dilemma or vicious circle where bias and debiasing approaches reinforce themselves. One of many differences between AI and human cognition is that humans can basically recognize and adapt to changed situations or contexts. AI systems usually do not have this capability. Technical approaches, such as reinforcing the adaptability of AI are not appropriate solutions to compensate this because it could even reinforce the problem: a system that frequently attempts to adapt to its environment might be even more unpredictable in behavior. Consequently, this would lead to an opaque, erratically functioning AI which would be even more of a threat than of use. To avoid such dilemmas, there is need for broader analytical perspectives that do not just focus on the technical issues of bias but bring together different views from multiple disciplines.

3. Employing AI Systems—A Set of Wicked Problems?

There are various difficulties behind the tempting (and misleading) assumption that every issue of AI, irrespective of whether it is technical or social, could be fixed by technical means. To realize that, it is useful to consider that AI can entail so-called wicked problems. Wicked problems are defined as a "class of social system problems which are ill-formulated, where the information is confusing, where there are many clients and decisionmakers with conflicting values, and where the ramifications in the whole system are thoroughly confusing" [19]. The idea of wicked problems originally stems from philosopher Karl Popper. Horst Rittel, a mathematician and designer, took up Poppers idea and, together with Marvin Webber, formulated ten properties of wicked problems [20]:

1. WPs have no definitive formulation.
2. WPs have no stopping rules.
3. Solutions to WPs cannot be true or false as they are context-dependent or value-laden.
4. There is no immediate and no ultimate test of a solution to a WP.
5. Solving a WP is a "one shot" operation with no room for trial and error.
6. There is no exhaustive list of admissible operations to solve WPs.

7. Every WP is essentially unique.
8. Every WP is a symptom of another "higher level" problem.
9. Every WP can be explained in various ways; the chosen way determines the nature of the problem's resolution.
10. Planners have no right to be wrong and are fully responsible for their actions.

From an engineering point of view, one may argue that some of these properties do not apply to AI systems. For instance, because AI systems have definitive formulations, clear rule sets and architecture, systemic boundaries, etc. Indeed, this may be valid if we look at AI systems from a sheer technical perspective. However, amongst other things, particularly AI systems imply a high degree of connectivity and interactivity with their environment. Therefore, it does not make sense to perceive them as closed systems, even though some systems might be closed in a narrow technical sense. In the original meaning, the framing of wicked problems basically refers to general issues of planning and policy-making. Their properties are thus not specifically applicable to the technical development of AI systems and might indeed be less relevant for ML developers. However, recalling the properties of the wicked problems might be supportive to come to a broader, problem-oriented perspective; especially for decision-makers and planners concerned with the use of AI systems in particular societal contexts. This could contribute to gain alternative views on the socio-technical complexity involved in the development and planning of AI systems. Akin to what Rittel and Webber [20] highlight for planning, for AI research and development, it is not just important to ask what AI systems are made of but also what AI systems do, as well as what should AI systems do? Neglecting these questions is akin to neglecting the fact that societal problems related to AI usage are different from the problems that AI architects and developers deal with. The crux is that AI research and development need more awareness and analytical perspectives that enable the grasp of technical as well as societal issues of AI systems. This includes accepting the fact that technical solutions are not always appropriate to fix societal problems but may require alternative approaches.

The core issue of a wicked problem is that there is a gap between determinism as a result of a given design or technology and indeterminism of the parts of reality which are affected by the use of the technology. This a general issue that can be found in every technology and which implies a conflicting relationship between determinacy and indeterminacy, or in other words: between the "artificial" and the "natural" [19]. AI makes this issue more severe as its use can have wider societal consequences: due to the high degree of automation inherent to AI, this conflict may intensify as AI has a transformative capacity where "natural" aspects of society are at risk of becoming reduced to machine-readable, datafied models that fit the logics of the artificial. AI in this regard bears the risk of an 'incremental "normalization" of society (. . .) for the sake of AI-based automation' [1]. This not only complicates transparency and accountability of AI-based applications but also reinforces the risks of undetected failure and self-fulfilling prophecies.

To deal with such risks, it is important to recall that automation is not an inevitable path we must take with AI; or as [11] points out (p. 5): "automation is not a straightforward perspective, but a choice." Consequently, "there is somebody who makes the choice and there should be reasons" justifying this choice. Otherwise, if there are no plausible reasons that justify automation through AI in particular contexts, automation should be questioned, critically assessed to comprehend its wider consequences and societal costs. More specifically, this means that we need better strategies to detect and avoid rule conflicts inherent to AI systems and conflicts between AI systems and their application contexts in real world settings. Based on the role of automation, the following sections discuss this aspect more thoroughly and propose potential ways to identify and handle eventual conflicts between system behavior and user practices.

