INTERNET POLICY REVIEW
Journal on internet regulation

Volume 9 | **Issue** 4

CONCEPT

OPEN ACCESS

PEER REVIEWED

Algorithmic bias and the Value Sensitive Design approach

Judith Simon *Universität Hamburg***Pak-Hang Wong** *Universität Hamburg*
Gernot Rieder *Universität Hamburg*

DOI: https://doi.org/10.14763/2020.4.1534

Published: 18 December 2020
Received: 29 October 2020 **Accepted:** 26 November 2020

Competing Interests: The author has declared that no competing interests exist that have influenced the text.

Citation: Simon, J. & Wong, P.-H. & Rieder, G. (2020). Algorithmic bias and the Value Sensitive Design approach. *Internet Policy Review*, *9*(4). https://doi.org/10.14763/2020.4.1534

Keywords: Value sensitive design, Algorithmic bias, Human values, Fairness, Fairness in Machine Learning

Abstract: Recently, amid growing awareness that computer algorithms are not neutral tools but can cause harm by reproducing and amplifying bias, attempts to detect and prevent such biases have intensified. An approach that has received considerable attention in this regard is the Value Sensitive Design (VSD) methodology, which aims to contribute to both the critical analysis of (dis)values in existing technologies and the construction of novel technologies that account for specific desired values. This article provides a brief overview of the key features of the Value Sensitive Design approach, examines its contributions to understanding and addressing issues around bias in computer systems, outlines the current debates on algorithmic bias and fairness in machine learning, and discusses how such debates could profit from VSD-derived insights and recommendations. Relating these debates on values in design and algorithmic bias to research on cognitive biases, we conclude by stressing our collective duty to not only detect and counter biases in software systems, but to also address and remedy their societal origins.

This article belongs to **Concepts of the digital society**, a special section of *Internet Policy Review* guest-edited by Christian Katzenbach and Thomas Christian Bächle.

1. Introduction

When, in 2016, investigative journalists at ProPublica published a report indicating that a software system used in US courts was racially biased, a lively debate ensued. In essence, the journalists had found that COMPAS, a decision support tool used by judges and parole officers to assess a defendant's likelihood to re-offend, was systematically overestimating the recidivism risk of black defendants while underestimating that of white defendants (see Angwin et al., 2016). Northpointe, the company that developed COMPAS, disputed the allegations, arguing that its assessment tool was fair because it predicted recidivism with roughly the same accuracy regardless of defendants' ethnicity (see Dieterich et al., 2016). The ProPublica journalists, in turn, held that an algorithmic model cannot be fair if it produces serious errors, that is, false positives (i.e., false alarms) and false negatives (i.e., missed detections), more frequently for one ethnicity than for another, triggering a debate about the very idea of programming fairness into a computer algorithm (see, e.g., Wong, 2019). To date, over 1,000 academic papers have cited the ProPublica article, [1] and its findings have been discussed in popular news outlets around the globe.

But the ProPublica case was not a one-off. Rather, it marked the beginning of a series of reports and studies that found evidence for algorithmic bias in a wide range of application areas: from hiring systems (Dastin, 2018) to credit scoring (O'Neil, 2016) to facial recognition software (Buolamwini and Gebru, 2018). Cases such as these, which highlight the potential for automated discrimination based on characteristics such as age, gender, ethnicity, or socio-economic status, have reinvigorated old debates regarding the relationship between technology and society (see, e.g., Winner, 1980), questioning the neutrality of algorithms and inviting discussions about their power to structure and shape, rather than merely reflect, society. However, if technologies are not morally neutral and if the values and disvalues embedded have tangible consequences for both individuals and society at large, would this not imply that algorithms should be designed with care and that one should seek not only to detect and analyse problems, but to proactively engage with them through mindful design decisions? [2] Such questions, which are now be-

[1] See the article's citation count on Google Scholar at https://scholar.google.com/scholar?cites=9718961392046448783&as_sdt=2005&sciodt=0,5&hl=en.

ing discussed within the computer science community, are not new, but have a long and often neglected history within computer science itself—e.g., through research in participatory design—but also in other fields and disciplines such as computer ethics, philosophy of technology, history of science, or science and technology studies (STS). The most principled attempt to design responsibly and sensitively to human values, however, is the Value Sensitive Design (VSD) approach, which emerged out of this intellectual landscape in the mid-1990s and has been expanded and refined ever since. More recently, and as result of increased awareness that "data is not a panacea" and that algorithmic techniques can "affect the fortunes of whole classes of people in consistently unfavorable ways" (Barocas and Selbst, 2016, p. 673), interest in the VSD methodology has been growing, begging the question: what insights can the approach offer to ongoing debates about bias and fairness in algorithmic decision-making and machine learning?

