Addressing Automation Bias through Verifiability

Lukas J. Hondrich[1,2], Hannah Ruschemeier[1]

[1]FernUniversität Hagen, Hagen, Germany
[2]Charité Universitätsmedizin Berlin, Berlin, Germany

Abstract

The phenomenon of human bias finds a new facet in the hybrid human-machine interaction of today's digitized decision-making systems: automation bias describes the fact that human decision-makers overly trust machine-generated decision proposals, sometimes against their better knowledge. Although human involvement in hybrid human-machine systems is the practical rule compared to fully automated systems in institutionalized decision-making processes, it is not clear how this can be operationalized in a safe, adequate and legally compliant way. In its current legislated form, human interaction does not ensure meaningful human involvement, it also represents a systemic avenue to shirk responsibility by decision support system manufacturers and deployers. In this paper, we analyze the literature on human performance in automated systems and automation bias and identify verification behavior as the key variable ameliorating automation bias. Based on the empirical evidence for automation bias and its cognitive-behavioral correlates, we propose verifiability as a minimum necessary requirement for meaningful human involvement. We argue that verifiability might be subdivided into 1) the intrinsic verification complexity of a system, 2) factors relating to the verification propensity of a user, and 3) the contextual factors influencing verification.

Keywords

automation bias, human centered computing, algorithmic regulation, human-in-the-loop, verification

1. Automation Bias, Safety and Fairness

The digital transformation has led to the ubiquity of algorithmically influenced decisions. Government and private actors are replacing formerly purely human decision-making processes entirely with computer systems or – in the plural – are pre-structuring human decision-making with automated decision proposals. Automated decision-making systems (ADMS) make existentially significant decisions about humans, such as the distribution of child benefits [1], the allocation of support measures in the labor market [2], or creditworthiness [3]. This is intended to make decision-making processes more effective, efficient, rational or neutral.

However, numerous examples of algorithmic errors, due to a deficient data basis, programming errors or incorrect application have shown that the digitization of decision-making processes is not efficient, desirable or compatible with the principles of the rule of law and the protection of fundamental rights in all areas. When it comes to the regulation of ADMS, current law draws a clear distinction between fully autonomous systems, practically prohibiting them by Article 22

EWAF'23: European Workshop on Algorithmic Fairness, June 07–09, 2023, Winterthur, Switzerland
lukas.hondrich@fernuni-hagen.de (L. J. Hondrich); hannah.ruschemeier@fernuni-hagen.de (H. Ruschemeier)
www.hondrich.github.io/ (L. J. Hondrich);
www.fernuni-hagen.de/prof-ruschemeier/team/hannah.ruschemeier.shtml (H. Ruschemeier)

GDPR, and decision support systems that are significantly less regulated. However, decades in psychological and human factors research show that humans perform comparably weak in passive monitoring tasks [4] and have difficulties calibrating reliance on decision support systems [5, 6]. This leads to a rift between the assumed human performance using decision support systems and the performance they exhibit in reality, contributing to the aforementioned consequential failures. Systematically overestimating humans and under-regulating human-machine systems leads to a situation in which humans become the legal and "moral crumble zone" [7]. This does injustice to the persons in question that have to carry responsibility beyond their capabilities and leads to unsafe systems. Unsafe systems, in turn, disproportionally cause harm to already marginalized groups, perpetuating and reinforcing systemic discrimination [8].

Therefore, to build safe, fair and legally compliant systems, we propose to better define and help operationalize meaningful and substantial human involvement [9]. Furthermore, we argue that for understanding what meaningful and substantial in the context of hybrid human-machine systems means, it is necessary to consult psychological, technical and human factors research. Based on this research, we propose the concept of verifiability as a minimal necessary requirement for meaningful and substantial human involvement in hybrid human-machine systems.

2. Automation Bias and Verification

Automation Bias has been defined "as a heuristic replacement for vigilant information seeking and processing" [10], resulting in the overreliance on decision support system (DSS) output. Research in cognitively demanding and multi-task settings pointed towards limited cognitive resources as important factors, i.e. a) withdrawal of attention in terms of incomplete cross-checking of information, (b) active discounting of contradictory system information, and (c) inattentive processing of contradictory information analog to a "looking-but-not-seeing" effect [11]. Other research, in single-task as well as social settings, still reported automation bias, pointing towards additional social and motivational aspects, such as the image of decision support systems as powerful agents with superior analytic capability [12], causing "diffusion of responsibility" [13]. Such social causes might add to or interact with cognitive limitations. It is likely that both cognitive and social-psychological mechanisms play a role in determining the user's cognitive strategy, that could be either be geared towards impression management and self-justification or an adaptive cognitive strategy, focused on self-critical thinking and taking a broader range of alternatives and options into account [14].