4. Why AI-Based Automation Makes a Difference

Apparently, not every use of AI provides the same form of automation. In the literature, there are various concepts for categorizing different levels of automation. A common distinction in the domain of autonomous vehicles, for instance, involves five or six levels [21]. A classical approach is from [22] who provide a comprehensive list of ten general automation levels; further concepts can be found in [23], offering a detailed literature review on automation in manufacturing. For AI-based automation, there are varying approaches that orientate on the levels of driving automation. For example, Ref. [24] adapted these levels to systems based on natural language processing. All these concepts are useful to systematize automation. However, the main focus of these approaches is on categorizing the extent to which a system operates automatically. This is relevant but it is of limited use in the context of this paper, i.e., how to raise problem awareness and critical AI literacy. Because the existing approaches rather describe *what* is automated but not *why and where*, human intervention or scrutiny is important *in spite of* or especially *because of* automation. As a consequence, the basic functioning and impact of automation remain inexplicable and uncontrollable, especially for people without technical expertise. Here, particularly ML is an important issue as the degree of automation of an AI system is not least determined by the data model and the ML approaches embedded in the system. Different approaches can have different implications for the degree of automation and thus also for DAB-related risks. Therefore, it is important, also for nontechnical users, to understand the implications of automation, including the general ML approach (for an overview see, e.g., [25]) of the system they interact with. Typically, supervised learning has a lower degree of automation than unsupervised, reinforcement, or deep learning. The latter currently offers the highest level of automation as it can feature self-optimizing algorithms or similar. Reinforcement learning usually interacts dynamically with the environment, e.g., through sensors and feedback to gather data, deep learning typically requires more training data and has higher feature engineering capabilities. Both approaches are highly dynamic and thus may tend to reinforce more unpredictable behavior compared to, e.g., supervised learning which is based on predetermined criteria and relatively stable training data. The main aspect here is that not every ML approach might be appropriate for any given application context. This is clear for developers and technical experts but not for standard users or decision-makers interacting with AI systems. To avoid problems and DAB, it is thus also important to raise awareness of the peculiarities of the ML approach in use and its potential issues in particular contexts among decision-makers and persons interacting with AI systems. The following examples highlight the relevance of this aspect:

The first example concerns a classical form of automation: an autopilot system of an airplane, which usually is a rule-based system that does not need to have a comprehensive model incorporating the whole complexity of reality. This would even be counterproductive, but it obviously needs to have a plausible model of the landscape, the environment of the aircraft, etc., in order to provide reliable information to the pilot so that he or she can comprehend whether the autopilot and the plane function correctly and, in case of problems, can rapidly correct failure so that the airplane remains in control and safe. Every autopilot system usually bears certain risks of AB [26,27] which can be reduced by extensive training and technical design mechanisms to reduce overreliance and foster controllability of the system. However, this gets more complicated the more complex and less predictably a system behaves. This is what DAB addresses. A worst case in this example would be, if the autopilot is based on a predictive deep learning approach permanently attempting to optimize the flight route without any effective option for human pilots to intervene. Consequently, the human pilot would then lag behind the decisions of the autopilot leading to an escalating vicious circle where the autopilot decisions and the human intervention to correct them collide. In other words: there is a risk that AI-based automation overrules human autonomy. The severity of this risk depends not just on general explicability but also on plausibility, reliability, predictability and intervenability of the automated system.

Because even though it might be explicable to the human pilot how the autopilot behaves, it is worthless without intervention in case of problems.

Another example concerns the use of AI systems for job applications in human resource management (HR). There are various evident cases of discrimination of job-seekers by an AI system in this domain. For instance, Ref. [28] revealed a case where the algorithm of such a system used a model that incorporated the background image of a person to compute the qualification score of an applicant. Hence, the system inter alia assumes persons sitting in front of a bookshelf would be better qualified for a job than persons with, e.g., a transparent background. Similar is the case for black persons or persons with lower contrasting visual backgrounds. This can also involve racial discrimination which is evident in other AI-based hiring tools as well as a large variety of algorithmic systems (see, e.g., [9,29–34]). Obviously, the color of the skin, the ethnicity as well as the background visuals during a job interview are completely irrelevant for the qualification of an applicant. This a typical case of algorithmic bias that also highlights a likely risk of DAB resulting from the use of AI systems in social domains that used to be nonautomated. Here, DAB occurs, if the HR employee inviting a job-seeker for an interview uncritically accepts the preselection of the algorithm, e.g., because he is either unaware of the problem or does not care. Consequently, those applicants without a bookshelf in the background are discriminated as they have lower chances to get the job. Particularly tricky in such a case is, that the HR employee may not even be aware of the algorithmic bias and the discrimination, which is deeply enshrined in the system and thus, is very difficult to reveal.

