Who is the fairest of them all? Public attitudes and expectations regarding automated decision-making

Helberger, N.; Araujo, T.; de Vreese, C.H.

DOI
10.1016/j.clsr.2020.105456

Publication date
2020

Document Version
Final published version

Published in
Computer Law & Security Review

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/CLSR

Computer Law & Security Review

Who is the fairest of them all? Public attitudes and expectations regarding automated decision-making

Natali Helberger[a,*], *Theo Araujo*[b], *Claes H. de Vreese*[b]

[a] *Institute for Information Law (IViR), University of Amsterdam, Amsterdam, the Netherlands*
[b] *Amsterdam School of Communication Research (ASCoR), University of Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam, the Netherlands*

ARTICLE INFO

Keywords:
Algorithmic fairness
Automated-decision Making
Public perceptions

ABSTRACT

The ongoing substitution of human decision makers by automated decision-making (ADM) systems in a whole range of areas raises the question of whether and, if so, under which conditions ADM is acceptable and fair. So far, this debate has been primarily led by academics, civil society, technology developers and members of the expert groups tasked to develop ethical guidelines for ADM. Ultimately, however, ADM affects citizens, who will live with, act upon and ultimately have to accept the authority of ADM systems.

The paper aims to contribute to this larger debate by providing deeper insights into the question of whether, and if so, why and under which conditions, citizens are inclined to accept ADM as fair. The results of a survey ($N = 958$) with a representative sample of the Dutch adult population, show that most respondents assume that AI-driven ADM systems are fairer than human decision-makers.

A more nuanced view emerges from an analysis of the responses, with emotions, expectations about AI being data- and calculation-driven, as well as the role of the programmer – among other dimensions – being cited as reasons for (un)fairness by AI or humans. Individual characteristics such as age and education level influenced not only perceptions about AI fairness, but also the reasons provided for such perceptions. The paper concludes with a normative assessment of the findings and suggestions for the future debate and research.

1. Introduction

AI is moving into the very essence of what constitutes us as a democratic society: the way we make decisions. Automated decision-making (ADM) systems are replacing human decision-makers in a whole range of areas, from governments and court rooms, to HR departments, financial institutions, media and politics. The ongoing integration of ADM into the bodies and institutions that take decisions in our society has triggered an intense debate among academics, policymakers and civil society about the conditions under which ADM is acceptable or unacceptable, the opportunities and, perhaps even more so, the risks that ADM poses, and also how we can en-

* Corresponding author. Natali Helberger Institute for Information Law (IViR), University of Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam, the Netherlands.
E-mail addresses: n.helberger@uva.nl (N. Helberger), t.b.araujo@uva.nl (T. Araujo), c.h.devreese@uva.nl (C.H. de Vreese).

https://doi.org/10.1016/j.clsr.2020.105456

sure that ADM systems respect the fundamental values that characterize our society.[1]

One such value is *fairness*. A notoriously difficult concept to define, automation and the algorithmic turn force us to revisit fairness and its meaning in the context of ADM. Different disciplines, from computer science to political philosophy and law, have begun to reconceptualize fairness to determine the potential but also the limits of integrating ADM into society. The importance of algorithmic fairness has been reemphasized in a range of high-level policy documents and ethical guidelines,[2] and the F from Fairness is a constitutive part of FACT and FAIR in the realm of Responsible Data Science, and more recently, the Responsible AI movement. Ultimately, the ongoing discussion about algorithmic fairness is a deep debate about whether we, as a society, are willing to accept ADM as a legitimate form of decision-making and, if so, under which conditions.

Thus far, this societal debate has been primarily led and informed by academics, civil society, technology developers and members of the many expert groups tasked to develop ethical guidelines for ADM, a fact that is not unreasonable given the immense complexity of the issue. Ultimately, however, ADM affects citizens. They will have to live with, act on and accept the authority of ADM systems. Any claim to the legitimacy of automated decisions will have to be recognized by them and algorithmic fairness accepted as justice. Citizens' perceptions and assumptions regarding fairness in ADM in contemporary societies, in general, and their expectations regarding ADM, in particular, remain critically understudied. The goal of this paper is therefore threefold: (1) to gain an initial understanding of citizens' intuitive perceptions of ADM fairness, and why and under which conditions people are inclined to perceive and accept ADM as fair or unfair in comparison to human decision-makers, (2) to ascertain to what extent the principles and concerns that citizens consider decisive in their judgement of ADM as fair or not corresponds with conceptions of fairness currently found in the academic literature, or whether we may have overlooked certain critical aspects, and (3) explore the extent to which individual characteristics such as age, gender and education levels influence such perceptions. In summary, we propose the following research question: *How do citizens perceive the potential fairness of ADM systems in comparison to human decision-makers; what principles lie behind their evaluations of fairness; and to what extent do individual characteristics influence such evaluations?*

In addressing this question, this article contributes to the current literature on fairness in ADM by adding empirically grounded insights into the perceptions of citizens as potential subjects of ADM. This includes their articulations of aspects related to procedural fairness in particular. Much of the current literature still focuses on aspects of substantive fairness (such as the lack of bias or respect for individual rights such as the right to privacy or non-discrimination) and the process of ADM itself. Because of the focus on people's perceptions, we are able to provide a more nuanced understanding of fairness, showing that people appreciate different aspects of fairness in ADM and value both modes of decision-making for different reasons, or in combination. Finally, the article seeks to broaden the debate around fairness in ADM by demonstrating that fairness does not automatically translate into justice, and that when implementing ADM in professional decision-making processes care must be taken not only regarding the fairness of the decisions themselves, but that automated judgements must be rendered in a way that respects human dignity and the potential need for human interaction. In other words, we argue that there is also an emotive or relational dimension of fairness that needs to be considered when implementing ADM in any decision-making process. For the purpose of this research, we approached ADM broadly in the sense of decision making in a professional capacity without referring to a particular sector. However, when elaborating on the qualities that could be expected in a (human) professional decision maker we borrowed from the literature on decision making in the judicial sector, as this is one of the proto-types of professional decision making, and an area where the question of what characterizes fairness in decision makers and decision making has been subject to extensive research, exactly because of the linkage between justice and fairness.

In the following, we will briefly outline the current discussion in the academic literature, before describing the methods and empirical findings of our research. We will conclude with a discussion of our findings and reflections for further research.