This article provides a brief overview of the key features of Value Sensitive Design (Section 2), examines its contributions to understanding and addressing issues around bias in computer systems (Section 3), outlines the current debates on algorithmic bias and fairness in machine learning (Section 4), and discusses how such debates could profit from VSD-derived insights and recommendations (Section 5). Relating these debates on values in design and algorithmic bias to research on cognitive biases, we conclude by stressing our collective duty to not only detect and counter biases in software systems, but to also address and remedy their societal origins (Section 6).

2. Value Sensitive Design: a brief overview

Value Sensitive Design as a theoretically grounded methodology emerged against the backdrop of the 1990s rapid computerisation and as a response to a perceived need for a design approach that would account for human values and social context throughout the design process (see Friedman and Hendry, 2019). Indeed, Friedman's (1997) seminal edited book *Human Values and the Design of Computer Technology* already provided an impressive demonstration on how to conceptualise and address issues around agency, privacy, and bias in computer systems, emphasising the need to "embrace value-sensitive design as part of the culture of computer science" (ibid.: p. 1). At its core, the VSD approach offers a concrete methodology for how to intentionally embed desired values into new technologies. It con-

[2] In this paper, we use the term *value* to refer to "[t]hose things that people find valuable that are both ideal and general" and the term *disvalue* to refer to "those general qualities that are considered to be bad or evil" (Brey, 2010, p. 46).

sists of three iterative phases, namely conceptual-philosophical, empirical, and technical investigations (see Friedman et al., 2006; Flanagan et al., 2008): [3]

Conceptual-philosophical investigations encompass both the identification of relevant human values and the identification of relevant direct and indirect stakeholders. Regarding the former, careful working conceptualisations of specific values are meant to (a) clarify fundamental issues raised by the project at hand and (b) enable comparisons across VSD-based studies and research teams. While VSD defines *human values* relatively broadly as "what is important to people in their lives, with a focus on ethics and morality" (Friedman and Hendry, 2019, p. 24), Friedman et al. (2006, p. 364f) have provided a heuristic list of human values with ethical import [4] that are often implicated in system design. Regarding the latter, by not only considering direct but also indirect stakeholders, VSD aims to counter the frequent neglect of non-users in technology design, that is, of groups which may not use a technology themselves, but who are nonetheless affected by it (see Oudshoorn and Pinch, 2005; Wyatt, 2005). Given that values are often interrelated—consider, e.g., the ongoing debate about the relationship between privacy and security—and that what is important to one group of stakeholders may or may not be important to another group, conceptual investigations are also concerned with the relative importance of different values as well as potential trade-offs between conflicting values.

Empirical investigations make use of a wide range of quantitative and qualitative social science methods (e.g., surveys, interviews, observations, experiments) to provide a better understanding of how stakeholders actually conceive and prioritise values in specific socio-technical contexts. Cultural, historical, national, ethnic, and religious affiliations may play a role in this process and can determine how value conflicts are handled and resolved (see Flanagan et al., 2008, p. 328). Moreover, empirical investigations may reveal differences between espoused practice (what is said) and actual practice (what people do), enabling a more nuanced analysis of design decisions and their impact on usage, thereby complementing the conceptual investigations outlined above. Ultimately, it is through this empirical mode of inquiry that a more situated understanding of the socio-technical sys-

[3] The following paragraphs are a reworked and expanded version of section 1 in "Value-Sensitive Design as a Methodology" (Simon, 2017).

[4] Examples of such "values with ethical import" include *privacy*, meaning "the right of an individual to determine what information about himself or herself can be communicated to others"; *autonomy*, meaning "people's ability to decide, plan, and act in ways that they believe will help them to achieve their goals"; or *informed consent*, which refers to "garnering people's agreement, encompassing criteria of disclosure and comprehension (for 'informed') and voluntariness, competence, and agreement (for 'consent')" (Friedman et al., 2006, p. 364).

tem can be derived, facilitating not only the observation of stakeholders' usage and appropriation patterns, but also whether the values envisioned in the design process are fulfilled, amended, or subverted.

Technical investigations are premised on the assumption that any given technological design provides "value suitabilities" (Friedman and Hendry, 2019, p. 34) in that it supports certain values and activities more readily than others. Following Friedman et al. (2008), investigations into these suitabilities can take one of two forms: in the first form, technical investigations focus on how existing technological properties can support or hinder specific human values. This approach bears similarities to the empirical mode, but instead of focusing on individuals, groups, or larger social systems, the emphasis is on the technology itself. In the second form, technological investigations involve the proactive design of systems to support and realise values identified in the conceptual investigation. If, for instance, privacy is a value that ought to be preserved, technical mechanisms must be implemented that further and promote privacy protections rather than diminish them. As specific designs will prioritise certain values over others, technical investigations can reveal both existing (first form) or prospective (second form) value hierarchies, thus adding another layer of insight to the analysis.

