PROCEDURAL FAIRNESS AS STEPPING STONE FOR SUCCESSFUL IMPLEMENTATION OF ALGORITHMIC DECISION-MAKING IN PUBLIC ADMINISTRATION: REVIEW AND OUTLOOK[1]

SVEN HOEPPNER, MARTIN SAMEK

Abstract: Algorithmic decision-making (ADM) is becoming more and more prevalent in everyday life. Due to their promise of producing faster, better, and less biased decisions, automated and data-driven processes also receive increasing attention in many different administrative settings. However, as a result of human mistakes ADM also poses the threat of producing unfair outcomes. Looming algorithmic discrimination can undermine the legitimacy of administrative decision-making. While lawyers and lawmakers face the age-old question of regulation, many decision-makers tasked with designing ADM for and implementing ADM in public administration wrestle with harnessing its advantages and limiting its disadvantages. "Algorithmic fairness" has evolved as key concept in developing algorithmic systems to counter detrimental outcomes. We provide a review of the vast literature on algorithmic fairness and show how key dimensions alter people's perception of whether an algorithm is fair. In doing so, we provide entry point into this literature for anybody who is required to think about algorithmic fairness, particularly in an public administration context. We also pinpoint critical concerns about algorithmic fairness that public officials and researchers should note.

Keywords: algorithmic decision-making; administration; procedural fairness; artificial intelligence

DOI: 10.14712/23366478.2024.24

1. INTRODUCTION

Algorithmic decision-making (ADM) increasingly shapes people's daily lives, albeit unbeknownst to many. However subtle, algorithms make important decisions not only in popularized contexts such as automated driving, selection of entertainment proposals on your favorite streaming platforms, and ChatGPT, but also in

[1] This article was written under the umbrella of the project "Regulatory Sandboxes: Mirage and Reality in Public Law", supported by Charles University's 4EU+ Mini-grants Programme.

potentially life-altering contexts such as hiring,[2] the legal system,[3] and also public administration.[4] Algorithms can be employed at almost every stage and branch of administration. Examples include the evaluation of claims for public benefits or the evaluation of publicly issued licenses for trade, weapons, or driving.

Employing ADM faces a trade-off. On the one hand, ADM can lead to faster and better decision outcomes.[5] For instance, in South Korea ADM was used to relocate ambulance units so that more people could receive emergency help within a five-minute time window of making an emergency call.[6] Algorithms can also reduce human biases in decision-making processes. For instance, they do not grow tired, have no agency, and are not distracted by emotional factors.[7] On the other hand, however, ADM often suffers from possible downsides in that (unfair) ADM systems can disparage certain members of society. Algorithms can decrease fairness.[8] For instance, ADM systems can systematically reinforce racial or gender stereotypes or marginalize minorities. But one example for such algorithmic discrimination is the – by now notorious – COMPAS algorithm, which disproportionally assigned a higher risk score of recidivism to black than to white defendants.[9] In the public administration context, ADM systems have also arbitrarily excluded citizens from food support programs, mistakenly reduced their disability benefits, or falsely accused them of fraud.[10]

Importantly, the negative biases of ADM are not a given. They result from human mistakes and are often unintended. First, human mistakes can occur in collecting and processing input data. When an algorithm learns from historical data about group features that is incomplete, unreliable, or biased, certain groups can be misrepresented by the data. This misrepresentation, in turn, can reproduce or exacerbate existing societal biases. Second, human mistakes can also occur in selecting, designing, specifying, and testing the algorithm. In such case an ADM system may perform fairly on some specific

[2] See e.g. ACIGKOZ, Y. – DAVIDSON, K. H. – COMPAGNONE, M. et al. Justice perceptions of artificial intelligence in selection. *International Journal of Selection and Assessment*. 2020, Vol. 28, No. 3, pp. 399–416; 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. *Business Research*. 2020, Vol. 13, No. 3, pp. 1–54.

[3] See e.g. CHOULDECHOVA, A. Fair Prediction with Disparate Impact: a Study of Bias in Recidivism Prediction Instruments. *Big Data*. Vol. 5, No. 2, pp. 153–163.

[4] See AlgorithmWatch. *Automating Society: Taking Stock of Automated Decision-Making in the EU* [online]. Bertelsmann Stiftung, 2019 [cit. 2023-12-14]. Available at: https://algorithmwatch.org/de/wp-content/uploads/2019/02/Automating_Society_Report_2019.pdf.

[5] See LEPRI, B. et al. Fair, Transparent, and Accountable Algorithmic Decision-making Processes: the Premise, the Proposed Solutions, and the Open Challenges. *Philosophy and Technology*. 2018, Vol. 31, No. 4, pp. 611–627.

[6] See NAM, T. Do the right thing right! Understanding the hopes and hypes of data-based policy. *Government Information Quarterly*. 2020, Vol. 37, No. 3, pp. 1–10.

[7] LEE, M. K. Understanding perception of algorithmic decisions: fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*. 2018, Vol. 5, No. 1, pp. 1–16.

[8] BAROCAS, S. – SELBST, A. D. Big data's disparate impact. *California Law Review*. 2016, Vol. 104, No. 1, pp. 671–729.

[9] CHOULDECHOVA, *c. d.*

[10] RICHARDSON, R. et al. Litigating Algorithms 2019 US Report: New Challenges to Government Use of Algorithmic Decision Systems. In: *AINow* [online]. 17. 9. 2029 [cit. 2024-01-02]. Available at: https://ainowinstitute.org/publication/litigating-algorithms-2019-u-s-report-2.

tasks but discriminate unfairly on others.[11] Third, transferring decision authority for sensitive issues from humans to ADM systems may in some circumstances be a human mistake in the first place.