Such problems cannot be fixed by technical means alone, such as improving FAT or debiasing. To point out why, we can imagine a typical approach of debiasing here: e.g., to adapt the algorithm so that it explicitly does not consider the background image at all. This would at least reduce bias resulting from background images, but it does not solve any other eventual issues resulting from other criteria the algorithm uses to preselect applicants. It would only be a one-shot solution to a symptom of a larger (wicked) problem. Beside the relatively simple aforementioned examples, there are several more complex issues like stereotypes and bias in large scale data models or ML frameworks that can significantly affect system behavior ([35,36]). Severe problems can occur with AI systems used in the health sector, where a variety of unintended consequences can lead to inappropriate decisions ([33,36,37]). In their study, Ref. [36] identified several problems with ML algorithms in clinical applications, particularly "overreliance on automation, algorithms based on biased data, and algorithms that do not provide information that is clinically meaningful". They thus conclude that automation can be useful but "overreliance on automation is not desirable". To reduce these problems, they argue that there is a need for measures to improve the understanding of the "intent behind the design" of AI systems "including choosing appropriate questions and settings for machine learning use, interpreting findings, and conducting follow-up studies" [36]. This could contribute to meaningful and ethically acceptable use of such systems with beneficial effects. Therefore, it is crucial to gain an analytical perspective on the problem that allows not just its technical but also societal dimensions to be considered. The following sections are dedicated to this aspect, elaborate on the main drivers of automation risks and then propose approaches that can facilitate the easing of the situation in real-world settings.

5. Influencing Factors of Deep Automation Bias

There are several guidelines and approaches that highlight the various issues of AI concerning ethics, human rights, bias, responsibility and trust, etc. (e.g., [38–42]). However, there is a general lack in practicability among those approaches as their focus is either on highlighting basic ethical requirements like avoiding harm and stimulating benefits for society, justice, and autonomy; or on technical challenges such as data quality or attempts to operationalize ethical principles in technology. Apparently, many of these approaches are relevant, but we also need more specific, problem-oriented approaches that provide useful for application contexts of AI without getting lost in ethical debates beyond practicability or misleading technocratic approaches to societal problems.

In order to circumvent this, a systemic perspective on AI that incorporates a major implication of its use in societal contexts could be fruitful to sharpen the analytical lens on the variety of sociotechnical issues. More precisely, this suggests more emphasis on automation which is a common denominator of AI systems irrespective of their particular design. AI basically implies a higher degree of automation than any other technology. Or in other words: AI-based technology represents a sociotechnical system that enables and fosters automation at different levels. Consequently, a problem-oriented assessment of the sociotechnical impact of an AI system involves gaining knowledge on the basic functionality including the degrees of automation of the system. The crux is that this is not easy to determine and highly context-dependent. Therefore, it makes sense to focus on how automation occurs in an AI system and how it affects its application context. Focusing on automation as an analytical frame puts greater emphasis on the procedural and structural changes resulting from AI usage, and, with appropriate concepts, also on the practical effects and societal implications of these changes in particular application contexts. Put simply, a key aspect to improving the sociotechnical understanding of AI-related problems is to think of these questions together: how does automation occur in an AI system and how does this lead to the generation of a particular output? What are the consequences of this output for human users and, in a wider view, for its societal context?

A general risk of automation is automation bias (AB), which describes the tendency to uncritically accept the computer-generated outcome (from a technical system). A basic reason for AB and the corresponding lack of critical reflection or scrutiny of the results from the lack of options to scrutinize how a system generated a certain outcome. This refers to, e.g., known issues like lacking transparency, accountability, explicability, verifiability, interpretability, etc. AI technology provides a higher form of automation than other technology which implies that AI can reinforce the risks of AB. In [1], I thus introduced deep automation bias (DAB) as a core problem of AI usage to highlight that AI technology significantly reinforces the risks of AB due to ML featuring adaptivity and unsupervised or self-learning capabilities (e.g., deep learning). DAB addresses the problem of increasing levels of complexity and opacity of AI that reinforce AB due to its dynamic, unpredictable and thus potentially uncontrollable behavior [1]. It includes the self-reinforcing problem of wrong expectations of AI functionality due to the human propensity to blindly trust in technology on the one hand, and the limits of technology due to its necessarily abstract/reductionistic model of societal reality on the other. This can lead to a vicious circle as unpredictable behavior of a system benefits unreliable decisions based on this system which then can reinforce pressure to (re-)act to adapt to system behavior. The result can be a nonoptimal acceptance of insufficient system functionality or failure due to a lack of other/better choices. As a wider consequence, society then becomes even more dependent on AI which increasingly produces an impact on society and individuals that hampers alternatives. DAB is thus a metarisk of AI systems and their intrinsic dynamics. A pressing question in this regard is: how to scrutinize a system and its societal impact that dynamically changes its algorithmic rule sets and consequently, its functionality?

Essentially, this requires heuristic approaches that enable the gaining of the aforementioned systemic perspective. In this regard, it is crucial to understand how DAB can emerge and what main factors influence the severity of DAB-related risks. A basic component is uncertainty, which can have different but interrelated reasons. According to studies on AB, humans tend to over-rely on automated systems due to: inexperience or lack of skills or confidence, lack of options to scrutinize an automated process (input, output, or action). This can be further complicated due to little practical experience or limited resources, e.g., due to pressure to act, lack of time and high workload [27,43–45]. All these issues are likely to aggravate in the case of DAB. In particular, as the higher complexity and opacity of AI systems reinforces the problem of lacking options to scrutinize the functionality and automated processes of the system. This can lead to further issues, i.e., a lack of agency, e.g., because of a generally lacking problem awareness and/or a lack of options to intervene in an automated decision-making process.