2. Fairness in decision-making

2.1. *Decision-making, justice and society*

Disagreement is a necessary characteristic of pluralist democratic societies, as is the existence of institutions and procedures to resolve them. This is done on the basis of rules and legal standards that reflect the central values and perceptions of justice and fairness in a society.[3] According to Bellamy, 'democracy embodies the "right to have rights" of citizens—it offers the mechanism through which their different views on justice and the good are treated with equal respect and their interests and ideals may be shown equal concern'. [4] Moreover, a central element of Rawls' theory of justice, which has strongly influenced the way decision-making processes are organized today, is that it is imperative on society that its citizens agree on certain 'rules of the game'. These rules must order civic life and ensure that citizens can live peacefully

[1] AlgorithmWatch, 'Automating Society: Taking Stock of Automated Decision-Making in the EU' (AlgorithmWatch, January 2019), https://algorithmwatch.org/wp-content/uploads/2019/01/Automating_Society_Report_2019.pdf; Michael Veale and Irina Brass, 'Administration by Algorithm? Public Management Meets Public Sector Machine Learning', in *Algorithmic Regulation*, ed. Karen Yeung and Martin Lodge (Oxford University Press, 2019); Meredith Whittaker et al., 'AI Now Report 2018' (AI Now, December 2008), https://ainowinstitute.org/AI_Now_2018_Report.pdf.

[2] For example, Ethics Guidelines of the High Level Expert Group, OECD Guidelines, AI for People Guidelines etc.

[3] John Rawls, 'Justice as Fairness', *The Philosophical Review* 67, no. 2 (1958): 175, 10.2307/2182612.

[4] Richard Bellamy, 'The Democratic Qualities of Courts: A Critical Analysis of Three Arguments', *Representation* 49, no. 3 (September 1, 2013): 333–46, 338 10.1080/00344893.2013.830485.

alongside each other,[5] or as Ginnis and Pearce have phrased it: 'Law is an information technology—a code that regulates social life'.[6] No less important than the rules themselves, is the role of decision-makers within a democratic society. Professional decision-makers, such as judges, doctors, politicians, journalists or government officials, have a central position.

What is the future of human decision-makers in a world of AI? AI and machine learning can, on the one hand, help to make rules and their application smarter in a range of areas. In law, for example, it can assist in information retrieval, such as the discovery of relevant case law, or the processing of large amounts of legal data, the automated generation of legal texts, the offering of new legal or counselling services and ways of interacting with citizens, or predictions of risks and case outcomes. Such automation of tasks can improve effectiveness in case handling.[7] Prins has even argued that decision-makers, and court rooms in particular, have a duty to embrace digitization.[8]

On the other hand, decision-making is also a deeply social and societal task. Once again, in law, for example, it involves individual judgement and social expertise, but also the ability to judge and understand human behaviour, the use of social intelligence, as well as the ability to engage with others, to listen and to explain issues in a way that brings the various parties involved to accept a judicial decision.[9] As Morison and Harkens have pointed out, judges operate in the context of 'a social process, mediated by judges and conditioned by a whole range of broader professional, social and economic factors within the overall legal system'.[10] Sourdin and Cornes have even claimed that '[a]n element of litigants' respect for judicial judgement, and the social legitimacy of the judiciary more broadly, must come from, we think, the fact that it is rendered by a fellow human being'.[11]

The centrality of decision-making institutions for the functioning of a democratic society necessarily imposes requirements on and social expectations of those making the decisions.[12] Typically, especially in contemporary society, decision-makers are trained experts with a sound understanding of the substantive as well as the procedural rules that guarantee fair decision-making.[13] Being able to internalize and follow rules is arguably one of the strong sides of ADM systems – which are essentially rule-based – providing it is possible to translate legal rules into machine rules.[14] In addition to being experts who base their activities on predefined written and unwritten rules, good decision-makers, it has been argued, generally also possess a number of intrinsic qualities. Take the example of judges, who are perhaps the most prototypical examples of societal decision-makers: based on an investigation of ethical codes for judges, literature in the field of political philosophy and interviews with practitioners, Domselaar developed a 'six-pack of judicial virtues', which 'are indispensable for realizing moral quality in adjudication: judicial perception, judicial courage, judicial temperance, judicial justice, judicial impartiality and judicial independency'.[15] Two of these judicial virtues (or sets of virtues) are particularly relevant to our context.

The first set is *judicial perception*, which refers to the ability of judges to investigate and observe the salient facts of an individual case, as the necessary factual basis for being able to apply the law. These facts are external to the legal rules but central to the concept of justice and the ability to make a fair decision. Many legal rules build on abstract concepts or 'instructions' (e.g. the prohibition against discrimination) that judges need to apply using their reason. As Blum has put it: 'It is not the rule but some other moral capacity of the agent which tells her that the particular situation she faces falls under a given rule', and as such requires a distinct moral capacity of the decision-maker.[16] Following this train of thought, justice not only concerns the application of rules but also the application of rules to a specific case. Accordingly, it could be argued that, as the facts of the case are an intrinsic part of a just decision, so is the human decision-maker, who is able to apply moral judgement to the case, which is an important element of a just decision. However, it might also be argued, to the contrary, that taking into account the facts of an individual case opens the door to subjective assessments and thus

[5] Rawls, 'Justice as Fairness', 171.

[6] John O. McGinnis and Russell G. Pearce, 'The Great Disruption: How Machine Intelligence Will Transform the Role of Lawyers in the Delivery of Legal Services', *Fordham Law Review* 82 (2014): 3041.

[7] Benjamin Alarie, Anthony Niblett, and Albert H Yoon, 'How Artificial Intelligence Will Affect the Practice of Law', *University of Toronto Law Journal* 68, no. supplement 1 (January 2018): 106–24, 10.3138/utlj.2017-0052; Nikolaos Aletras et al., 'Predicting Judicial Decisions of the European Court of Human Rights: A Natural Language Processing Perspective', *PeerJ Computer Science* 2 (2016): e93; McGinnis and Pearce, 'The Great Disruption: How Machine Intelligence Will Transform the Role of Lawyers in the Delivery of Legal Services'; Tania Sourdin and Richard Cornes, 'Do Judges Need to Be Human? The Implications of Technology for Responsive Judging', in *The Responsive Judge* (Springer, 2018), 87–119.

[8] Corien Prins, 'Digital Justice', *Computer Law & Security Review* 34, no. 4 (August 1, 2018): 920–23, 10.1016/j.clsr.2018.05.024.