How to harness the advantages while limiting the disadvantages of ADM in the administration context? How to constrain implementation difficulties that give rise to the downside of ADM? Avoiding each and every human mistake in setting up ADM systems is impossible. In light of these implementation difficulties, lawyers evoke their favorite answer: regulation. The European Union legislators are currently overwhelmed with the proposition of the AI Act,[12] discussing the uses of automation, algorithms, large language models and similar technologies in excruciating detail. The soon-to-be-adopted AI Act, upon which the EU Trialogue recently reached a political agreement,[13] also touches on the subject of AI and algorithm usage in public administration. In its explanatory memorandum section, the AI Act states that one of its main goals is to ensure a high level of protection of fundamental rights, among others, non-discrimination, the right to a fair process and also the general principle of good administration.[14] The newly adopted text, for example, bans social scoring and biometric categorization using data such as sexual orientation or religion. In some cases the AI Act also prohibits predictive policing for individuals.[15] There are also exemptions for law enforcement with prior judicial authorization and only for specific types of crimes.[16] Regarding the use of AI in public administration, the AI act specifically mentions employing systems similar to credit scoring in applications for social assistance, benefits, and services.[17] Since these ADM systems may be used to determine whether such benefits and services should be denied, reduced, revoked, or reclaimed by authorities they may have a significant impact on persons' livelihood and may infringe their fundamental rights. Therefore, under the AI Act such ADM systems automatically categorize them as "high-risk", which is the second-highest tier and the highest tier that is not prohibited.[18]

[11] See VEALE, M. – BINNS, R. Fairer machine learning in the real world: mitigating discrimination without collecting sensitive data. *Big Data & Society*. 2017, Vol. 4, No. 2, pp. 1–17; LEPRI, *c. d.*; EUBANKS, V. Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. New York: St. Martin's Press, 2018; KÖCHLING – WEHNER, *c. d.*

[12] See Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act), COM(2021) 206 final, 2021/0106(COD).

[13] See Commission welcomes political agreement on Artificial Intelligence Act. In: *European Commission: Press release* [online]. 11. 12. 2023 [cit. 2024-01-29]. Available at: https://digital-strategy.ec.europa.eu/en/news/commission-welcomes-political-agreement-artificial-intelligence-act.

[14] See Explanatory memorandum, article 3.5 of Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act), COM(2021) 206 final, 2021/0106(COD).

[15] See Artificial Intelligence Act: deal on comprehensive rules for trustworthy AI. In: *News: European Parliament: Press releases* [online]. 9. 12. 2023 [cit. 2024-01-29]. Available at: https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai.

[16] Ibid.

[17] Ibid. rec. (37).

[18] According to the AI Act proposal, there are two main categories of AI systems: prohibited and high-risk. The rest of AI systems are either under "certain AI systems" with specific obligations, or are out of scope of the AI Act. See point 5.2 of the explanatory memorandum.

In legal academia the AI Act has initially been met with less than lukewarm enthusiasm.[19] Legal scholarship has since spent much attention to debating *to what extent* new technologies can fruitfully be employed in administration, *how* legal administration can employ ADM in various tasks,[20] and which legal challenges their use brings forth.[21]

In ADM research, by contrast, one concept has become the key element in developing algorithmic systems to counter detrimental outcomes. That concept is algorithmic fairness.[22] Consequently algorithmic fairness has also become endorsed by as one of the main principles for trustworthy AI by the OECD[23] and the European Commission[24]. It has also been featured in more than 80% of guidelines for AI ethics.[25]

Much like abstract and open legal terms need continuous interpretive completion, the concept of algorithmic fairness requires more than just a technological solution. Employing algorithmic fairness in designing, implementing, and relying upon an ADM system requires a sophisticated empirical understanding of when, why, and how people perceive an algorithmic decision as fair or unfair. Only if algorithm-subjective individuals perceive the algorithmic decision as fair will they accept it as legitimate and, therefore, empirical insights into people's fairness perceptions are a *conditio sine qua non* for human-centric AI that informs developers entrusted with designing and users entrusted with implementing ethical ADM systems. Therefore, the social sciences have a huge potential to contribute to research on societal consequences of ADM, thus informing policymaking.[26]

[19] See e.g. VEALE, M. – BORGESIUS, F. Z. Demystifying the Draft EU Artificial Intelligence Act: analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*. Vol. 22, No. 4, pp. 97–112.

[20] See e.g. LIU, X. – LORINI, E. – ROTOLO, A. – SARTOR, G. Modelling and Explaining Legal Case-based Reasoners through Classifiers. *Frontiers in Artificial Intelligence and Applications, in corso di stampa*. 2022, pp. 1–13.

[21] In a nutshell, this line of research concentrates *on the procedure*. Scholars only recently started noticing the importance of possible attitude and behaviour changes that result from the use of new technologies. See e.g. COGLIANESE, C. Administrative Law in the Automated State. *Daedalus*. 2021, Vol. 150, No. 3, pp. 104–120.

[22] HUTCHINSON, B. – MITCHELL, M. 50 years of test (un)fairness: lessons for machine learning. In: *FAT* '19: Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. New York: Association for Computing Machinery, 2019, pp. 49–58.

[23] Recommendation of the Council on Artificial Intelligence. In: *OECD Legal Instruments* [online]. [cit. 2024-01-08]. Availabe at: https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449, also see: https://oecd.ai/en/ai-principles.

[24] Ethics guidelines for trustworthy AI. In: *European Commission: Shaping Europe's digital future* [online]. 8. 4. 2019 [cit. 2024-01-08]. Available at: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[25] JOBIN, A. – IENCA, M. – VAVERNA, E. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*. 2019, Vol. 1, No. 9, pp. 389–399.