These factors are not to be misinterpreted as human deficits which could be fixed by technical means only like, e.g., fostering usability. Besides these human factors there are also technical factors that affect the severity of DAB-related risks. Figure 1 illustrates the interplay of these different factors from both the technical (blue circled area) and the human (green circled area) perspective:

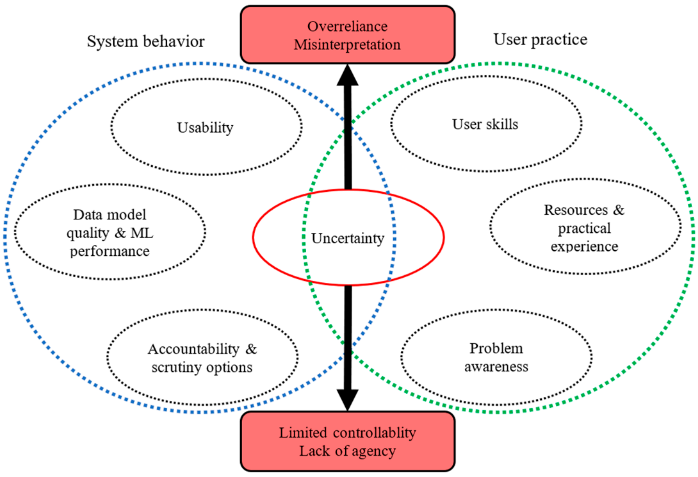

Figure 1. Main influencing factors of deep automation bias.

The two circles represent the interplay between system behavior and user practice. Each circle contains different but interrelated factors that affect behavior and uncertainty on each side. Every factor also relates to its counterpart in the opposed circle on the same level (e.g., usability is linked to user skills, etc.).

As shown, there are at least two dimensions that determine the risk of DAB: at the top is the “classical” AB risk of overreliance on the behavior of an AI system which can involve 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 sociotechnical factors. The main connecting factor is uncertainty which is shaped by technical as well as social issues. Both can reinforce mutually, as the intersection highlights. From a technical perspective, system complexity, opacity or unpredictable system behavior strongly depend on the quality of data models and ML performance, as well as from usability features. The interplay of these factors basically affects system behavior and the degree of uncertainty. This can have a negative effect on the transparency and accountability of the system and the options to verify and scrutinize system behavior. From the social perspective, the aforementioned factors of AB, such as user skills, resources or practical experience, affect uncertainty and particularly the problem awareness of the human agent. Resources here typically mean the time and knowledge to comprehend, interpret or verify the system behavior, such as the generated output or action triggered based on the system performance, etc. The interplay of all these factors influences the ability to comprehend and interpret the functioning of the system. This typically involves input, output, or action, and if there is a need for scrutiny and evaluation of system behavior. Hence, the risk of DAB is potentially higher when there is either a lack of accountability and options to scrutinize system behavior and/or low problem awareness due to lack of options to comprehend how and what foundation the system generates an output or how the system triggers particular actions, etc. These issues can mutually reinforce and increase uncertainty (e.g., low options for scrutiny benefit lacking agency, low problem awareness hampers the ability to scrutinize and thus agency). The interplay of all these factors affects the severity of DAB-related risks.

The core aspect the model highlights is that the higher the mismatch between both perspectives (system behavior and user practices), the higher is the risk of DAB. Hence, DAB ultimately bears the risks of undetected mismatches or conflicts between the AI model of reality and the human users' model of reality in a particular application context. DAB-related risks are basically higher, when it is hampered or impossible to critically assess whether the functioning and behavior of an AI system is transparent, explicable, plausible, reliable or legitimate. The degree of automation plays a particularly important role in this regard. To assess the risks of an AI system thus also requires consideration of the extent to which automated procedures are involved. This supports the raising of problem awareness and critical AI literacy in order to deal with DAB-related risks from a sociotechnical perspective.

6. A Draft Heuristic Model to Assess DAB

In Section 4, I presented the main factors that affect DAB. This section now outlines some aspects of system behavior that can affect DAB. A basic determinant for DAB-related risks is the occurrence of contradictions or rule conflicts between factual system behavior and the expectations of the human user of how the system behaves. This distinction is important because the increasing use of AI systems in societal contexts also entails an incremental reduction of complex social issues for the sake of algorithmic representation, measurement and operation. As any computer model, also data models and algorithmic rule sets of AI are necessarily abstract and reductionist as they cannot completely represent the whole complexity of a real-world context. Humans have better skills in this regard. Consequently, AI models are less complex than social reality. This is unproblematic as long as system behavior is predictable, explicable and interpretable for human entities and options for intervention are available if necessary. Problems occur in the case of a mismatch that remains undetected or uncorrected, leading to problematic decisions or actions (as also discussed in Section 4). The following model in Figure 2 illustrates an approach to raise awareness on these issues and to reveal eventual rule conflicts or mismatches:

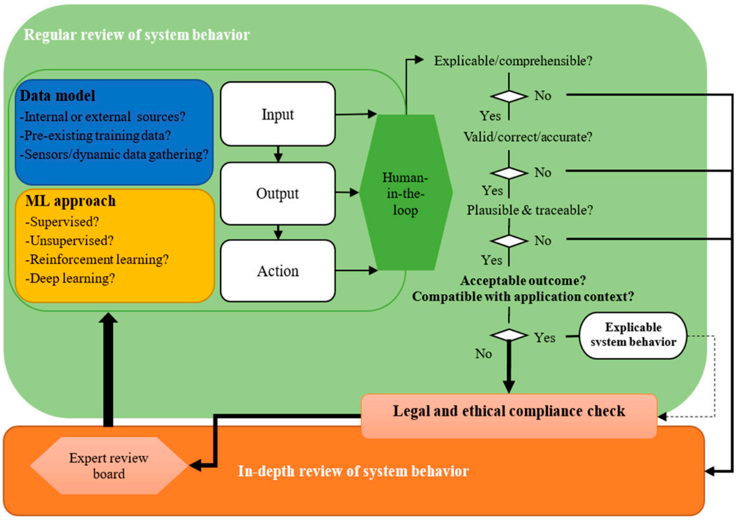
Regular review of system behavior
Data model
-Internal or external sources?
-Pre-existing training data?
-Sensors/dynamic data gathering?
ML approach
-Supervised?
-Unsupervised?
-Reinforcement learning?
-Deep learning?
Input
Output
Action
Human-in-the-loop
Explicable/comprehensible?
No
Yes
Valid/correct/accurate?
No
Yes
Plausible & traceable?
No
Yes
Acceptable outcome?
Compatible with application context?
Yes
No
Explicable system behavior
Legal and ethical compliance check
Expert review board
In-depth review of system behavior

Figure 2. Draft heuristic model to assess deep automation bias.

The simplified core of the AI system is represented by the data model and the ML approach which build a functional unit that determines input, output and action of the system. To some extent, the specific features of this unit predetermine the degree of automation of the system and thus the general susceptibility to DAB. Therefore, it makes a difference, e.g., if training data stems from external or internal sources, how accurate and reliable the data is, and if data is dynamically gathered through sensors during system operation, etc. The same applies to the ML approach which also affects the degree of automation and system behavior. As mentioned before, DAB basically occurs in the case of mismatch between factual system behavior and human interpretation of this behavior. This may concern the input, output or action of the system for different reasons. For instance, the input might be incorrect entailing further issues; or correct but not comprehensible; or the output is invalid but the error remains uncorrected; or the system operates formally correctly but produces outcome or leads to actions that are opaque or incompatible with the application context of the system. This can involve a violation of legal obligations, human rights (such as privacy, security, autonomy, dignity), reinforcement of prejudice and discrimination or other ethical issues.

To consider that the occurrence of DAB can have different reasons, the model involves different levels of human scrutiny and oversight. A precondition for this is a human-in-the-loop (HITL) concept [46] as an integral part of the AI system. HITL here is not limited to technical issues like supporting ML feature learning or the like. Depending on the specifics of an AI-system, this can be a typical part of the review as well. However, basically, HITL here has a broader scope, i.e., to facilitate human review loops during system operation in order to assess DAB-related risks at different sociotechnical levels. Therefore, it is modeled at the threshold between the AI system and the review queries that concern the application context. Consequently, humans that review the system are not necessarily system developers but persons who comprehend the functioning of the system and its wider sociotechnical effects in a particular application context. Ideally, more than one person is involved in the review. How this is operationalized in practice depends on the features of an AI-system in particular. However, the model highlights the necessity to foresee review processes that consider the whole functioning and performance of the system (including, input, output and action).

There are four basic levels which are not necessarily iterative but apply to input, output as well as outcome/action of the system: (1) Explicability concerns the assessment of how explicable and comprehensible the system behavior is. (2) Validity involves the question of whether the system operates correctly, uses or produces accurate information that is formally valid for the processes it relates to. (3) Plausibility comprises a check, if the system behavior is not just formally correct in a technical sense but also plausible, traceable and interpretable to human users without uncertainties or doubts concerning generation and reproducibility of the results. For each case applies: the next check can be performed if there is no significant doubt or uncertainty concerning the issue remaining. Otherwise, if something in the system operation is questionable or uncertain, then a more detailed review of the system behavior is necessary with different experts involved. The final level, (4) acceptability/compatibility concerns the question of whether the behavior and overall outcome of the system is acceptable and compatible with the context to which it is applied, also in a legal and ethical sense. In the case of uncertainties or open issues regarding legal or ethical compliance, the system needs additional compliance checking. Until this review is proceeding and not finished, the AI system should be taken out of operation to avoid legal or ethical issues or any further risks and harmful events.