[9] Lili Barna, Dorottya Juhász, and Soma Márok, 'What Makes a Good Judge' (Budapest: European Judicial Training Network Themis Competition, 2017), http://www.ejtn.eu/Documents/Team%20HU%20semi%20final%20D.pdf; Adam Feldman and Elli Menounou, 'What Motivates the Justices: Utilizing Automated Text Classification to Determine Supreme Court Justices' Preferences', in *Annual Meeting of the Southern Political Science Association (SPSA)* (New Orleans, LA, 2015); Sourdin and Cornes, 'Do Judges Need to Be Human? The Implications of Technology for Responsive Judging'.

[10] John Morison and Adam Harkens, 'Re-Engineering Justice? Robot Judges, Computerised Courts and (Semi) Automated Legal Decision-Making', *Legal Studies*, 2019, 19, 10.1017/lst.2019.5.

[11] Sourdin and Cornes, 'Do Judges Need to Be Human? The Implications of Technology for Responsive Judging', 98.

[12] R. Cranston, 'What Do Courts Do', *Civil Justice Q* 5 (1986): 124; Sourdin and Cornes, 'Do Judges Need to Be Human? The Implications of Technology for Responsive Judging'.

[13] Extensive legal training is required to become a judge.

[14] The question of whether this is possible or not is subject to a whole body of research in legal informatics, which is beyond the scope of this contribution, see instead Morison and Harkens, 'Re-Engineering Justice?', 15.

[15] Iris van Domselaar, 'Moral Quality in Adjudication: On Judicial Virtues and Civic Friendship', *Netherlands Journal of Legal Philosophy*, no. 1 (2015): 27.

[16] Lawrence Blum, 'Moral Perception and Particularity', *Ethics* 101, no. 4 (July 1, 1991): 708–9, 10.1086/293340.

biases, personal preferences and values, or even simply to being influenced by the time of the day, or what a judge may have eaten the day before.[17]

Human decisions can have characteristics that may not readily accord with the idea of a just decision-maker, hence the importance of the other set of values in Domselaar's six-pack: *judicial temperance, impartiality* and *independence*. Common to these three virtues is that they elevate the human judge above a normal human being, suggesting that they are able to operate free from external influences or attempts at manipulation (such as bribes and financial advantages but also politics and lobbying), as well as the judge's own internal passions, biases and subjectivity. ADM systems are often credited with such characteristics as impartiality, objectivity and lack of emotions, which could thus be an argument in favour of considering machines the better and fairer decision-makers.[18]

It is not self-evident, however, that typical human traits, such as the ability to have emotions and affects, necessarily stand in the way of fairer decision-making. As Domselaar argued, emotions could actually be a necessary ingredient of fairer decision-making because they are means of cognitive perception.[19] Sourdin and Cortes found that emotional understanding can even be decisive in the outcome of cases.[20] In a comparative, multi-method study of the ideal characteristics of judges, more inherently human characteristics, such as charisma (in the sense of the ability to inspire and convince) and empathy (the ability to share and understand someone else's feelings), were mentioned as critical to the ability of judges to render judgements that citizens were willing to accept.[21] Thus, the question is to what extent citizens perceive such typical human traits as an advantage or disadvantage of human decision-makers vis-à-vis machines. While we have discussed the example of judges and perceptions of the professional and human traits that judges require in order to qualify as decision-makers whose decisions are authoritative, it stands to reason that many of these traits, such as impartiality, objectivity, the ability to consider the characteristics of an individual case and the absence (or presence) of emotions such as empathy, will also play a role in shaping expectations of other, judge-like decision-makers, such as public servants and doctors.

2.2. *Fairness as an intuitive and malleable concept*

Closely connected to the question of the qualities of a just and fair decision-maker is the interpretation of fairness as one element, and perhaps the most intuitive, critical element of justice itself (in the broadest sense of just decision-making in a society). What aspects of ADM lead someone to consider it fair or unfair? Do people judge human and AI decision-makers according to the same conception of fairness? Fairness as an ordering concept in liberal societies is as much intuitive as it is a malleable concept. Often associated with justice and the legal field, fairness is an issue across all areas of society. We can feel fairly or unfairly treated, for example, by decisions of a judge, doctor or government official.

There is no clear definition of fairness, and perceptions of fairness may differ between different cultures, jurisdictions, contexts and sectors, and even individuals. The lack of a concrete definition can be explained by the fact that fairness is not so much a predefined principle but far more a judgement call, or what Angelopoulos called a 'rational discourse'.[22] The essence of this rational discourse is the balancing of interests or values. As such, fairness is deeply contextual. Moreover, as a part of their training, professional decision-makers, such as judges, doctors, politicians, journalists or government officials, are taught to make well-founded 'fair' decisions taking into account *substantive fairness* safeguards (e.g. the right to non-discrimination, privacy or other fundamental rights), as well as a well-developed set of rules that must guarantee *procedural fairness* (e.g. transparency, due process). With the arrival of automated decision-makers, the extent to which they might replace or augment human decision-makers and the lack of clear models or metrics to determine fairness have become pertinent and of interest to a growing group of researchers.

2.2.1. *ADM and substantive fairness*

Perhaps the most common conceptualization of fairness in ADM is related to the idea of *fairness as non-discrimination and differential treatment*. Rawls's conception of fairness is deeply rooted in ideals of liberty and equality. This may readily explain why fairness is often conceptualized as, or even equated to, lack of bias or non-discrimination; that is, in terms of principles of substantive fairness. This seems to be particularly true for the computer science literature on fairness-aware algorithms. Here, one influential definition of fairness conceptualizes it as a form of equal treatment, thus concerning parity or distance metrics, in the sense that 'similarly situated people are given similar treatment – that is, a fair process will give similar participants a similar probability of receiving each possible outcome'.[23] Another important way of looking at algorithmic fairness focuses on equal opportunities,[24] or

[17] Tania Sourdin, *Alternative Dispute Resolution*, 5th ed. (Pyrmont: Thomson Reuters, 2016).

[18] Daniel Martin Katz, 'Quantitative Legal Prediction-or-How I Learned to Stop Worrying and Start Preparing for the Data-Driven Future of the Legal Services Industry', *Emory LJ* 62 (2012): 909.

[19] Van Domselaar, 'Moral Quality in Adjudication', 33.