[26] Compare BARABAS, CH. et al. Studying Up: Reorienting the study of algorithmic fairness around issues of power. In: *Proceedings of the ACM Conference on Fairness Accountability, and Transparency*. New York: Association for Computing Machinery, 2020, pp. 167–176; SLOANE, M. – MOSS, E. AI's social sciences deficit. *Nature Machine Intelligence*. 2019, Vol. 1, No. 8, pp. 330–331; KIESLICH, K. – KELLER, B. – STARKE, CH. Artificial intelligence ethics by design: evaluating public perception on the importance of ethical design principles of artificial intelligence. *Big Data & Society*. 2022, Vol. 9, No. 1, pp. 1–19. By drawing on an empirical understanding of citizens' fairness perceptions, algorithmic fairness can contribute to answering the call for a "society-in-the-loop" approach that embeds societal values in the design of ADM systems. See RAHWAN, I. Society-in-the-loop: programming the algorithmic social contract. *Ethics and Information Technology*. 2018, Vol. 20, No. 1, pp. 5–14;

In this paper, we want to inform both regulators of ADM systems and decision-makers in public administration tasked with deploying algorithms into administrative decision-making about main topics and main observations in ADM research on fairness perceptions regarding algorithmic decisions. The literature on this topic is vast, has witnessed a tremendous growth in the past decade, and continuous to strongly expand.[27] Our survey of the literature can, therefore, not be conclusive and our view on and selection of the literature is certainly biased by our own research interests. We intend this paper to serve as a mere entry point – or better yet: flood gate – to essential insights about people's perceptions of algorithmic fairness. In the context of new technologies in administration and administrative law, we contribute to the current discussion by providing a taxonomy of which features are crucial for fairness perceptions of ADM systems, which we hope will inform deployment of ADM systems in administration.

As an organizing concept, this paper relies on the topic procedural fairness as it has been studied in behavioral economics and organizational psychology.[28] In contrast to distributive or outcome-based fairness, procedural fairness in behavioral economics refers to the sensitivity of individuals towards differences in expected payoffs. Procedural fairness differs from other types of fairness, such as distributive fairness, in that it focuses on the fairness attributes of the decision-making process and the perceived fairness of the outcomes based on the procedures followed. Individuals who care about procedural fairness take additional factors of the decision-making process into account.[29] Here, we focus on the accuracy of algorithmic decisions, their transparency, and to what extent addressees of ADM systems have agency over the decision procedure and the outcome as core elements of the (algorithmic) allocation procedure.

This paper proceeds as follows. The next section will differentiate different fairness approaches that have been used in ADM research about fairness perceptions. The further sections then discuss the importance of (1) the accuracy of algorithmic decisions, (2) their transparency, and (3) to what extent addressees of ADM systems have agency over the

GERDON, F. – BACH, R. L. – KERN, CH. – KREUTER, F. Social impacts of algorithmic decision-making: a research agenda for the social sciences. *Big Data & Society*. 2022, Vol. 9, No. 1, pp. 1–13.

[27] Other authors have compiled fantastic and much more detailed surveys of this literature. We borrow heavily from: STARKE, CH. – BALEIS, J. – KELLER, B. – MARCINKOWSKI, F. Fairness perceptions of algorithmic decision-making: a systematic review of the empirical literature. *Big Data & Society*. 2022, Vol. 9, No. 2, pp. 1–16; KORDZADEH, N. – GHASEMAGEI, M. Algorithmic bias: review, synthesis, and future research directions. *European Journal of Information Systems*. 2022, Vol. 31, No. 3, pp. 388–409; WANG, X. – ZHANG, Y. – ZHU, R. A brief review on algorithmic fairness. *Management System Engineering*. 2022, Vol. 1, No. 7, pp. 1–13.

[28] Of course, the idea of procedural fairness also matters strongly in administrative law, which on general level mostly comprises a fair hearing rule and a rule against bias. For more on procedural fairness and its elements, see e.g. Recommendation CM/Rec(2007)7 of the Committee of Ministers to member states on good administration. Here, however, we are not interested in discussing when there is a duty to accord procedural fairness or what procedural fairness entails. Rather, we are interested in when people evaluate an administrative decision (by an algorithm) that allocates resources as fair. Therefore, we rely on the concept of procedural fairness as investigated in behavioral economics.

[29] Compare TRAUTMANN, S. T. Procedural fairness and equality of opportunity. *Journal of Economic Surveys*. 2023, Vol. 37, No. 5, pp. 1697–1714; KURZ, V. – ORLAND, A. – POSADZY, K. Fairness versus efficiency: how procedural fairness concerns affect coordination. *Experimental Economics*. 2018, Vol. 21, pp. 601–626; BOLTON, G. E. – BRANDTS, J. – OCKENFELS, A. Fair procedures: evidence from games involving lotteries. *The Economic Journal*. 2005, Vol. 115, No. 506, pp. 1054–1076.

decision procedure and the outcome for people's algorithmic fairness perceptions. The last section discusses the findings in the context of public administration and regulation, points out limitations, hints at possible venues for future research, and concludes.

2. ALGORITHMIC PROCEDURAL FAIRNESS

DIFFERENTIATION OF FAIRNESS APPROACHES

There appears to be no clear consensus on a precise definition of algorithmic fairness to date.[30] The stark heterogeneity among notions of algorithmic fairness is welcome, as it allows the concept to grow and evolve. The downside, for now, is that some terminological clarification is required, if only to avoid human errors in designing, specifying, and implementing fair ADM systems. This is, in fact, not dissimilar to typical legal tasks.