If all queries were positively answered, the behavior of the system seems to be basically reasonable and explicable in a way that is appropriate and acceptable for the application context. This means that no opaque or incomprehensible modes of system behavior were detected during a regular review. Consequently, there was no DAB risk observable. As the term suggests, this regular review is meant to be on a regular basis, and in any case of irregularities or uncertainties, a further, more detailed in-depth review process (the orange block at the bottom) is triggered. To ensure that the system behavior remains stable with low DAB-risks, it is advisable to conduct frequent in-depth review processes also if no irregularities were detected (as represented by the dotted line on the right). This apparently includes legal and ethical compliance checking which requires additional expertise. Therefore, this block is modeled at the threshold between regular and in-depth review. In any case, in-depth review ideally involves a specific review board of different experts with technical as well as organizational, legal and ethical skills or a mix of system developers, planners, decision-makers and practitioners. In-depth review means that the system as a whole (including its design and behavior and outcome in different contexts) is inspected and analyzed by the expert board.

7. Summary and Conclusions

A main argument of this paper is that technical approaches to improve FAT or debiasing are not sufficient to address the sociotechnical and ethical risks related to the use of AI systems. Particularly, because these risks may involve several wicked problems that may require alternative approaches beyond technical solutions. Certainly, bias in ML is a critical issue of AI systems with various reasons and effects. Bias can result from pre-existing prejudice during technical development, technical issues such as poor data quality, insufficient modeling, or inappropriate operation of ML-algorithms, etc. Ultimately, though, bias can also result from rule conflicts between AI design and AI application contexts due to complexity gaps between (algorithmic/statistical) assumptions in the AI system and user practices. Hence, although different forms of bias can occur in an AI system, automation is their common ground. Therefore, the paper focusses on deep automation bias (DAB) as a metarisk of AI. There is indeed a lot of relevant research in the field of FAT and debiasing with several important approaches in this regard. Furthermore, it is important to consider ethical implications of a technology at an early stage of development. However, incorporating ethical values into technology design as part of automated procedures as such can be counterproductive and could worsen the situation by making existing societal risks of AI even more severe. Related to that, there is also a risk of "agency-laundering" meaning that instead of taking responsibility for unethical actions and measures to correct them, ethical and societal problems remain unsolved [5]. Accordingly, techno-fixes for FAT and the like might be misused to avoid taking responsibility by reframing the problems to sheer technical issues. To avoid this, we need to be aware of the fact that the problems that bias in ML can induce, like unfair/discriminating, false, misleading, unethical or socially inacceptable outcomes, are not sheer technical, but sociotechnical issues. Otherwise, there is no chance to understand the societal implications of AI systems, its risks and thus also no option to effectively avoid different forms of bias, discrimination, system malfunction or unethical consequences.