[20] Sourdin and Cornes, 'Do Judges Need to Be Human? The Implications of Technology for Responsive Judging', 97.

[21] Barna, Juhász, and Márok, 'What Makes a Good Judge', 17; Morison and Harkens, 'Re-Engineering Justice?'

[22] Christina Angelopoulos, 'MTE v Hungary : A New ECtHR judgement on Intermediary Liability and Freedom of Expression', *Journal of Intellectual Property Law & Practice* 11, no. 8 (August 1, 2016): 582–84, 10.1093/jiplp/jpw081.

[23] Reuben Binns, 'Data Protection Impact Assessments: A Meta-Regulatory Approach', *International Data Privacy Law* 7, no. 1 (2017): 22–35; Joshua A. Kroll et al., 'Accountable Algorithms', *University of Pennsylvania Law Review*, no. 3 (2017 2016): 685, as well as the influential article by Cynthia Dwork et al., 'Fairness through Awareness', in Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ACM, 2012), 214–26.

[24] Hoda Heidari et al., 'A Moral Framework for Understanding Fair ML through Economic Models of Equality of Opportunity', in *Proceedings of the Conference on Fairness, Accountability, and Transparency – FAT* '19* (the Conference, Atlanta, GA, USA: ACM Press, 2019), 181–90, 10.1145/3287560.3287584.

the disparate impact that algorithmic processes may have.[25] This perspective is again related to the broader issue of non-discrimination,[26] and it is often used interchangeably with the notion of bias in computer systems.

Computer science's focus on framing substantive fairness in terms of non-discrimination reflects the legal conceptualization of fairness in such terms. This means fairness is both a fundamental right (Art. 14 ECHR) and guaranteed by various non-discrimination laws[27] that prohibit the treatment of people or groups differently based on various sensitive characteristics, including race, gender, nationality, sexual preference, political or religious convictions, etc.

The conceptualization of fairness as non-discrimination is related to another important conceptualization of fairness, namely *distributive fairness*, or the question of how benefits and burdens in a society can be distributed fairly.[28] Distributive justice is central, for example, to the Rawlsian account of fairness and his two principles of equal liberty and difference.[29] Matters of distributive justice are also central to concerns about fairness in ADM.[30] For example, ADM systems can be used for predictive credit, risk or price scoring, with the result that certain parts of the population will receive higher prices or fees than others – a practice that is intuitively considered by many consumers to be unfair.[31] New digital inequalities between data haves and have-nots are possible, but also between those who are able to play the ADM system and those who are not.[32]

Beyond these two conceptualizations of fairness, the term can have many other meanings.[33] Fair data processing is, for example, a key notion in the GDPR, including its provisions on ADM.[34] Fairness-related provisions in data protection law concern, for example, a fair balance between the interests of users and data controllers (e.g. in the form of consent requirements, general principles such as data minimalization and purpose limitation, or by giving users concrete rights with respect to their data).[35] Fairness as a matter of balancing rights and interests in asymmetric relationships is also an important operationalization in consumer law. Examples include the control of contractual provisions that unfairly disadvantage consumers, but also the exploitation of information asymmetries with the goal of misleading or otherwise unfairly manipulating consumers into certain economic behaviour.[36]

2.2.2. *Procedural fairness*

Less prominent thus far in the debate about algorithmic fairness are arguments concerning procedural justice, which is central to many legal accounts of fairness. To return to Rawls, '[a] practice is just or fair, then, when it satisfies the principles which those who participate in it could propose to one another for mutual acceptance under the aforementioned circumstances'.[37] Traditional decision-making by courts (and also insurance companies, doctors, teachers, journalists and others) is governed by extensive and stringent rules that dictate how decisions must be taken in order to be considered and accepted as fair. Elements of procedural fairness or due process include transparency and the kind of information that citizens have about decisions, as well as questions of algorithmic accountability and the allocation of responsibilities, the checks and balances in place and the ability of people to challenge decisions.

Procedural fairness is important for the given context, as it essentially concerns the conditions under which automated decisions will be acceptable and perceived as fair. A growing body of literature is concerned with precisely this aspect of procedural fairness, primarily from a legal/regulatory perspec-

[25] Solon Barocas and Andrew D. Selbst, 'Big Data's Disparate Impact', *Calif. L. Rev.* 104 (2016): 671.

[26] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy, 'Techniques for Discrimination-Free Predictive Models', in *Discrimination and Privacy in the Information Society* (Springer, 2013), 223–39.

[27] Council Directive 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation, OJ L 303, 2.12.2000, pp. 16–22; Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin, OJ L 180, 19.7.2000, pp. 22–26; Council Directive 2004/113/EC of 13 December 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services, OJ L 373, 21.12.2004, pp. 37–43; Directive 2006/54/EC of the European Parliament and of the Council of 5 July 2006 on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation (recast), OJ L 204, 26.7.2006, pp. 23–36; *Proposal* for a Council Directive on implementing the principle of equal treatment between persons irrespective of religion or belief, disability, age or sexual orientation, COM/2008/0426final.

[28] Serena Olsaretti, 'The Idea of Distributive Justice', in *The Oxford Handbook of Distributive Justice*, ed. Serena Olsaretti (Oxford University Press, 2018).

[29] '[F]irst, each person participating in a practice, or affected by it, has an equal right to the most extensive liberty compatible with a like liberty for all; and second, inequalities are arbitrary unless it is reasonable to expect that they will work out for everyone's advantage, and provided the positions and offices to which they attach, or from which they may be gained, are open to all', Rawls 'Justice as Fairness'.

[30] Danielle Keats Citron, 'Technological Due Process', *Washington University Law Review*, no. 6 (2008 2007): 1249–1314; Tal Zarsky, 'The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making', *Science, Technology, & Human Values* 41, no. 1 (January 1, 2016): 118–32, 10.1177/0162243915605575.

[31] Joost Poort and Frederik J. Zuiderveen Borgesius, 'Does Everyone Have a Price? Understanding People's Attitude towards Online and Offline Price DiscriminationDoes Everyone Have a Price? Understanding People's Attitude towards Online and Offline Price Discrimination', *Internet Policy Review*, 2019, 10.14763/2019.1.1383.

[32] Zarsky, 'The Trouble with Algorithmic Decisions', 125.

[33] Zarsky, 'The Trouble with Algorithmic Decisions'.