Algorithmic fairness generally is a consequentialist concept. In its very core, it entails that algorithmic decisions should not lead to unjust, discriminatory, or disparate consequences.[31] The literature distinguishes two broad approaches to algorithmic fairness. First, one part of the literature employs an axiomatic approach to algorithmic fairness and formalizes fairness criteria mathematically.[32] Second, another part of the literature draws on fairness concepts advanced in philosophy and the social sciences and applies them to questions of algorithmic fairness.[33]

The latter approach to algorithmic fairness can be further differentiated. On the one hand are researchers concerned with algorithmic predictors of algorithmic fairness, i.e., how an ADM procedure's technical design affects people's fairness perceptions. On the other hand are researchers who investigate human predictors of algorithmic fairness, i.e., what socio-economic, cultural, or other features of individuals affected by an ADM procedure determine their fairness perceptions about the ADM system. In what follows we intentionally bypass the latter because we assume that administrative services are extended to a broad public such that it may be too costly or otherwise difficult to cater to a heterogeneity of human predictors. Note however that decision-makers in public administration tasked with designing and implementing ADM systems should account

[30] SRIVASTAVA, M. – HEIDARI, H. – KRAUSE, A. Mathematical notions vs. Human perception of fairness: A descriptive approach to fairness for machine learning. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: Association for Computing Machinery, 2019, pp. 2459–2468.

[31] See SHIN, D. – PARK, Y. J. Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior*. 2019, Vol. 98, No. C, pp. 277–284.

[32] See e.g. GAJANE, P. – PECHENIZKIY, M. On Formalizing Fairness in Prediction with Machine Learning. In: *arXiv* [online]. 2017 [cit. 2023-12-13]. Available at: http://arxiv.org/abs/1710.03184; VERMA, S. – RUBIN, J. Fairness Definitions Explained. In: *Proceedings of the International Workshop on Software Fairness*. New York: Association for Computing Machinery, 2018, pp. 1–7; WANG – ZHANG – ZHU, *c. d.*; ŽLIOBAITĖ, I. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*. 2017, Vol. 31, No. 4, pp. 1060–1089.

[33] Seminal for this approach is BINNS, R. Fairness in machine learning: Lessons from political philosophy. *Proceedings of Machine Learning Research*. 2018, No. 81, pp. 149–159; BINNS, R. What can political philosophy teach us about algorithmic fairness? *IEEE Security & Privacy*. 2018, Vol. 16, No. 3, pp. 73–80.

for human predictors of algorithmic fairness if the ADM system provides a service to a specific subset of the population that can be characterized by specific features that matter for the fairness perception of individuals under the ADM system.

GENERAL OBSERVATIONS ON ALGORITHMIC PROCEDURAL FAIRNESS

Generally, fairness in is a crucial factor when evaluating algorithms.[34] However, research investigating people's general notions of algorithmic fairness yields mixed results.[35] Sometimes respondents perceive the very idea of ADM for important decisions based on past data unfair.[36] Other times participants argue that algorithms are by definition impartial.[37]

In additional to the mixed results on directionality, general fairness perceptions about algorithmic decision-making are highly dependent on the context. Some studies revealed that algorithmic fairness is perceived as more problematic in some domains than in others.[38] For instance, discrimination by ADM in housing, job recommendations, health care, or finance is viewed as more harmful than in music or movie recommendations.[39] Also, less complex algorithmic tasks elicited higher fairness perceptions than more complex ones.[40]

ACCURACY AND ALGORITHMIC FAIRNESS

Most studies go beyond people's general perception of algorithmic fairness and investigate how fairness perceptions are related to specific attributes of the

[34] See e.g. BANKINS, S. – FORMOSA, P. – GRIEP, Y. – RICHARDS, D. AI decision making with dignity? Contrasting workers' justice perceptions of human and AI decision making in a human resource management context. *Information Systems Frontiers*. 2022, Vol. 3, No. 2, pp. 1–19; ZHOU, J. – VERMA, S. – MITTAL, M. – CHEN, F. Understanding Relations Between Perception of Fairness and Trust in Algorithmic Decision Making. In: *Proceedings of the International Conference on Behavioral and Social Computing (BESC 2021)*. IEEE, 2021, pp. 1–5.

[35] See, e.g. DODGE, J. – LIAO, V. Q. – ZHANG, Y. – BELLAMY, R. K. E. – DUGAN, C. Explaining models: an empirical study of how explanations impact fairness judgment. In: *Proceedings of the International Conference on Intelligent User Interfaces*. New York: Association for Computing Machinery, 2019, pp. 275–285; SHIN – PARK, *c. d.*

[36] See BINNS, R. – VAN KLEEK, M. – VEALE, M. – LYNGS, U. – ZHAO, J. – SHADBOLT, N. It's Reducing Human Being to a Percentage: Perceptions of Justice in Algorithmic Decisions. In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. New York: Association for Computing Machinery, 2018, pp. 1–14.

[37] See LEE, M. K. – RICH, K. Who Is Included in Human Perceptions of AI?: Trust and Perceived Fairness around Healthcare AI and Cultural Mistrust. In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. New York: Association for Computing Machinery, 2021, pp. 1–14.

[38] HANNAN, J. – CHEN, H.-Y. W. – JOSEPH, K. Who Gets What According to Whom? An Analysis of Fairness Perceptions in Service Allocation. In: *Proceedings of the 2021 AAAI/ACMConference on AI, Ethics, and Society*. New York: Association for Computing Machinery, 2021, pp. 555–565.

[39] SMITH, J. – SONBOLI, N. – FIESLER, C. – BURKE, R. Exploring User Opinions of Fairness in Recommender Systems. In: *CHI'20 Workshop on Human-Centered Approaches to Fair and Responsible AI*. New York: Association for Computing Machinery, 2020, pp. 1–4.