The conceptualization of DAB in this paper is an attempt of an alternative analytical frame to improve practical problem awareness in this regard. As argued, automation plays a particular role in AI with different degrees and implications for the application context of AI systems. A stronger focus on this role is thus important to tackle the variety of sociotechnical risks and broaden the analytical view. To stimulate this focus with an analytical lens, the paper presented and discussed DAB as a metarisk inherent to AI systems. As shown in Section 5, DAB is a multifaceted problem that is shaped by a number of driving factors comprising technical as well as social issues that together affect uncertainty on how to comprehend and scrutinize the functionality and behavior of an AI system. The main issue of DAB concerns conflicts or undetected mismatches between the behavior of an AI system and human user practices. Risks of DAB are basically given in cases, where it is hampered or impossible to critically assess whether the functioning and behavior of an AI system is transparent, explicable, plausible, reliable or legitimate. Particularly the degree of automation plays an important role in this regard. To highlight this issue, the proposed heuristic model in Section 6 provides an approach to assess DAB-related risks of AI systems. The basic idea here is a to arrive at approaches that ease regular review to detect DAB-related risks due to conflicts or mismatches between system behavior and user practices. In case of problems or uncertainties, this could be linked to in-depth reviews where different experts can then analyze the system behavior in detail. This heuristic model can be also of practical use in real world settings to support users or non-technical experts in detecting DAB-related risks of AI systems. Depending on the identified issue, further, more detailed review phases might then be triggered. In this regard, the model also aims at supporting "human-in-the-loop" concepts which are basically essential. However, to be effective, they require more awareness on the role of automation in AI systems, including its effects, limits and societal implications. This paper is a contribution in this respect. Further research is necessary which may also involve a stronger focus on the variable degrees of automation in AI systems and how to foster explicability and scrutiny on a general level. This is not least relevant in order to come to harmonized approaches that combine technical solutions for FAT and debiasing with advanced concepts of training and education for human actors interacting with AI or being confronted with the use of AI systems in particular application contexts. This could be supportive to avoid AI shaping society as blackbox-systems beyond human control and come to a responsible implementation and use of AI in accordance with ethical principles, societal values and human wellbeing.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: I want to thank my colleagues for their support, the anonymous reviewers for their valuable comments that helped to improve the quality of the paper and the editorial team for the efficient publication process.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Strauß, S. From Big Data to Deep Learning: A Leap towards Strong AI or 'Intelligentia Obscura'? *Big Data Cogn. Comput. (BDCC)* **2018**, *2*, 16. [CrossRef]
2. Edwards, L.; Veale, M. Slave to the Algorithm? Why a 'Right to Explanation' Is Probably Not the Remedy You Are Looking For. *Duke Law Technol. Rev.* **2017**, *16*, 18–84. [CrossRef]
3. Selbst, A.D.; Boyd, D.; Friedler, S.A.; Venkatasubramanian, S.; Vertesi, J. Fairness and Abstraction in Sociotechnical Systems. In Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAT*), Atlanta, GA, USA, 29–31 January 2019; pp. 59–68. Available online: https://ssrn.com/abstract=3265913 (accessed on 12 March 2021).
4. Wong, P.-H. Democratizing Algorithmic Fairness. *Philos. Technol.* **2019**, *33*, 225–244. [CrossRef]
5. Tsamados, A.; Aggarwal, N.; Cowls, J.; Morley, J.; Roberts, H.; Taddeo, M.; Floridi, L. The ethics of algorithms: Key problems and solutions. *AI Soc.* **2021**. [CrossRef]
6. Friedman, B.; Nissenbaum, N. Bias in Computer Systems. *ACM Trans. Inf. Syst.* **1996**, *14*, 330–347. [CrossRef]
7. Pasquale, F. *The Blackbox Society—The Secret Algorithms that Control Money and Information*; Harvard University Press: Cambridge, UK, 2015.
8. Annany, M.; Crawford, K. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media Soc.* **2016**. [CrossRef]
9. O'Neil, C. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*; Crown: New York, NY, USA, 2016.
10. Caplan, R.; Donovan, J.; Hanson, L.; Matthews, J. Algorithmic Accountability: A Primer. Research Report, Data & Society. 2018. Available online: https://datasociety.net/wp-content/uploads/2018/04/Data_Society_Algorithmic_Accountability_Primer_FINAL-4.pdf (accessed on 7 March 2021).
11. Tsoukiàs, A. Social Responsibility of Algorithms: An Overview. 2020. Available online: https://arxiv.org/pdf/2012.03319.pdf (accessed on 12 March 2021).
12. Wieringa, M. What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20), Barcelona, Spain, 27–30 January 2020; pp. 1–18. [CrossRef]
13. Holstein, K.; Vaughan, J.; Daumé, H., III; Dudík, M.; Wallach, H. Improving fairness in machine learning: What do industry practitioners need? In Proceedings of the CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), Scotland, UK, 4–9 May 2019. [CrossRef]
14. Arduini, M.; Noci, L.; Pirovano, F.; Zhang, C.; Raj, Y.; Paudel, B. Adversarial Learning for Debiasing Knowledge Graph Embeddings. *arXiv* **2020**, arXiv:2006.16309. Available online: https://arxiv.org/abs/2006.16309 (accessed on 12 March 2021).
15. Eid, F.-E.; Elmarakeby, H.A.; Chan, Y.A.; Fornelos, N.; ElHefnawi, M.; Van Allen, E.M.; Heath, L.S.; Lage, K. Systematic auditing is essential to debiasing machine learning in biology. *Commun. Biol.* **2021**, *4*, 183. [CrossRef] [PubMed]
16. Simon, J.; Wong, P.-H.; Rieder, G. Algorithmic bias and the Value Sensitive Design approach. *Internet Policy Rev.* **2020**, *9*, 4. [CrossRef]
17. Turilli, M.; Floridi, L. The Ethics of Information Transparency. *Ethics Inf. Technol.* **2009**, *11*, 105–112. [CrossRef]