[34] Michael Butterworth, 'The ICO and Artificial Intelligence: The Role of Fairness in the GDPR Framework', *Computer Law & Security Review* 34, no. 2 (April 1, 2018): 257–68, 10.1016/j.clsr.2018.01.004; Damian Clifford and Jef Ausloos, 'Data Protection and the Role of Fairness', *Yearbook of European Law* 37 (January 1, 2018): 130–87, 10.1093/yel/yey004.

[35] Clifford and Ausloos, 'Data Protection and the Role of Fairness'. Article 5(1)(a) GDPR provides that personal data must be 'processed lawfully, fairly and in a transparent manner'.

[36] Josse Gerard Klijnsma, *Contract Law as Fairness: A Rawlsian Perspective on the Position of SMEs in European Contract Law* (Universiteit van Amsterdam, 2014); Malek Radeideh, *The Principle of Fair Trading in EC Law: Information and Consumer Choice in the Internal Market* (Rijksuniversiteit Groningen, 2004).

[37] Rawls, 'Justice as Fairness', 178.

tive.[38] Discussions of procedural fairness can also be found in the computer science literature, although the focus there is often on the aspects of formalizing algorithmic transparency and explainability, or auditing and monitoring algorithms or tools for procedural regularity.[39] In addition to concerns about how ADM relates to existing safeguards, such as in data protection law or procedural law, others point to new challenges, for example the use of personal data to draw inferences that are beyond the control of an individual.[40]

The importance of aspects of procedural fairness is further emphasized by the growing insight that in order to understand ADM and its implications for justice and fairness in society, it is not sufficient to consider the workings of algorithms separately, but to understand them as part of a broader socio-technical system,[41] similar to the notion of the legal system as a social process.[42] In other words, decision-making, including ADM, is not so much a process but rather a system that has a multitude of actors, all with their own motives, incentives and ambitions for power. The safeguards of procedural fairness have an important role in defining the rules of engagement, restricting decision-making power and ensuring access to justice for all.

2.3. *Perceptions of ADM fairness*

While discussions about fairness in any kind of decision-making often focus on the role of the decision-maker, the process or data used to reach such decisions, or the consequences of these decisions for individuals and society as a whole, it is also important to understand the perceptions that *individuals* have about ADM systems, and their expectations and assumptions regarding the *fairness* of such systems. Gaining this knowledge is critical, as pre-existing attitudes, assumptions and expectations about automation or ADM systems can influence the extent to which trust is built. Pre-existing attitudes and expectations also shape subsequent usage decisions,[43] and as has been argued, upfront assumptions about what 'algorithms are capable of and their comparison with human decision makers play important roles in people's judgements of trustworthiness and fairness, as well as their emotional responses'.[44] An important aspect underlying much of this discussion is the role of trust.[45] In summary, as Lee and See have argued, 'people tend to rely on automation they trust and tend to reject automation they do not. By guiding reliance, trust helps overcome the cognitive complexity people face in managing increasingly sophisticated automation'.[46] Nonetheless, evidence from previous research is mixed.

On the one hand, the emerging literature on *algorithmic appreciation*[47] has introduced the idea that in several instances individuals may prefer the decisions or recommendations made by algorithms over those made by other humans. This stems from a general tendency to believe that statistical or system-driven processes may outperform humans in rational decisions.[48] This may be further reinforced by perceptions of computers as being autonomous sources of information,[49] overlooking the role of the programmer. Moreover, in line with the notion of the *machine heuristic*,[50] the existing research has found that people – and some individuals more than others – tend to be more willing to disclose personal information to machines than to humans, in the belief that machines are unbiased.[51]

On the other hand, surveys have found general concerns about and a reluctance to accept ADM by algorithms (e.g. by the Pew Research Center),[52] as well as the emergence of *algo-*

[38] Binns, 'Data Protection Impact Assessments: A Meta-Regulatory Approach'; Citron, 'Technological Due Process'; Bryce Goodman and Seth Flaxman, 'European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation"', *AI Magazine* 38, no. 3 (2017): 50–57; Marion Oswald, 'Algorithm-Assisted Decision-Making in the Public Sector: Framing the Issues Using Administrative Law Rules Governing Discretionary Power', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376, no. 2128 (2018): 20170359.

[39] See e.g. the distinction in the overview by Kroll et al., 'Accountable Algorithms'.

[40] Sandra Wachter and Brent Mittelstadt. 'A Right to Reasonable Inferences', *Columbia Business Law Review*, 2019 (2), 494–620.

[41] Andrew D. Selbst et al., 'Fairness and Abstraction in Sociotechnical Systems', in *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19* (the Conference, Atlanta, GA, USA: ACM Press, 2019), 59–68, 10.1145/3287560.3287598; Linnet Taylor, 'What Is Data Justice? The Case for Connecting Digital Rights and Freedoms Globally', *Big Data & Society* 4, no. 2 (December 1, 2017): 1–14, 10.1177/2053951717736335.

[42] Morison and Harkens, 'Re-Engineering Justice?'

[43] Kevin Anthony Hoff and Masooda Bashir, 'Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust', *Human Factors: The Journal of the Human Factors and Ergonomics Society* 57, no. 3 (May 2015): 407–34, 10.1177/0018720814547570.

[44] Min Kyung Lee, 'Understanding Perception of Algorithmic Decisions: Fairness, Trust, and Emotion in Response to Algorithmic Management', *Big Data & Society* 5, no. 1 (January 1, 2018): 13, 10.1177/2053951718756684.

[45] For a comprehensive review, see: Hoff and Bashir, 'Trust in Automation'; John D Lee and Katrina A See, 'Trust in Automation: Designing for Appropriate Reliance', *Human Factors* 46, no. 1 (2004): 50–80.

[46] Lee and See, 'Trust in Automation: Designing for Appropriate Reliance', 51.

[47] Jennifer Logg, Julia Minson, and Don A. Moore, 'Algorithm Appreciation: People Prefer Algorithmic To Human judgement', SSRN Scholarly Paper (Rochester, NY: Social Science Research Network, April 24, 2018), https://papers.ssrn.com/abstract=2941774.

[48] R. M. Dawes, D. Faust, and P. E. Meehl, 'Clinical versus Actuarial judgement', *Science* 243, no. 4899 (March 31, 1989): 1668–74, 10.1126/science.2648573; Jaap J. Dijkstra, Wim B. G. Liebrand, and Ellen Timminga, 'Persuasiveness of Expert Systems', *Behaviour & Information Technology* 17, no. 3 (January 1998): 155–63, 10.1080/014492998119526.