[40] HSU, S. – LI, T. W. – ZHANG, Z. – FOWLER, M. – ZILLES, C. – KARAHALIOS, K. Attitudes Surrounding an Imperfect AI Autograder. In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. New York: Association for Computing Machinery, 2021, pp. 1–15.

algorithmic decision-making process. One element that has garnered a lot of attention is decision accuracy. Most if not all ADM systems are imperfect, just as human decision-makers. And, after all, high decision accuracy also strengthens decision consistency. How does ADM accuracy affect fairness perceptions?

Srivastava et al. use recidivism risk prediction and skin cancer risk predictions as examples for and ADM evaluation and elicit people's fairness choices. Interestingly, their data suggest that demographic parity best matched the fairness preferences of most participants, i.e., people favored algorithms aiming to equalize the positive rate across different groups. For instance, if ten percent of all applicants to a university get admitted, this rate should be equal for all gender groups. Regarding the issue of accuracy versus inequality, more importantly, the authors presented participants with three algorithms, each offering a different trade-off between accuracy and equality, and asked them to choose the one they consider ethically most desirable. For the case of medical risk prediction a high-stakes scenario described predicting the risk of skin cancer whereas a low-risk scenario described predicting the severity of flu symptoms. Similarly, for the case of recidivism risk prediction a high-stakes scenario describes that predictions are used to determine jail time whereas a low-stakes scenario describes that predictions are used to set bail amounts. Regardless of the decision context, the authors find that in high-stakes situations respondents attached a higher importance to accuracy than to inequality and vice versa.[41]

Three other studies inform the relationship between perceived fairness and decision accuracy. In another high-stakes context, i.e., an ADM system to assist child abuse hotline call workers in their screening decisions (child maltreatment prediction), Cheng et al. find that participants are willing to accept disparities in accuracy across groups than give up overall accuracy.[42] The results of Hsu et al. also highlight that accurate ADM is perceived as fairer than inaccurate ADM.[43] While their prediction context, i.e., automated college-level grading, may be viewed as low-stake, college grades and how accurately an ADM system decides about them are not low-stake for college students. Finally, in a tax fraud detection context, the results of Kieslich et al. suggest that their German participants weigh fairness and accuracy as equally important.[44]

Beyond accuracy in general, algorithmic fairness perceptions can be determined by the source of inaccuracy. And ADM system with an unbiased accuracy of 90% overall still suffers from 10% false positives and 10% false negatives. When an algorithm decided whether to grant bail to criminal defendants, participants of Harrison et al. were asked to decide about pairwise trade-offs between an ADM system that equalized one potentially desirable model property and that let another property vary across different racial groups and an ADM system that did the opposite. Harrison et al. observe that

[41] SRIVASTAVA – HEIDARI – KRAUSE, *c. d.*

[42] CHENG, H.-F. – STAPLETON, L. – WANG, R. – BULLOCK, P. – CHOULDECHOVA, A. – WU, Z. S. S. – ZHU, H. Soliciting Stakeholders' Fairness Notions in Child Maltreatment Predictive Systems. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. New York: Association for Computing Machinery, 2021, pp. 1–17.

[43] HSU – LI – ZHANG – FOWLER – ZILLES – KARAHALIOS, *c. d.*

[44] KIESLICH – KELLER – STARKE, *c. d.*

participants favor an algorithm that equalizes the false positive rate between groups over one that equalizes general accuracy.[45]

Yet another strand in this line of research focuses on how specific input features of an ADM system relate to its perceived fairness. Grgić-Hlača et al. use recidivism risk estimation and predictive policing as scenarios for their investigation. The authors focus on how the perceived fairness of an input feature is affected by additional knowledge about a desirable effect, i.e., an increase in accuracy, and an undesirable effect, i.e., an increase of disparity. Accordingly, they can define three measure of input-based process fairness: (1) feature-apriori fairness, i.e., a feature is perceived as fair, independent of its effect on the outcome; (2) feature-accuracy fairness, i.e., a feature is perceived as fair if it increases the accuracy of an algorithm; and (3) feature-disparity fairness, i.e., a feature is perceived as fair even if it increases disparity in the outcomes of an algorithm. Regardless of the decision scenarios, participants evaluated feature-accuracy fairness as most important, followed by feature-a-priori fairness, and feature-disparity fairness.[46] Similarly, Albach and Wright find that relevance of an input feature but also increases accuracy are the essential characteristics when deciding whether it is fair to use a feature in an ADM system.[47]

TRANSPARENCY AND ALGORITHMIC FAIRNESS

Transparency is central to the information dimension of procedural fairness. Without transparency other crucial aspects such as consistency, accountability, and revisability/revocability will be extraordinarily difficult to obtain. From our review regarding algorithmic procedural fairness, however, the effect of transparency on fairness perceptions appears to be understudied. Notable exceptions are Wang, who finds that algorithmic transparency increased perceptions of fairness, but not to an extent that algorithms are perceived to be preferable to human decision-makers[48], and Wang et al. who additionally observe that different degrees of transparency have no differential effect on perceived algorithmic fairness.[49] The reason for this lack of research may be inherent in the topic under investigation. Predictive systems often rely on sophisticated

[45] HARRISON, G. – HANSON, J. – JACINTO, CH. – RAMIREZ, J. – UR, B. An Empirical Study on the Perceived Fairness of Realistic, Imperfect Machine Learning Models. In: *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. New York: Association for Computing Machinery, 2020, pp. 392–402.

[46] GRGIĆ-HLAČA, N. – ZAFAR, M. B. – GUMMADI, K. P. – WELLER, A. Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning. In: *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*. New York: Association for Computing Machinery, 2018, pp. 51–60.