Through these three modes of investigation, VSD aims to contribute to the critical analysis of socio-technical systems and the values that have been—intentionally or unintentionally—embedded into them. Accordingly, VSD on the one hand serves as an analytical tool to open up valuation processes within technology design and development that are usually black-boxed or neglected. On the other hand, it provides a constructive tool that enables and supports the realisation of specific desired values in the design and development of new technologies. [5]

3. Bias in computer systems

Long before the current debate about algorithmic bias and its consequences, Friedman and Nissenbaum (1996) had already pioneered an analysis of bias in computer systems, arguing that such systems are biased if they "systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others [by denying] an opportunity for a good or [assigning] an undesirable outcome to an individual or groups of individuals on grounds that are unreasonable or inappropriate" (ibid.: p. 332). For Friedman and Nissenbaum it was important to

[5] Of course, like any mature methodology, the Value Sensitive Design approach has also been subject to a good deal of critique (see Friedman and Hendry, 2019, p. 172f). For a comprehensive review of these critiques, see Davis and Nathan (2014).

develop a better understanding of bias in computer systems, not least because they considered biased systems to be "instruments of injustice" and stressed that "freedom from bias should be counted among the select set of criteria according to which the quality for systems in use in society should be judged" (ibid.: p. 345f). A good understanding of biases would allow us to identify potential harms in a system and either avoid them in the process of design or correct them if the system is already in use. To this end, Friedman and Nissenbaum provided a taxonomy of biases that remains highly relevant and useful for today's debate on algorithmic bias and discrimination (see, e.g., Dobbe et al., 2019; Cramer et al., 2018). Based on the respective origin of bias, they specified three different types of biases, namely *preexisting bias*, *technical bias*, and *emergent bias*.

According to Friedman and Nissenbaum (1996), *preexisting bias* has its roots in social institutions, practices, and attitudes and usually exists prior to the creation of the system. It can either originate from individuals who have significant input into the design of the system (individual preexisting bias) or from prejudices that exist in society or culture at large (societal preexisting bias). Importantly, such biases mostly enter a system implicitly and unconsciously rather than through conscious effort.

Technical bias, in turn, arises from technical constraints or considerations. Sources of technical bias may include limitations of computer tools (e.g., in terms of hardware, software, or peripherals), the use of algorithms that have been developed for a different context, and the unwarranted formalisation of human constructs, that is, the attempt to quantify the qualitative and discretise the continuous.

Finally, *emergent bias* is bias that arises in a context of use, typically some time after a design is completed, as a result of (a) new societal knowledge or changing cultural values that are not or cannot be incorporated into the system design or (b) a mismatch between the users—their expertise and values—assumed in the system design and the actual population using the system.

In sum, Friedman and Nissenbaum's taxonomy of biases is meant to enable designers and researchers to identify and anticipate bias in computer systems by considering individual and societal worldviews, technological properties, and the contexts of use. Their analysis of biases foregrounds the value-laden nature of computer systems and stresses the possibility to mitigate or eliminate potential harms through *proactively* engaging with the design and development of the systems, which is one of the main objectives of the Value Sensitive Design approach. Consequently, their analysis reflects the double function of VSD as a tool for the analysis of (dis)values in existing technologies and for the construction of novel technologies that account for specific desired values. These two functions, the analytical and the constructive, are also central in recent research on bias and fairness in machine learning.

4. Algorithmic bias and fairness in machine learning

When mathematician Cathy O'Neil published her popular book *Weapons of Math Destruction* in 2016, the message was clear: Mathematical models, she wrote, can "encod[e] human prejudice, misunderstanding, and bias into the software systems that increasingly manag[e] our lives. [...] Their verdicts, even when wrong and harmful, [are] beyond dispute or appeal. And they ten[d] to punish the poor and the oppressed in our society, while making the rich richer" (O'Neil, 2016, p. 3). To support this claim, O'Neil works through a number of cases—from crime prediction software and personalised online advertising to college ranking systems and teacher evaluation tools to credit, insurance, and hiring algorithms—demonstrating the punitive power such systems can have on those who already suffer from social inequalities and emphasising the task to "explicitly embed better values into our algorithms, creating 'Big Data' models that follow our ethical lead" (ibid.: p. 204). O'Neil's book, along with a few other academic and non-academic texts, was at the forefront of a movement that sought to push back against the depiction of algorithms as fair and objective, showcasing their potential to "so[w] injustice, until we take steps to stop them" (ibid.: p. 203).