[49] S. Shyam Sundar and Clifford Nass, 'Source Orientation in Human-Computer Interaction Programmer, Networker, or Independent Social Actor', *Communication Research* 27, no. 6 (December 1, 2000): 683–703, 10.1177/009365000027006001.

[50] S. Shyam Sundar, 'The MAIN Model: A Heuristic Approach to Understanding Technology Effects on Credibility', *Digital Media, Youth, and Credibility* 73100 (2008), http://www.marketingsociale.net/download/modello_MAIN.pdf.

[51] S. Shyam Sundar and Jinyoung Kim, 'Machine Heuristic: When We Trust Computers More Than Humans with Our Personal Information', in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19 (New York, NY, USA: ACM, 2019), 538:1–538:9, 10.1145/3290605.3300768.

[52] Aaron Smith, 'Public Attitudes Toward Computer Algorithms' (Pew Research Center, 2018), http://www.pewinternet.org/2018/11/16/public-attitudes-toward-computer-algorithms/.

rithm aversion. This is the notion that users will not use algorithms or ADM, for example, if they are aware that it can make – or see it making – mistakes.[53]

The question then becomes: What types of assumptions and expectations do citizens have when it comes to ADM systems, especially compared to human decision-makers? While most research on trust in automation has tended to focus more on humans as *operators* of the ADM system than as *subjects* of the decisions, emerging research has begun to explore perceptions of algorithmic decision-making in a broader manner. As a result, some key themes have begun to emerge.

In a study exploring perceptions about the justice of algorithmic decisions,[54] participants in a lab study (N=19) evaluated the justice of different scenarios concerning automated decisions according to six main themes: (1) lack of a human touch, (2) attempts to interpret how the system worked/reasoned, (3) acceptability of the usage of statistical inference in the scenario (or lack thereof), (4) the extent to which the explanation provided by the system was actionable, (5) aspects of the decision that may have been overlooked by the system, and (6) discussions about the meaning and the level of relevance of moral concepts such as *fairness* within the scenario. Moreover, research has also indicated that the *type of context or task* for which the ADM system is being used has an influence on perceptions about ADM,[55] as well as considerable disagreement when it comes to which types or features of information an ADM system should use for fair decision-making.[56]

While this stream of research offers important insights about perceptions regarding ADM decisions, the findings still need to be replicated and expanded to a broader sample of the population, with a comparison between ADM and human decision-makers, in particular, needing to be explicitly made. The present study, drawing from a representative sample of the Dutch adult population, aims to extend this line of research, and proposes three research sub-questions:

RQ1. To what extent are citizens inclined to perceive and accept ADM as fair in comparison to human decision-makers?

RQ2. Which principles and concerns do citizens consider to be decisive in their acceptance of ADM as fair?

RQ3. To what extent do individual characteristics such as age, gender and education level influence such perceptions?

3. Methods

3.1. Sample

To investigate public perceptions and assumptions regarding ADM fairness, we conducted a survey of a nationally representative sample of the Dutch adult population (18 years or older). Participants were recruited from a public opinion research company's database, which has over 115,000 registered respondents. The survey is part of a larger project investigating ADM by AI. We began by inviting a random sample, reflective of the national population for age, gender, region and educational level, to complete the survey (n = 3072). A total of 1069 panel participants accepted the invitation, out of which 958 completed the survey and provided informed consent. The final sample (N = 958) had an average age of 50.9 years (SD = 16.7) and was composed of 49% females.

3.2. Procedure

As part of a larger questionnaire investigating public perceptions about ADM by AI, respondents were asked: 'Who would, according to you, make a fairer decision: a human or artificial intelligence/computer? Could you please briefly explain why and provide an example of the type of decision you were considering when answering this question?' It is important to note that, before answering this question, participants had already read a definition of ADM by AI.[57] Responses to this question were open-ended and had an average of 19.39 words (SD =19.43).

3.2. Content analysis

Given the exploratory nature of this study, we combined qualitative and quantitative content analysis to investigate public perceptions of ADM fairness.

3.2.1. Qualitative content analysis

In the first step, inspired by the framework method,[58] the first two authors reviewed 100 responses independently in an open-coding stage, creating codes based on the statements provided by each participant. After these responses were inductively reviewed, the authors discussed their individual codes (approximately 180) and combined them in a list of 31 themes. These themes were used as a starting point for the

[53] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey, 'Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err'., *Journal of Experimental Psychology: General* 144, no. 1 (2015): 114–26, 10.1037/xge0000033; Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey, 'Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them', *Management Science* 64, no. 3 (March 2018): 1155–70, 10.1287/mnsc.2016.2643.

[54] Michael Veale, Max Van Kleek, and Reuben Binns, 'Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making', in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18 (New York, NY, USA: ACM, 2018), 440:1–440:14, 10.1145/3173574.3174014.

[55] Lee, 'Understanding Perception of Algorithmic Decisions', 13.

[56] N. Grgic-Hlaca, E. M. Redmiles, K. P. Gummadi and A. Weller, Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. Proceedings of the 2018 World Wide Web Conference, 903–912.

[57] The definition that was given to participants was: 'automated decision-making by artificial intelligence or computers can be defined as computer programs that can make decisions that were previously made by humans. These decisions are made automatically by computers based on data'.

[58] Nicola K. Gale et al., 'Using the Framework Method for the Analysis of Qualitative Data in Multi-Disciplinary Health Research', *BMC Medical Research Methodology* 13 (2013): 117, 10.1186/1471-2288-13-117.

Table 1 – Who is fairer?

	Responses	%	% valid
Humans	223	23%	33%
AI	362	38%	54%
Depends on circumstance	57	6%	9%
Both are equally fair/unfair	18	2%	3%
Both should work together	9	1%	1%
Unclear or no response[1]	289	30%	

[1] This category also included instances in which the respondent indicated that they did not know or could not provide an answer to the question of who was fairer (humans or AI).

development of a codebook for a quantitative content analysis, and later used for a qualitative review of the complete sample.

3.2.2. *Quantitative content analysis*

In the second step, two coders who were not part of the study were recruited to categorize the responses to the survey, using categories derived from the most frequent themes emerging from the qualitative content analysis. A total of four rounds of coder training and the double coding of random subsamples were needed to reach sufficient levels of reliability for the most common categories.