[47] ALBACH, M. – WRIGHT, J. R. The Role of Accuracy in Algorithmic Process Fairness Across Multiple Domains. In: *Proceedings of the ACM Conference on Economics and Computation*. New York: Association for Computing Machinery, 2021, pp. 29–49.

[48] WANG, A. J. Procedural Justice and Risk-Assessment Algorithms. In: *SSRN* [online]. 2018 [cit. 2024-01-04]. Available at: https://ssrn.com/abstract=3170136.

[49] WANG, R. – HARPER, M. F. – ZHU, H. Factors Influencing Perceived Fairness in Algorithmic Decision-Making: Algorithm Outcomes, Development Procedures, and Individual Differences. In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. New York: Association for Computing Machinery, 2020, pp. 1–14.

yet opaque machine learning models, which do not – or: hardly – facilitate an understanding of the general public, i.e., lay people who most often are subjected to a decision and who are study participants, how or why a given decision was arrived at.

Related to ex-ante transparency about the ADM system is providing information about input features, the decision-making process, and other reasons for the specific decision of an algorithm ex-post, i.e., after an algorithmic decision is made. Such ex-post transparency can be achieved through explanations for a decision. In fact, the European Union's General Data Protection Regulation (GDPR)[50] requires organizations deploying certain predictive systems to provide "meaningful information about the [decision] logic" to individuals affected by those predictive systems.[51] However, academic authors cannot agree whether "meaningful information" shall be interpreted without prejudice as a right to information. Selbst and Powles provide an informative summary and critical overview of this debate.[52] This right also is and will be further developed in practice with rulings of Court of Justice of the European Union, such as the case C634/21 which interpreted the provision in a context of an ADM used in a third party decision.[53] Furthermore, some authors think that the GDPR does not effectively regulate the use of predictive models, especially since these models can be trained on anonymized data, which falls completely out of scope of the GDPR.[54] As for the AI Act proposal, the transparency is covered under Article 4a on the general level[55] and especially for high-risk AI systems under Article 13[56] which in its current form seems very permitting as a right. The transparency is a necessary requirement for human oversight mentioned in our concluding remarks.

The set of studies that investigated the relation of explanations and perceived algorithmic fairness is surprisingly large, given that our reading of the literature suggests that transparency and algorithmic fairness remains somewhat of a research gap. As a general result, explanations for an algorithmic decision increase people's perception of procedural fairness. This is good news as administrative decisions, human or algorithmic, need to come with motivated explanations. By contrast, perceptions of interpersonal and

[50] Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (GDPR).

[51] See Article 13, para. 1, let. f) of the GDPR.

[52] SELBST, A. D. – POWLES, J. Meaningful information and the right to explanation. *International Data Privacy Law*. 2017, Vol. 7, No. 4, pp. 233–242. Available online in: *Oxford Academy* [online]. [cit. 2024-02-21]. Available at: https://doi.org/10.1093/idpl/ipx022.

[53] In Judgement of the Court from 7 December 2023, OQ v. Land Hessen, (SCHUFA Holding AG), C634/21, ECLI:EU:C:2023:957 the court interpreted Article 22 of the GDPR in such way, that the article is applicable for situations where the information from ADM system is used to influence a decision of a third party.

[54] MÜHLHOFF, R. Predictive privacy: Collective data protection in the context of artificial intelligence and big data. *Big Data & Society* [online]. 2023, Vol. 10, No. 1 [cit. 2023-01-04]. Available at: https://doi.org/10.1177/20539517231166886.

[55] Article 4 of AI Act proposal states that: *"AI systems shall be developed and used in a way that allows appropriate traceability and explainability while making humans aware that they communicate or interact with an AI system as well as duly informing users of the capabilities and limitations of that AI system and affected persons about their rights."*

[56] Article 13 of the AI Act proposal states that *"high-risk AI systems shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable providers and users to reasonably understand the system's functioning"*.

distributive fairness do not seem to be affected by the provision of explanations.[57] The results are very nuanced, however. For instance, the results of Shulner-tal et al. suggest that providing some kind of explanation contributes to participants' understanding of the outcome. Yet, while explanations provided by the system are important for increasing participants' perception of fairness, their fairness perceptions mainly depend on the decision of the system. In other words, the effect of providing explanations can get overwritten by other characteristics of the ADM.[58] Moreover, Lee et al. find that the effect of outcome explanations for perceived algorithmic fairness may well be context-dependent. When outcome explanations helped participants in their qualitative study to understand biased outcome distributions, perceived fairness decreased. By contrast, when outcome explanations for helped them to understand rather equal distributions, perceived fairness increased.[59]

Other studies investigate the effects of different styles of explanations for people's algorithmic fairness perceptions. Binns et al. provide respondent's with five application contexts (personal financial loan, promotion at work, car insurance premiums, overbooking of airline flights, and freezing of bank accounts) and created different scenarios in which an algorithmic decision negatively affected one individual. They varied the scenarios regarding whether explanation for the negative decision is provided and, if so, also varied the explanation using four different explanation styles: (1) presenting an algorithm's input variables and a quantitative measure of their influence; (2) for each input variable used in a decision, providing a sensitivity analysis that shows how much the value of that variable would have to differ in order to change the decision; (3) presenting a case from the model's training data which is most similar to the decision being explained; and (4) presenting aggregate statistics on the decisions for people in the same demographic categories as the decision-subject, such as age, gender, income level or occupation. The results are not as straight-forward as one would hope, but rather depend on the experimental design. On the one hand, participants strongly engaged with the details of each explanation (when discussing a case in the qualitative part of the study). This occurred within-subjects, i.e., when participants were presented with different explanation styles. In particular, case-based explanation styles impacted fairness perceptions negatively, especially compared to sensitivity-based explanation styles. On the other hand, however, in between-subjects designs, i.e., when participants are exposed to only one explanation style across multiple scenarios, these explanation effects largely disappeared.[60] Similar results have been found in a recidivism risk con-

[57] See SCHLICKER, N. – LANGER, M. – ÖTTING, S. K. – BAUM, K. – KÖNIG, C. J. – WALLACH, D. What to expect from opening up 'black boxes'? Comparing perceptions of justice between human and automated agents. *Computers in Human Behavior*. 2021, Vol. 122, No. 4, pp. 1–16.