18. Strauß, S. *Privacy and Identity in a Networked Society: Refining Privacy Impact Assessment*; Routledge: Abingdon, UK; New York, NY, USA, 2019.
19. Buchanan, R. Wicked Problems in Design Thinking. *Des. Issues* **1992**, *8*, 5–21. [CrossRef]
20. Rittel, H.; Webber, M. Dilemmas in a General Theory of Planning. *Policy Sci.* **1973**, *4*, 155–169. [CrossRef]
21. SAE J3016 Levels of Driving Automation. Available online: https://www.sae.org/news/2019/01/sae-updates-j3016-automated-driving-graphic (accessed on 12 March 2021).
22. Sheridan, T.B.; Verplank, W.L. *Human and Computer Control of Undersea Teleoperators*; Technical Report; Massachusetts Institute of Technology, Man-Machine Systems Laboratory: Cambridge, MA, USA, 1978.
23. Frohm, J.; Lindström, V.; Winroth, M.; Stahre, J. Levels of Automation in Manufacturing. *Ergon. Int. J. Ergon. Hum. Factors* **2008**, *30*, 181–207. Available online: https://core.ac.uk/download/pdf/70575908.pdf (accessed on 12 March 2021).
24. Edwards, J.; Perrone, A.; Doyle, P.R. Transparency in Language Generation: Levels of Automation. In Proceedings of the 2nd Conference on Conversational User Interfaces CUI'20, Bilbao, Spain, 9–10 July 2020. [CrossRef]
25. Fumo, D. Types of Machine Learning Algorithms You Should Know. 15 June 2017, Towards Data Science. Available online: https://towardsdatascience.com/types-of-machine-learning-algorithms-you-should-know-953a08248861 (accessed on 12 March 2021).
26. Parasuraman, R.; Manzey, D. Complacency and Bias in Human Use of Automation: An Attentional Integration. *J. Hum. Factors Ergon. Soc.* **2010**, *52*, 381–410. [CrossRef]
27. Goddard, K.; Roudsari, A.; Wyatt, J.C. Automation bias: Empirical results assessing influencing factors. *Int. J. Med Inform.* **2014**, *83*, 368–375. [CrossRef]
28. Harlan, E.; Schnuck, O. Objective of Biased? On the Questionable Use of Artificial Intelligence for Job Applications, BR Online. 2021. Available online: https://web.br.de/interaktiv/ki-bewerbung/en/ (accessed on 12 March 2021).
29. Osoba, O.; Welser, I.V.W. *An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence*; RAND Corporation, 2017. Available online: https://www.rand.org/content/dam/rand/pubs/research_reports/RR1700/RR1744/RAND_RR1744.pdf (accessed on 12 March 2021).
30. *Noble SU, Algorithms of Oppression: How search Engines Reinforce Racism*; NYU Press: New York, NY, USA, 2018.
31. Borgesius. *Discrimination, Artificial Intelligence, and Algorithmic Decision-Making; Study for the Council of Europe: 2018*. Available online: https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73 (accessed on 12 March 2021).
32. Kodiyan, A.A. An Overview of Ethical Issues in Using AI Systems in Hiring with a Case Study of Amazon's AI Based Hiring Tool. Researchgate Preprint 2019. Available online: https://www.researchgate.net/publication/337331539_An_overview_of_ethical_issues_in_using_AI_systems_in_hiring_with_a_case_study_of_Amazon%27s_AI_based_hiring_tool (accessed on 12 March 2021).
33. Obermeyer, Z.; Powers, B.; Vogeli, C.; Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **2019**, *336*, 447–453. [CrossRef]
34. Köchling, A.; Wehner, M.C. Discriminated by an algorithm: A systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. *Bus. Res.* **2020**, *13*, 795–848. [CrossRef]
35. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. *arXiv* **2020**, arXiv:2005.14165. Available online: https://arxiv.org/abs/2005.14165 (accessed on 12 March 2021).
36. Gianfrancesco, M.A.; Tamang, S.; Yazdany, J.; Schmajuk, G. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. *JAMA Intern Med.* **2018**, *178*, 1544–1547. [CrossRef]
37. Cabitza, F.; Rasoini, R.; Gensini, G.F. Unintended consequences of machine learning in medicine. *JAMA* **2017**, *318*, 517–518. [CrossRef]
38. HLEG—High-Level Expert Group on Artificial Intelligence (2019), Ethics Guidelines for Trustworthy AI. European Commission. Available online: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (accessed on 12 March 2021).
39. AlgorithmWatch. *Automating Society: Taking Stock of Automated Decision-Making in the EU*, 1st ed.; AlgorithmWatch in Cooperation with Bertelsmann Stiftung; 2019. Available online: https://algorithmwatch.org/en/automating-society-2019/ (accessed on 12 March 2021).
40. Council of Europe—CoE. Unboxing Artificial Intelligence: 10 Steps to Protect Human Rights. Council of Europe, Commissioner for Human Rights. Available online: https://rm.coe.int/unboxing-artificial-intelligence-10-steps-to-protect-human-rights-reco/1680946e64 (accessed on 12 March 2021).
41. Bertelsmann. From Principles to Practice—An Interdisciplinary Framework to Operationalise AI Ethics. AI Ethics Impact Group, Bertelsmann Stiftung. Available online: https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/WKIO_2020_final.pdf (accessed on 12 March 2021).
42. Floridi, L.; Cowls, J.; Beltrametti, M.; Chatila, R.; Chazerand, P.; Dignum, V.; Luetge, C.; Madelin, R.; Pagallo, U.; Rossi, F.; et al. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds Mach.* **2018**, *28*, 689–707. [CrossRef]

43. Alberdi, E.; Strigini, L.; Povyakalo, A.A.; Ayton, P. Why Are People's Decisions Sometimes Worse with Computer Support? In *Computer Safety, Reliability, and Security SAFECOMP 2009*; Lecture Notes in Computer Science. 5775; Springer: Berlin/Heidelberg, Germany, 2009; pp. 18–31. [CrossRef]
44. Goddard, K.; Roudsari, A.; Wyatt, J.C. Automation bias: A systematic review of frequency, effect mediators, and mitigators. *J. Am. Med. Inform. Assoc.* **2012**, *19*, 121–127. [CrossRef] [PubMed]
45. Lyell, D.; Coiera, E. Automation bias and verification complexity: A systematic review. *J. Am. Med. Inform. Assoc.* **2016**, *24*, 424–431. [CrossRef]
46. Emmanouilidis, C.; Pistofidis, P.; Bertoncelj, L.; Katsouros, V.; Fournaris, A.; Koulamas, C.; Ruiz-Carcel, C. Enabling the human in the loop: Linked data and knowledge in industrial cyber-physical systems. *Annu. Rev. Control* **2019**, *47*, 249–265. [CrossRef]