In the computer science community, where research on bias and discrimination in computational processes was conducted even prior to the current debate on the impacts of "Big Data" and *artificial intelligence* (see, e.g., Custers et al., 2013), attempts to detect and prevent such biases intensified. An example for this would be the organisation of the yearly *FAT/ML*[6] annual meeting from 2014 onwards, which in light of a growing recognition that techniques such as machine learning raise "novel challenges for ensuring non-discrimination, due process, and understandability in decision-making," sought to "provid[e] researchers with a venue to explore how to characterize and address these issues with computationally rigorous methods" (FAT/ML, 2018). Other events such as the *DAT (Data and Algorithmic Transparency) Workshop* in 2016 or the *FATREC Workshop on Responsible Recommendation* in 2017 followed, and the *FAT/ML* meeting was eventually succeeded by the *FAT** and later the *ACM FAccT Conference,* which seeks to bring together "researchers and practitioners interested in fairness, accountability, and transparency

[6] The acronym FAT/ML stands for Fairness, Accountability and Transparency in Machine Learning.

in socio-technical systems" (ACM FAccT Conference, 2021). As mentioned, this research and the VSD approach find common ground in their twofold objective to (a) identify bias and discrimination in algorithmic systems (analytical objective) and (b) to create and design fair algorithmic systems (constructive objective). [7]

With respect to (a), researchers of algorithmic bias have proposed different frameworks for understanding and locating the sources of algorithmic biases, thereby delineating ways to mitigate or correct them (see, e.g., Baeza-Yates, 2018; Mehrabi et al., 2019; Olteanu et al., 2019). Barocas and Selbst (2016), for instance, provide a detailed description of the different ways that biases can be introduced into a machine learning system, including (i) through problem specification, where the definition of target variables rests on subjective choices that may systematically disadvantage certain populations over others; (ii) through the training data, where biased data sets can lead to discriminatory models and harmful results; (iii) through feature selection, where the reductive representation of real-world phenomena may result in inaccurate determinations and adverse effects; and (iv) through proxy variables, where specific data points are highly correlated with class membership, facilitating disparate treatment and potentially implicating less favorable outcomes for members of disadvantaged groups. In a similar vein, Danks and London (2017) identify different forms of algorithmic bias in autonomous systems, namely (i) training data bias, (ii) algorithmic focus bias, (iii) algorithmic processing bias, (iv) transfer context bias, and (v) interpretation bias. The parallels between such recent approaches and Friedman and Nissenbaum's earlier work become most apparent when considering their common goal to sound a "call for caution" (Barocas and Selbst, 2016, p. 732), provide "a taxonomy of different types and sources of algorithmic bias" (Danks and London, 2017, p. 4691), and offer a "framework for understanding and remedying it" (Friedman and Nissenbaum, 1996, p. 330). In either case, the designation and characterisation of different types of biases is thus seen as a key element of the common analytical objective to recognise and remedy such biases in existing algorithmic systems.

With respect to (b), and in addition to the analytical task of identifying and mitigating bias, there is also a more constructive aspiration in the machine learning community to design *fair algorithms*. Kearns and Roth, for instance, describe this aspiration as the "science of socially aware algorithm design" that looks at how algorithms can "incorporate – in a quantitative, measurable, verifiable manner – many of the ethical values we care about as individuals and as a society" (2019, p.

[7] From an VSD perspective, the development of a "fair" algorithmic system would entail the embedding of specific values such as fairness, accountability, or transparency into the system.

18). Alternatively, research on algorithmic fairness has been characterised as "translat[ing non-discrimination] regulations mathematically into non-discrimination constraints, and develop[ing] predictive modeling algorithms that would be able to take into account those constraints, and at the same time be as accurate as possible." (Žliobaitė, 2017, p. 1061) In other words, algorithmic fairness research does not only aim at identifying and mitigating bias, but more proactively at building the value of fairness into algorithmic systems. Such research generally proceeds from some predefined fairness metrics or fairness constraints, and then aims to develop algorithmic systems that are optimised according to the proposed metrics or satisfy the specified constraints. This process can either take place (i) in the pre-process stage, where input data are modified to ensure that the outcomes of algorithmic calculations when applied to new data will be fair, (ii) during the in-process stage, where algorithms are modified or replaced to generate fair(er) output, or (iii) in the post-process stage, where the output of any model is modified to be fairer. [8] Once again, there are obvious parallels between such computational approaches and VSD's goal of "influencing the design of technology early in and throughout the design process" (Friedman and Hendry, 2019, p. 4). In both cases, the adoption of a proactive orientation is indicative of a shared commitment to progress and improvement through ethical, value-based design. It is a constructive agenda that aims at contributing to responsible innovation rather than taking a purely analytical, after-the-fact approach. As Friedman and Hendry (2019, p. 2) put it: "While empirical study and critique of existing systems is essential, [VSD] is distinctive for its design stance – envisioning, designing, and implementing technology in moral and ethical ways that enhance our futures."