Importantly, this adaptive cognitive strategy was in turn associated with more "verification behavior" [14] and naturally to less automation bias. Verification behavior is a broad term referring to checking for the correctness of DSS output. In the context of technical properties of DSS a related term, verification complexity, and implicitly verification behavior, have been defined as the number of necessary steps (acquire, transform, interpret, or use steps) to check for correctness of DSS output [15].

Performing these steps requires cognitive resources - in line with this, Lyell and Coiera [15] report that interventions aimed at reducing cognitive load were most effective in reducing automation bias. Such interventions were aimed at the DSS itself, e.g. better user interfaces that

integrated well with user's prior training as well as the DSS environment, for instance reducing distractions.

In the context of machine and deep learning, the subfield of explainability gained popularity. While explainability and verifiability are not synonymous, they ultimately share the same goal - making it easier for humans to check for correctness of processing and output. Not only is this verifiability a prerequisite for fairness, but in certain areas of decision-making, such as public administration, justification is required by the rule of law. Such justifications in turn require knowledge and understanding of the reasons for the decision.

To summarize, the importance of verifiability of machine output for human involvement is well reflected in the technical and human factors research. We argue that verifiability - if expanded to encompass all aspects that drive verification behavior (technical, social, contextual) - can represent a valuable design principle for decision support systems and should be taken into account by attempts to formulate legislation around decision support systems.

3. Verifiability as a Minimal Necessary Requirement for Meaningful and Substantial Human Involvement

Empirical research on automation bias and generally human performance in automated systems, showed a range of different factors that drive the severity of automation bias ([12, 6, 15]). Parasuraman et al. [4, 12] subdivided these factors into properties of the system (degree of automatization, reliability, etc.), properties of the situational context (degree of personal responsibility, cognitive load, etc), properties of personality (self-efficacy, attitude towards technology, etc.) and the psychological state of the user (tiredness, motivation, etc.). Goddard et al. [6], similarly, subdivided important factors into DSS factors, environment factors and user factors. Accordingly we propose to structure aspects that determine verifiability into DSS factors, i.e the intrinsic verification complexity [15] of a system, user factors, i.e. verification propensity of a user [10], e.g. psychological traits (e.g., confidence of in their ability to scrutinize decisions[16]) and states (e.g., tiredness), as well as context factors, e.g. time constraints or cognitive load [15].

3.1. Non-compensatory Nature of Verifiability Factors

A sufficiently verifiable hybrid human-machine system in our view necessitates the DSS to be verifiable *and* the user having the capacity *and* the context being favorable to verification. In this scenario the "the weakest link", i.e. the lowest verifiability factor, would limit the overall verifiability. This is suggested by previously mentioned studies in which manipulation of individual DSS, user and context factors was associated with a significant effect on automation bias. Additionally, recent critiques of explainability may be interpreted in going into a similar direction; i.e. that explainability of a system in itself may not be sufficient and not even be properly definable without taking the specific user and context into consideration. For instance [17] raised the point that explanations without allowing for integration with a user's expertise may be counterproductive and risky as seen in the case of the COMPAS-system, where judges did not have the information that severity of crime was not included in the 130+ input variables of the

system, and thus likely excluding this critical piece of information in their risk assessment. This seems particularly important when hybrid decision-making procedures are used in traditionally grown and analogue institutions, such as the judicial system. Accordingly, newer frameworks formulating standards for explanations take user and context factors into consideration (compare to *usability requirements*, *operational requirements* and *validation requirements* proposed for explainability factsheets [18]).

3.2. Interaction Effects between DSS, User and Contextual Factors

Another important characteristic are possible interaction effects between DSS, user and context factors that may lead to unexpected overall verifiability and thus occurance of automation bias. Various studies, e.g. [19, 20], reported that explanations may lead to worsening of automation bias due to increased, unwarranted trust in the DSS, pointing at interactions between DSS- and user factors. Other studies, e.g. [10], found that social accountability may lead to differing cognitive strategies, i.e. impression management or verification behavior, putatively depending whether users felt responsible for the outcome or the process of the decisions, pointing at an interaction effect of context factors (social accountability) with user factors (construction of accountability, self-efficacy and self-trust). Another meta-analysis ([21]) found that advantageousness of the locus of accountability depended on the task difficulty, with easier tasks benefiting from accountability on the process, while more difficult ones benefiting from outcome accountability - pointing at an additional interaction with the factor of verification complexity of the DSS.