[58] See SHULNER-TAL, A. – KUFLIK, T. – KLIGER, D. Fairness, explainability and in-between: Understanding the impact of different explanation methods on non-expert users' perceptions of fairness toward an algorithmic system. *Ethics and Information Technology*. 2022, Vol. 24, No. 1, pp. 1–13.

[59] See LEE, M. K. – JAIN, A. – CHA, H. J. – OJHA, S. – KUSBIT, D. Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. *Proceedings of the ACM on Human-Computer Interaction*. 2019, Vol. 3, No. CSCW, Art. 182, pp. 1–26.

[60] BINNS – VAN KLEEK – VEALE – LYNGS – ZHAO – SHADBOLT, *c. d.*

text, in that certain explanations are considered inherently less fair, while others can enhance people's confidence in the fairness of the algorithm.[61]

We think that one result from the literature deserve special attention, especially in a journal with a substantial audience with a legal background. Nyarko et al. study attitudes towards "blinding" algorithms. Many scholars, algorithm developers, and – not least of all – policymakers posit that algorithmic fairness requires excluding information about certain characteristics of individuals, such as their race or gender, as input variables, especially regarding minority groups. "Blinding" algorithms in this way is often conveyed as an unconditional ethical imperative, i.e., a minimal requirement of fair treatment, and any contrary practice is assumed to be morally and politically untenable.[62] However, excluding information about race or gender from algorithmic decisions can in fact lead to worse outcomes for racial minorities and women in some circumstances, which complicates the rationale for blinding. In a set of randomized studies, Nyarko et al find that people are generally averse to the use of sensitive information such as race and gender in algorithmic predictions of pretrial risk. They also find, however, that this preference for excluding sensitive information shifts in response to a relatively mild intervention, namely when respondents are provided with factually correct information about the possibility that this exclusion of sensitive input variables could lead to higher detention rates for black and female defendants, respectively.[63]

CONTROL, PARTICIPATION, OR AGENCY: VOICE AND REVOCABILITY

Thibaut and Walker developed the control model of procedural fairness: procedural fairness is a function of the degree of control over the decision that individuals receive. When individuals perceive to have more control over the processes that lead to decision outcomes ("process control") and the decision outcomes ("outcome control"), they perceive the results to be fairer.[64] Process control is the ability to influence what evidence or data is considered by the decision-maker, how that evidence is presented, and the rules by which the evidence is interpreted. In the realm of ADM, process control may include allowing individuals to determine input data or giving individuals the ability to influence the rules and logics of the algorithm itself. Outcome control refers to the ability to appeal or modify the outcome of a decision once it has been made and enables correctability and possible recourse against decisions that are

[61] See DODGE – LIAO – ZHANG – BELLAMY – DUGAN, *c. d.*

[62] Compare KLEINBERG, J. – LUDWIG, J. – MULLAINATHAN, S. – RAMBACHAN, A. Algorithmic fairness. *AEA papers and proceedings*. 2018, Vol. 108, pp. 22–27.

[63] NYARKO, J. – GOEL, S. – SOMMERS, R. Breaking Taboos in Fair Machine Learning: An Experimental Study. In: *Proceedings of Equity and Access in Algorithms, Mechanisms, and Optimization*. New York: Association for Computing Machinery, 2021, pp. 1–11.

[64] THIBAUT, J. W. – WALKER, L. *Procedural justice: a psychological analysis*. New York: L. Erlbaum Associates, 1975. Also compare LIND, A. E. – LISAK, R. I. – CONLON, D. E. Decision Control and Process Control Effects on Procedural Fairness Judgments 1. *Journal of Applied Social Psychology*. 1983, Vol. 13, No. 4, pp. 338–350.

wrong or improper.[65] In the realm of ADM, outcome control will enable individuals to reject algorithmic decisions through appeal or by finding alternative outcomes.

Again, while there is recent recognition that this is an important area,[66] our reading of the literature suggests that the effect of control (or participation, or agency) on perceptions of algorithmic fairness remains relatively unexplored. One exception is the interview study of Hsu et al., who find that the discontent expressed by participants about the false negative rate and the subsequent decrease in fairness perceptions, could be mitigated by an appeal procedure to some extent.[67] In a set of vignette studies, the results of Sun and Tang also suggest that participants who have control over avoiding algorithmic discrimination on a booking website for airline flights increases perceived algorithmic fairness.[68] Finally, both process control and outcome also increased perceptions of algorithmic fairness in the interview study of Lee et al.[69]

3. DISCUSSION & CONCLUSION

As even more ADM system permeate public administration and society in general, decision-makers tasked with implementing ADM should be concerned about algorithmic fairness beyond the technical properties of a given machine learning model, if only to instill more acceptability by addressees of and legitimacy of ADM systems. Most technology-specific regulation aim at remedying potential downsides stemming from the interaction of technology and society.[70] In the ADM context, legislators have not been sleepy either. Intentionally or not, the proposal for the AI Act by the EU features some elements that may strengthen perceived algorithmic fairness. For example, Article 14 of the AI Act proposal states that AI systems must be designed such that they can be effectively overseen by human decision-makers. However, the proposal remains vague on responsibilities of human overseers, despite that "human-in-the-middle" principle has been prominent in talks of EU legislators and in the explanatory section of proposal as well.[71] The Article 14 sets out certain parameters or qualities that human overseers should possess *in theory*. And while the Recital 48 proposes some level of

[65] Compare LEVENTHAL, G. S. What should be done with equity theory? In: GERGEN, K. J. – GREENBERG, M. S. – WILLIS, R. H. (eds.). *Social exchange: Advances in Theory and Research*. Boston: Springer, 1976, pp. 27–55; and also compare HOULDEN, P. – LATOUR, S. – WALKER, L. – THIBAUT, J. Preference for modes of dispute resolution as a function of process and decision control. *Journal of Experimental Social Psychology*. 1978, Vol. 14, No. 1, pp. 13–30.