5. Discussion

Despite the conceptual similarities outlined above and the fact that the VSD literature is often cited by the FAT (Fairness, Accountability, and Transparency in sociotechnical systems) community, the uptake and integration of some of VSD's core ideas in computer science remains inadequate in several important aspects.

First, concerns have been raised that the current literature on fairness in machine learning tends to focus too narrowly on how individual or societal biases can enter algorithmic systems—concentrating mostly on what Friedman and Nissenbaum refer to as "preexisting bias"—while ignoring other sources of bias such as technical bias or emergent bias. In response to this, Dobbe and co-authors (2019) have

[8] See Lepri et al. (2018), Bellamy et al. (2019), and Friedler et al. (2019) for an overview of these techniques.

stressed the need for a broader view on algorithmic bias that takes into account all the categories of Friedman and Nissenbaum's (1996) taxonomy and considers "risks beyond those pre-existing in the data" (Dobbe et al., 2019, p. 2). Thus, in order to better fulfill the analytical objective of identifying and mitigating bias in algorithmic systems, it is important that the academic machine learning community does not resort to VSD in an eclectic, piecemeal manner, but rather draws on the full breadth of the proposed frameworks.

Second, it is important to remember that concepts such as *fairness* are by no means self-explanatory or clear-cut. Verma and Rubin (2018), for instance, point out that more than twenty different notions of fairness have been proposed in AI-related research in the last few years, a lack of agreement that calls into question the very idea of operationalising fairness when seeking to design fair algorithms. Although the idea of fairness and the related concept of 'equality of opportunity' have been extensively discussed in philosophical research (see, e.g., Ryan, 2006; Hooker, 2014; Arneson, 2018), Binns (2018) has argued that most fairness measures in machine learning research tend to be undertheorized from a philosophical perspective, resulting in approaches that focus "on a narrow, static set of prescribed protected classes [...] devoid of context" (ibid.: p. 9). Last but not least, Corbett-Davies and Goel (2018) have highlighted the divergence between formalised notions of fairness and people's common understanding of fairness in everyday decision contexts. What follows from these objections is that attempts to formalise and operationalise fairness in specific ways can be contested on numerous grounds.

Unfortunately, this contestability is often disregarded or downplayed in the presentation of technical solutions, [9] even though recent years have shown a trend toward more interdisciplinary approaches that are conscious of the need to broaden the analytical scope. Proper utilisation of VSD could support such efforts as the method not only requires diligent investigations of the values at stake (see, in particular, the philosophical and technical investigations in the VSD method), but also calls for the involvement of interdisciplinary research teams that include, for example, philosophers, social scientists, or legal scholars. Of course, such interdisciplinary approaches can be challenging and resource intensive, but ethical design ultimately demands more than mechanical, recipe-based treatments of FAT requirements (see Keyes et al., 2019). Striving for truly *value-sensitive* designs implies being sensitive to the manifold meanings of values in different societal and

[9] For a detailed discussion of the concept of contestability and the importance of contestable design, see Kluttz et al., 2020.

cultural contexts and requires recognising, relating, and applying different disciplinary competences.

Finally, and on a related note, there is not only a need to expand the breadth of disciplinary perspectives, but also to widen the scope of the object of investigation itself. Simply put, instead of focusing more narrowly on fairness, accountability, and transparency in machine learning, research on algorithmic bias should also account for (a) the broader socio-technical system in which technologies are situated and (b) the different logics and orders that these algorithmic technologies produce and engender. Regarding the former, Gangadharan and Niklas (2019) have warned that the techno-centric focus on embedding fairness in algorithms, which is based on the idea that technical tweaks will suffice to prevent or avoid discriminatory outcomes, runs the danger of ignoring the wider social, political, and economic conditions in which unfairness and inequality arise. Regarding the latter, Hoffmann (2019, p. 910) reminds us that work on algorithmic bias does not only demand sustained attention to system failures but also to "the kinds of worlds being built – both explicitly and implicitly – by and through design, development, and implementation of data-intensive, algorithmically-mediated systems". What would thus be needed is greater attention to the "broader institutional, contextual, and social orders instantiated by algorithmically mediated systems and their logics of reduction and optimization" (ibid.). The FAT community has already made strides in this direction, with the *ACM FAT* Conference 2020* explicitly seeking "to sustain and further improve the high quality of computer science research in this domain, while simultaneously extending the focus to law and social sciences and humanities research" (ACM FAcct Conference, 2020). Nevertheless, we believe that a more comprehensive uptake of VSD, which has been conceptualised as an interdisciplinary approach from the very start, could support this process.