The responses were coded according to two main dimensions. The first dimension captured whether the respondent indicated a human or an AI decision-maker as fairer, with six mutually exclusive options, as outlined in Table 1. This dimension was reliable in the first two rounds of coding (round 1: 84% agreement, with 50 responses double coded, Krippendorff's $\alpha = 0.78$; round 2: 88% agreement, with 100 responses coded, $\alpha = 0.83$).

The second dimension encompassed the reasons for and assumptions about fairness. For these categories, the coders were instructed to categorize the responses as much as possible based on the concepts as expressed by the respondents and, should the response lack sufficient clarity, not to assign it to a specific category. Nine categories reached acceptable ($\alpha \geq 0.667$),[59] or nearly acceptable ($\alpha \geq 0.60$) levels, and were used in the analysis.[60] It is important to note that both dimensions were coded independently. One respondent, for example, might have indicated that they did not know whether AI or a human decision-maker would be fairer (first dimension), but could still have indicated that a human touch was important in the process (second dimension).

4. Results

4.1. *Who is the fairest of them all?*

The largest share of the respondents indicated that AI would be fairer than a human in making decisions (see Table 1), in response to RQ1. Some answers, however, were more nuanced, suggesting that fairness would depend on the circumstances; or that both AI and humans are equally fair/unfair; or that they should work together.

4.2. *Conditions for fairness*

In addition to indicating which kind of decision-maker they considered to be fairer, respondents also elaborated on the conditions under which, or reasons why, humans or AI would be considered fair/unfair. These statements referred to how the decision-making process itself should occur, expectations about how AI would perform as a decision-maker, or general considerations about ADM or fairness. In response to RQ2, below we report the categories that had a higher frequency of response and were sufficiently reliable to allow for a quantitative analysis of the content, as well as quotes that were considered illustrative during the qualitative content analysis. Given the combination of quantitative and qualitative approaches to content analysis, we also report on categories with a relatively small share of responses, which we consider provides a deeper level of insight into the topic.

4.2.1. *Decision-making process*

A first set of themes that emerged concerned the decision-making process and the conditions under which it would be considered fair. Four categories can be highlighted.

(1) The most frequent response category concerned the role of *emotions* in the decision-making process, mentioned in 25.4% of the responses. The ability to feel emotions was discussed both as a reason why a decision-maker would be considered fair: '*A person will generally make fairer decisions because a person has feelings. Feelings are also very important when making decisions. Compassion should never be lost. If a computer or a certain form of artificial intelligence, such as a robot, for example, is going to replace a judge, then there will never be a judgement based on any mitigating circumstances. It then becomes a tough society, without an ounce of humanity*' (R679); or unfair: '*The computer would be the fairest. No feelings play a role (…)*' (R516). From the responses that could be clearly identified as indicating that either humans or AI were fairer ($n = 585$), 21.9% considered AI to be fairer because of the lack of emotions, and 9.4% indicated that humans would be fairer because they had emotions. The association between both variables was significant (X^2 (1, $N = 585$) $= 6.85$, $p < .01$).

(2) The *risk of manipulation* in the decision-making process was mentioned by 4.3% of the respondents. In this category, AI was seen to be both (a) immune to manipulation: '*When it comes to fairness, the computer wins because, once programmed and activated, the computer will be absurdly fair (honest). A human, on the other hand, can be manipulated also when it comes to fairness*' (R61); and (b) also vulnerable to manipulation, being dependent on the data that is available: '*In principle, the computer could make a fairer decision based on all kinds of data, but who checks which data is correct and whether it has been manipulated or not?*' (R197). In relation to fairness, 3.8% of the responses associated AI with being fairer due to the lack of manipulation possibilities, versus 0.51% for humans. The

[59] Klaus Krippendorff, *Content Analysis: An Introduction to Its Methodology*, 2nd ed. (Thousand Oaks, Calif: Sage, 2004).

[60] Details about the content analysis and an overview of all categories (with levels of intercoder reliability) are available in Appendix A.

association between both variables was significant (X^2 (1, $N = 585$) = 6.44, $p < .05$).

(3) The *need for a human touch* to achieve fair decisions was also mentioned by 4.3% of the respondents. In this category, fair decisions were associated with those in which: '*the person, physically present, eye contact*' (R204) was needed, especially for: '*(…) decisions that require the human dimension, because objectivity alone is not enough*' (R339). Moreover, human dignity: '*the human decision will be fairer because human dignity plays an important role*' (R423); and human control of the process: '*I have no idea [who is fairer], but I would still like human control*' (R660), were also mentioned. For this category, 2.6% of the responses associated humans being fair with the need for a human touch, versus 1.4% for AI. The association between both variables was significant (X^2 (1, $N = 585$) = 6.30, $p < .05$).

(4) Finally, the need for the decision-maker to be able to *compare multiple arguments and take the context into account* was mentioned by 3.6% of the responses. This category usually associated humans with the ability to look deeper into the situation and thus make fairer decisions: '*Humans can make a more balanced decision. For example, someone drives through a red light. An AI system can then impose a fine, a person can take other circumstances into account (medical necessity or cause)*' (R153). In the case of AI, this was usually associated with processing or computing ability: '*I think artificial intelligence or a computer [is fairer] because a person can sometimes forget things. Artificial intelligence can of course also weigh things against each other to the minimum [level of detail], which sometimes falls short of a person*' (R45). Overall, 3.2% of the responses associated this category with humans being fairer, versus 1% with AI. The association between both variables was significant (X^2 (1, $N = 585$) = 14.25, $p < .001$).

4.2.2. *Assumptions about AI as a decision-maker*

Several responses elaborated on perceptions or assumptions about AI as a decision-maker. Two main categories could be identified.

(1) *AI being data and calculation-driven* was discussed in the context of fairness by 10.5% of the responses. This concerned both the advantage or claim to objectivity of AI as data or fact-driven: '*artificial intelligence will, I believe, be able to make objective decisions based on facts, which can be fairer*' (R192); but also as a limitation of AI: '*People can think and make decisions that are humane. A computer works based on data and cannot think, it is impossible to enter all of the data, you can hardly put social emotional [aspects] into data*' (R257). When contrasting humans and AI, 8.5% of the responses associated this category with AI being fairer, versus 3.2% with humans being fairer. The association between both variables was not significant (X^2 (1, $N = 585$) = 3.22, $p = .07$), indicating that this characteristic was not determinant when considering who was fairer (humans or AI).