[66] Compare HIRSCH, T. – MERCED, K. – NARAYANAN, S. – IMEL, Z. E. – ATKINS, D. C. Designing contestability: Interaction design, machine learning, and mental health. In: *Proceedings of the 2017 Conference on Designing Interactive Systems*. New York: Association for Computing Machinery, 2017, pp. 95–99.

[67] HSU – LI – ZHANG – FOWLER – ZILLES – KARAHALIOS, *c. d.*

[68] SUN, L. – TANG, Y. Data-Driven discrimination, perceived fairness, and consumer trust: the perspective of consumer attribution. *Frontiers in Psychology*. 2021, No. 12, pp. 1–13.

[69] LEE – JAIN – CHA – OJHA – KUSBIT, *c. d.*

[70] For a more detailed discussion about determinism and technology and law, see e.g. SCHREPEL, T. Law and Technology Realism. *MIT Computational Law Report* [online]. 2020 [cit. 2024-01-03]. Available at: https://law.mit.edu/pub/lawandtechnologyrealism/release/3.

[71] E.g. Rec. 4a), 6), 14a), 32a), 43), 48) or 70a) of the AI Act proposal.

responsibility of human overseers, recitals are, by their nature, not binding and the proposal in its current form lacks a clear guidance on the scope of overseers' responsibility. Moreover, as pointed out above ADM systems might not be very conducive to being overseen – at least not by third parties without an extremely high of expertise. Even the most explainable AI (XIA) system may feature too much data and a "black box" learning model to warrant a fair and transparent decision. Note also that the AI Act proposal does not provide a clear recommendation addressees of ADM systems should request assistance from human overseers, which limits addressees' agency. In our view, public officials regulating and implementing ADM should pay much closer attention to drivers for perceived fairness in ADM and also try harder to find ways of employing insights from that line of research.

On a conciliatory note, regulators and ADM practitioners are confronted with the patchwork nature of current research, which complicates implementing and regulating procedural algorithmic fairness. One key takeaway from our review is that people's fairness perceptions regarding ADM systems are highly dependent on context, such as are of application and the specific decision task at hand. Moreover, existing theoretical fairness concepts (e.g. distributional, procedural, etc.) are not consistently used in the literature. More research on the theoretical underpinnings of (procedural) algorithmic fairness is required, not only as an intellectual exercise but also to facilitate smart regulation. As regulators may regulate different domains differently, we also call for extending research beyond the classical use cases of ADM in, e.g., the criminal justice system, human resources, and airline ticket booking. ADM has surged in many other areas of society, such as medical decision-making, health management, insurance pricing, credit scoring, and distributing social services and benefits. We also argue for a higher diversity of research methods to be employed by researchers studying algorithmic fairness perceptions. While the literature features an appealing mix on qualitative and quantitative methods, in our reading the quantitative studies almost exclusively rely on vignette studies in which participants are exposed to varied scenario descriptions and then asked for their fairness evaluations regarding these scenarios. We propose to add to the method mix much more controlled choice experiments in which subjects themselves are exposed to algorithmic decisions. Lastly, we encourage researchers to study the comparative effects of standalone ADM systems versus ADM systems that serve as decision support on perceived algorithmic fairness, which we did not find in the section of the literature that we reviewed.

As ADM increasingly extends into different sectors of society, concerns about the distributive and procedural fairness of those systems arise. This is especially crucial in the relation between citizens and state, which is governed by public administration. Luckily, in western democracies human public administration already sports a large degree of procedural fairness. Citizens participating in administrative procedures possess information right, objection rights, the right to contribute evidence and express their opinion, all of which are beneficial to the agency dimension of procedural fairness. Similarly, public officials are required to provide motivated reasoning for their decisions, which is helpful for the transparency dimension of procedural fairness. When implementing ADM in administrative procedures, public officials need to ensure that fairness perceptions are not crowded-out and carefully think about the societal implications of employing ADM in their routine and non-routing tasks and procedures. Suppose public institutions fail to design and implement ADM systems that are perceived as fair, perhaps because they deny citizens a voice in the decision process or produce seemingly arbitrary results. In that case, citizen could become alienated or lose trust in public institutions.[72] Once trust is lost, they might become more vulnerable to populist rethoric. Decision-makers tasked with designing and implementing ADM in public administration are, therefore, well advised to engage empirical researchers who work on human-machine transactions and jointly conduct a field test of the to-be-implemented ADM system first. Only if satisfied, not only with the technical functionality of the system, but also the addressees fairness perceptions, and other societal consequences, should public officials implement ADM on a full-size scale.

Sven Hoeppner, M.Sc., LL.M., Ph.D.
Charles University, Faculty of Law
hoeppnes@prf.cuni.cz
ORCID: 0000-0003-2697-4420

Mgr. Martin Samek
Charles University, Faculty of Law
samek@prf.cuni.cz
ORCID: 0000-0002-0302-0520

[72] COGLIANESE, *c. d.*