6. Concluding remarks

This paper has offered a concise review of the methodology of Value Sensitive Design and the taxonomy of biases proposed by Friedman and Nissenbaum (1996). It has shown that both VSD and the taxonomy of biases remain highly relevant for current research on bias and fairness in socio-technical systems. Despite its usefulness, however, VSD is often taken up only partially and crucial insights—e.g., regarding the conceptual underpinnings of values, the need to consider both users and non-users of a technology, [10] or the importance of interdisciplinarity—are lost.

[10] For a more detailed discussion on the need to also take non-users into account, see Wong (2019) and Wong and Simon (2020).

Consequently, it would be advisable to intensify efforts to revitalise and deepen the uptake of Value Sensitive Design in Fairness, Accountability, and Transparency (FAT) and related research. Fortunately, there is indeed a trend to expand the debates and move the discussion beyond the technical domain.

Clearly, the review of VSD and research on algorithmic bias in this paper does not fully capture the evolving debate. Moreover, it is important to note that research on biases goes well beyond the purview of VSD and computer science. Indeed, psychology and the cognitive sciences have long studied *cognitive biases* (Gigerenzer et al., 2012; Kahneman, 2011) and *implicit biases* (Holroyd et al., 2017). [11] While Friedman and Nissenbaum's notion of preexisting bias has, to some extent, accounted for implicit biases, the relationship between human cognitive biases and bias in computer systems requires further analyses. Especially in the context of automated decision-making (ADM), where *human* decisions are complemented—or even replaced—by *machine* decisions, human cognitive biases can have interesting ramifications for the design and use of ADM systems.

Firstly, cognitive biases can be *causally* related to biased automated decision-making. Cognitive limitations and biases may for instance contribute to the formation of societal stereotypes, prejudices and unwarranted preferences, or poor decision-making practices (e.g., through the defective interpretation of probabilities), which are fed into ADM systems through training data, thereby hiding while at the same time reproducing and reinforcing such biases in seemingly neutral machines.

Secondly, and conversely, ADM systems can also reduce and/or eliminate cognitive biases by accounting for and possibly correcting flaws in human reasoning (see, e.g., Savulescu and Maslen, 2015; Sunstein, 2018). In this respect, if designers and researchers of ADM systems can a) identify the sources of cognitive biases and b) counter them through specific methodological choices in designing and implementing the system, such systems can be conceived as tools to both disclose cognitive biases in human decision-making and to reduce or even prevent their negative impacts through sophisticated human-machine interaction in decision-making.

Finally, unwarranted delegation of *human* decision-making to machines can be a cognitive bias in itself, known as automation bias (Mosier et al., 1996) or automation complacency (Parasuraman and Manzei, 2010). Automation bias is characterised by the human tendency to over-trust and over-rely on allegedly neutral

[11] It should be noted that cognitive bias and implicit bias do not necessarily have the negative moral connotation as in the case of bias in VSD.

machines in that they follow wrong (or questionable) 'decisions' from the machines without seeking further corroborative or contradictory information, or even discount information from other existing sources (Skitka et al., 1999). Relatedly, automation complacency describes human operators' belief in the system's reliability, thereby causing them to pay insufficient attention to monitoring the process and to verifying the outputs of the system. Thus, recognising the dangers of automation bias and automation complacency—i.e., of overreliance on automated decision-making—brings us right back to Friedman and Nissenbaum's early warnings regarding biases in seemingly accurate, neutral, and objective computer systems, and their timely request to actively expose and counter them for better design and informed public discourse on the merits and limitations of such software tools. However, improving our tools will only bring us so far—accounting for values and countering bias also requires us to acknowledge and remedy existing inequalities and injustices in our societies and to concede that not all decision-making processes should be conducted by algorithms.

References

A.C.M.FAcct Conference. (2020). *ACM FAT* Conference 2020*. ACM FAccT Conference. https://facctconference.org/2020/

A.C.M.FAccT Conference. (2021). *ACM FaccT Conference 2021*. ACM FaccT Conference. https://facctconference.org/2021/index.html

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine Bias. In *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Arneson, R. (2018). Four Conceptions of Equal Opportunity. *The Economic Journal*, *128*(612), 152–173,. https://doi.org/10.1111/ecoj.12531

Baeza-Yates, R. (2018). Bias on the web. *Communications of the ACM*, *61*(6), 54–61. https://doi.org/10.1145/3209581