4. Conclusion

We argue for a stronger focus on the risks of hybrid decision-making systems and for the translation of psychological findings into normative guidelines. The current differentiation in EU law between fully automated and partially automated decision-making systems is not convincing in view of the comparable risks for the protected legal interests (legal protection of the individual, prohibition of discrimination, equal opportunities, legal obligation of the administration). With concrete implementation, procedural requirements such as different levels of review, ex ante assessments of systems or rotation requirements are just as conceivable as rights of data subjects, facilitation of evidence or substantive requirements along the lines of Article 22 of the GDPR.

References

[1] M. van Bekkum, F. Z. Borgesius, Digital welfare fraud detection and the dutch syri judgment, European Journal of Social Security 23 (2021) 323–340. URL: https://journals.sagepub.com/doi/full/10.1177/13882627211031257. doi:10.1177/13882627211031257.

[2] P. Lopez, Reinforcing intersectional inequality via the ams algorithm in austria (2019).

[3] A. Ben-David, Rule effectiveness in rule-based systems: A credit scoring case study, Expert Systems with Applications 34 (2008) 2783–2788. doi:10.1016/J.ESWA.2007.05.003.

[4] R. Parasuraman, M. Mouloua, Automation and Human Performance : Theory and Applications, 1996. doi:10.1201/9781315137957.

[5] J. D. Lee, K. A. See, Trust in automation: Designing for appropriate reliance, Human Factors (2004). doi:10.1518/hfes.46.1.50_30392.

[6] K. Goddard, A. V. Roudsari, J. C. Wyatt, Automation bias: a systematic review of frequency, effect mediators, and mitigators., Journal of the American Medical Informatics Association (2012). doi:10.1136/amiajnl-2011-000089.

[7] M. C. Elish, Moral crumple zones: Cautionary tales in human-robot interaction, Engaging Science, Technology, and Society 5 (2019) 40–60. URL: https://estsjournal.org/index.php/ests/article/view/260. doi:10.17351/ESTS2019.260.

[8] A. Birhane, The impossibility of automating ambiguity, Artificial life 27 (2021) 44–61. URL: https://pubmed.ncbi.nlm.nih.gov/34529757/. doi:10.1162/ARTL_A_00336.

[9] B. Wagner, Liable, but not in control? ensuring meaningful human agency in automated decision-making systems, Policy & Internet (2019). doi:10.1002/poi3.198.

[10] K. L. Mosier, K. L. Mosier, L. J. Skitka, M. Burdick, S. T. Heers, Automation bias, accountability, and verification behaviors, null (1996). doi:10.1177/154193129604000413.

[11] D. Manzey, J. Reichenbach, L. Onnasch, Human performance consequences of automated decision aids: The impact of degree of automation and system experience, Journal of Cognitive Engineering and Decision Making (2012). doi:10.1177/1555343411433844.

[12] R. Parasuraman, D. Manzey, Complacency and bias in human use of automation: an attentional integration., Human Factors (2010). doi:10.1177/0018720810376055.

[13] S. J. Karau, K. D. Williams, Social loafing: A meta-analytic review and theoretical integration., Journal of Personality and Social Psychology (1993). doi:10.1037/0022-3514.65.4.681.

[14] L. J. Skitka, K. L. Mosier, M. Burdick, Accountability and automation bias, International Journal of Human-computer Studies International Journal of Man-machine Studies (2000). doi:10.1006/ijhc.1999.0349.

[15] D. Lyell, E. Coiera, Automation bias and verification complexity: a systematic review., Journal of the American Medical Informatics Association (2016). doi:10.1093/jamia/ocw105.

[16] L. Chong, G. Zhang, K. Goucher-Lambert, K. Kotovsky, J. Cagan, Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of ai advice, Computers in Human Behavior (2022). doi:10.1016/j.chb.2021.107018.

[17] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, arXiv: Machine Learning (2018). doi:10.1038/s42256-019-0048-x.

[18] K. Sokol, P. A. Flach, Explainability fact sheets: a framework for systematic assessment of explainable approaches, FAT* (2020). doi:10.1145/3351095.3372870.

[19] M. Schemmer, N. Kühl, C. Benz, G. Satzger, On the influence of explainable ai on automation bias, European Conference on Information Systems (2022). doi:10.48550/arxiv.2204.08859.

[20] A. Bussone, S. Stumpf, D. O'Sullivan, The role of explanations on trust and reliance in clinical decision support systems, 2015 International Conference on Healthcare Informatics

(2015). doi:10.1109/ichi.2015.26.

[21] I. Sharon, A. Drach-Zahavy, E. Srulovici, The effect of outcome vs. process accountability-focus on performance: A meta-analysis, Frontiers in Psychology (2022). doi:10.3389/fpsyg.2022.795117.