Barcoas, S., & Selbst, A. D. (2016). Big data’s disparate impact. *California Law Review*, *104*(3), 671–732. https://doi.org/10.15779/Z38BG31

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., & Zhang, Y. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, *63*(4/5), 1–15. https://doi.org/10.1147/JRD.2019.2942287

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. *Proceedings of Machine Learning Research*, *81*, 149–159. http://proceedings.mlr.press/v81/binns18a.html

Brey, P. (2010). Values in technology and disclosive computer ethics. In L. Floridi (Ed.), *The Cambridge handbook of information and computer ethics* (pp. 41–58). Cambridge Univerity Press. https://doi.org/10.1017/CBO9780511845239.004

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of Machine Learning Research*, *81*, 1–15. http://proceedings.mlr.press/v81/buolamwini18a.html

Corbett-Davies, S., & Goel, S. (2018). The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. *ArXiv*. https://arxiv.org/abs/1808.00023

Cramer, H., Garcia-Gathright, J., Springer, A., & Reddy, S. (2018). Assessing and Addressing Algorithmic Bias in Practice. *Interactions*, *25*(6), 58–63. https://doi.org/10.1145/3278156

Custers, B., Calders, T., Schermer, B., & Zarsky, T. (Eds.). (2013). *Discrimination and Privacy in the Information Society Data Mining and Profiling in Large Databases*. Springer. https://doi.org/10.1007/978-3-642-30487-3

Danks, D., & London, A. J. (2017). Algorithmic bias in autonomous systems. *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 4691–4697. https://dl.acm.org/doi/10.5555/3171837.3171944

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. In *Reuters*. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Davis, J., & Nathan, L. P. (2014). Value sensitive design: Applications, adaptions, and critique. In J. Hoven, P. E. Vermaas, & I. Poel (Eds.), *Handbook of Ethics, Values, and Technological Design* (pp. 1–26). Springer. https://doi.org/10.1007/978-94-007-6970-0_3

Dieterich, W., Mendoza, C., & Brennan, T. (2016). *COMPAS Risk Scales: Demonstrating accuracy equity and predictive parity* [Technical report]. Northpointe. https://www.documentcloud.org/documents/2998391-ProPublica-Commentary-Final-070616.html

Dobbe, R., Dean, S., Gilbert, T., & Kohli, N. (2018). *A Broader View on Bias in Automated Decision-Making: Reflecting on Epistemology and Dynamics*. 2018 Workshop on Fairness, Accountability and Transparency in Machine Learning. http://arxiv.org/abs/1807.00553

F.A.T./M.L. (2018). *Fairness, Accountability, and Transparency in Machine Learning*. https://www.fatml.org/

Flanagan, M., Howe, D. C., & Nissenbaum, H. (2008). Embodying Values in Technology: Theory and Practice. In J. Hoven & J. Weckert (Eds.), *Information Technology and Moral Philosophy* (pp. 322–353). Cambridge University Press. https://doi.org/10.1017/CBO9780511498725.017

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., & Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning. *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT*, 19*, 329–338. https://doi.org/10.1145/3287560.3287589

Friedman, B. (Ed.). (1997). *Human Values and the Design of Computer Technology*. Cambridge University Press.

Friedman, B., & Hendry, D. G. (2019). *Value Sensitive Design: Shaping Technology with Moral Imagination*. MIT Press.

Friedman, B., Kahn, P. H., & Borning, A. (2006). Value Sensitive Design and Information Systems. In P. Zhang & D. Galetta (Eds.), *Human-Computer Interaction in Management Information Systems: Foundations* (pp. 348–372). M.E. Sharpe.

Friedman, B., & Nissenbaum, N. (1996). Bias in Computer Systems. *ACM Transactions on Information Systems*, *14*(3), 330–347. https://doi.org/10.1145/230538.230561

Gangadharan, S. P., & Niklas, J. (2019). Decentering technology in discourse on discrimination. *Information, Communication & Society*, *22*(7), 882–899. https://doi.org/10.1080/1369118X.2019.1593484

Gigerenzer, G., Fiedler, K., & H, O. (2012). Rethinking Cognitive Biases as Environmental Consequences. In P. M. Todd & G. Gigerenzer (Eds.), *Ecological Rationality. Intelligence in the World* (pp. 80–110). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195315448.001.0001

Hoffmann, A. L. (2019). Where fairness fails: Data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society*, *22*(7), 900–915. https://doi.org/10.1080/1369118X.2019.1573912

Holroyd, J., Scaife, R., & Stafford, T. (2017). What is Implicit Bias? *Philosophy Compass*, *12, e12437*. https://doi.org/10.1111/phc3.12437

Hooker, B. (2014). Utilitarianism and fairness. In B. Eggleston & D. Miller (Eds.), *Cambridge Companion to Utilitarianism* (pp. 280–302). Cambridge University Press. https://doi.org/10.1017/CCO9781139096737.015

Kahneman, D. (2011). *Think, Fast and Slow*. Allen Lane.

Kearns, M., & Roth, A. (2020). *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press.

Keyes, O., Hutson, J., & Durbin, M. (2019). A Mulching Proposal: Analysing and Improving an Algorithmic System for Turning the Elderly into High-Nutrient Slurry. *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–11. https://doi.org/10.1145/3290607.3310433

Kluttz, D. N., Kohli, N., & Mulligan, D. K. (2020). Shaping Our Tools: Contestability as a Means to Promote Responsible Algorithmic Decision Making in the Professions. In K. Werbach (Ed.), *After the Digital Tornado: Networks, Algorithms, Humanity* (pp. 137–152). Cambridge University Press. https://www.cambridge.org/core/books/after-the-digital-tornado/shaping-our-tools-contestability-as-a-means-to-promote-responsible-algorithmic-decision-making-in-the-professions/311281626ECA50F156A1DDAE7A02CECB

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology*, *31*, 611–627. https://doi.org/10.1007/s13347-017-0279-x

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A Survey on Bias and Fairness in Machine Learning. *arXiv*. https://arxiv.org/abs/1908.09635.

Mosier, K. L., & Skitka, L. J. (1996). Human Decision Makers and Automated Decisions Aids: Made for Each Other? In R. Parasuraman & M. Mouloua (Eds.), *Automation and Human Performance: Theory and Applications* (pp. 201–220). NJ. Lawrence Erlbaum Associates.

Olteanu, A., Castillo, C., Diaz, F., & Kıcıman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, *2*. https://doi.org/10.3389/fdata.2019.00013

O'Neil, C. (2016). *Weapons of Math Destruction. How Big Data Increases Inequality and Threatens Democracy*. Crown Publishers.

Oudshoorn, N., & Pinch, T. (2005). How Users and Non-users Matter. In N. Oudshoorn & T. Pinch (Eds.), *How Users Matter: The Co-Construction of Users and Technology* (pp. 1–28). MIT Press.

Parasuraman, R., & Manzey, D. H. (2010). Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors*, *52*(3), 381–410. https://doi.org/10.1177/0018720810376055

Ryan, A. (2006). Fairness and Philosophy. *Social Research*, *73*(2), 597–606. https://www.jstor.org/stable/40971838

Savulescu, J., & Maslen, H. (2015). Moral enhancement and artificial intelligence: Moral AI? In J. Romportl, E. Zackova, & J. Kelemen (Eds.), *Beyond Artificial Intelligence: The Disappearing Human-Machine Divide* (pp. 79–95). Springer. https://doi.org/10.1007/978-3-319-09668-1_6

Simon, J. (2017). Value-sensitive design and responsible research and innovation. In S. O. Hansson (Ed.), *The ethics of technology methods and approaches* (pp. 219–235). Rowman & Littlefield.

Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does Automation Bias Decision-Making? *International Journal of Human–Computer Studies*, *51*, 991–1006. https://doi.org/10.1006/ijhc.1999.0252

Sunstein, C. R. (2019). Algorithms, correcting biases. *Social Research*, *86*(2), 499–511. https://www.muse.jhu.edu/article/732187

Verma, S., & Rubin, J. (2018). Fairness definitions explained. *Proceedings of the International Workshop on Software Fairness*, 1–7. https://doi.org/10.1145/3194770.3194776

Winner, L. (1980). Do Artifacts Have Politics? *Daedalus*, *109*(1), 121 136. https://www.jstor.org/stable/20024652

Wong, P. H. (2019). Democratizing Algorithmic Fairness. *Philosophy & Technology*, *33*, 225–244. https://doi.org/10.1007/s13347-019-00355-w

Wong, P. H., & Simon, J. (2020). Thinking About 'Ethics' in the Ethics of AI. *IDEES*, *48*. https://revistaidees.cat/en/thinking-about-ethics-in-the-ethics-of-ai/

Wyatt, S. (2005). Non-Users Also Matter: The Construction of Users and Non-Users of the Internet. In N. Oudshoorn & T. Pinch (Eds.), *How Users Matter: The Co-Construction of Users and Technology* (pp. 67–79). MIT Press.

Žliobaitė, I. (2017). Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, *31*(4), 1060–1089. https://doi.org/10.1007/s10618-017-0506-1

Published by

Alexander von Humboldt
INSTITUTE FOR INTERNET
AND SOCIETY

in cooperation with

CREATe

centre internet et société

UOC R&I IN3 Internet interdisciplinary Institute
Universitat Oberta de Catalunya