(2) *AI being programmed by humans* was mentioned in 6.5% of the responses. Respondents who mentioned this category often provided a more detailed level of explanation in their answer. For example: '*People [are fairer], because AI is an illusion; humans indicate which parameters should be used. AI is so bad that they did not know how AI had taught itself things. If this kind of situation can arise with AI, then no one is responsible and the one who has the most money or is the strongest will simply take advantage*' (R763); or, as another respondent explained: '*This is the most important question in this discussion. An expert system is made by people, all rules come from a person; whether it is fair depends on the people who make the system (write down the algorithms). Just because it is a computer, does not mean that it is also fair. We have to realize that*' (R748). Other responses also indicated that AI could be fair(er) if programmed *correctly*, or *because* it is programmed by humans. When contrasting humans and AI, 2.4% of the responses associated this category with AI being fairer, versus 2.2% with humans being fairer. The association between both variables was not significant (X^2 (1, $N = 585$) = 0.80, $p = .37$).

4.2.2. *General considerations about ADM and fairness*

The final set of themes included general considerations about the acceptability of ADM in society, and general remarks about the notion of fairness itself.[61]

(1) The most frequent category within this set was related to *fairness being conditional on the type of decision*, mentioned in 4.2% of the responses. The importance of considering the context and type of decision was often one part of a response, such as: '*In my experience that depends very much on the person and the situation. A computer thinks logically, business-like, without emotion. Ideal in some situations, totally undesirable and inhuman in other situations*' (R198); or, as another respondent explained: '*It depends on the case. Emotional decisions, people, because, for example, they can incorporate intrinsic values. With technical decisions, computers, because they make a decision purely on the basis of data and people also include other aspects such as emotions in the decision*' (R839).

(2) A small share (1%) of responses offered general considerations about the actual notion of fairness. For example that: '*fairness is subjective. Every culture has different standards of fairness. For an American, the death penalty is fair, while in Europe it no longer exists*' (R6); or: '*I think that fair and fair can be different. A computer can be fairer in the sense that everyone is given equal opportunities or that the rules are met. A human can go outside the rules and have more insight into when an exception should be made, and therefore perhaps be fairer in another sense, in the form of favouring someone*' (R236).

(3) Finally, some respondents (0.9%) indicated a *non-acceptance of ADM for principled reasons*. Within this category, the responses that provided further elaboration indicated that being objective is not sufficient and questioned the overall consequences of ADM being implemented. For example: '*The computer is not biased, (…) works efficiently, accurately. But important decisions must always be made by people. I think it is a scary prospect that all decisions will be taken over by computers. Soon we will have nothing to add, everything will be taken by computers. It's already slipping in slowly, I think we haven't even*

[61] Categories in this set of themes were either almost always associated with humans being fairer (e.g. non-acceptance of ADM for principled reasons), or were not associated with either humans or AI being clearly fairer. As such, Chi-Square tests of association are not reported.

Table 2 – The influence of personal characteristics (*N* = 585).

AI fairer than humans	Estimate	Standard Error	Odds Ratios[CI: 2.5–97.5]
Intercept	0.22	0.43	1.24 [0.54–2.89]
Gender (Female)	−0.03	0.18	0.97 [0.68–1.39]
Age (years)	−0.02***	0.005	0.98 [0.97–0.99]
Level of Education	0.29***	0.05	1.34 [1.20–1.49]

*** $p < .001$.

Table 3 – Personal characteristics and conditions for fairness (*N* = 958).

	Age	Gender	Education
Decision-making aspects			
The role of emotion	−0.01 (0.004)*	0.03 (0.15)	0.03 (0.04)
The risk of manipulation	−0.02 (0.01)	−0.17 (0.32)	0.07 (0.10)
Need for a human touch	0.02 (0.01)*	0.22 (0.33)	0.01 (0.09)
Comparison of arguments	0.004 (0.01)	0.32 (0.35)	0.26 (0.11)*
Assumptions about AI			
Data/calculation-driven	0.004 (0.01)	−0.07 (0.22)	0.21 (0.07)**
Programmed by humans	0.01 (0.01)	0.06 (0.27)	−0.05 (0.08)
Considerations about ADM and fairness			
Fairness conditional	−0.01 (0.01)	0.97 (0.36)**	0.41 (0.12)***
What is fairness?	0.02 (0.02)	0.29 (0.66)	0.32 (0.21)
Non-acceptance of ADM	0.04 (0.02)	0.35 (0.69)	−0.20 (0.19)

Notes: Intercept and odds-ratios not reported for conciseness. Standard errors in parentheses.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

noticed' (R525); or as another respondent explained: '*Perhaps initially artificial intelligence [is] more objective, but that is not everything. It's about people, their lives, and you can't just leave that to a computer. Every situation is unique. A human must make the final judgement*' (R614).

Two quotes further demonstrate the complexity of the topic, as articulated in the considerations made by the respondents regarding ADM and fairness. On the one hand, one respondent clearly highlighted the idea that once a societal position on a rule has been reached, AI will implement it consistently. This was a reason to favour ADM: '*I assume that artificial intelligence or a computer makes decisions based on overall decision-making, while a "person" makes a personal decision that is absolutely is not based on overall decision-making. In such a situation, I prefer decisions made by artificial intelligence or a computer*' (R500). On the other hand, another respondent argued that even if ADM might be considered fair, it may not be sufficient: '*A computer will, in my opinion, always make a fairer decision than a human being. A computer would need to "learn" unfairness, unless such an "algorithm" is designed as such [as unfair]. [However], whether decisions made by a computer are always ethically acceptable is questionable*' (R552).

4.2. *The influence of personal characteristics*

Finally, in response to RQ3, we investigated the extent to which personal characteristics – age, gender and level of education – were associated with perceptions regarding ADM fairness. When considering responses that were clearly in favour of AI or of humans ($N = 585$), the results of a logistic regression model indicated that age was negatively associated with the likelihood of AI being considered fairer than humans, while levels of education were positively associated (see Table 2). There was no significant association with gender.

With respect to the specific categories, similar logistic regression models were also executed. Table 3 summarizes the effects based on the category, and highlights how education levels, primarily, but also age and to a lesser extent gender, influence the likelihood of these conditions for fairness being mentioned.

5. Discussion

We now turn to a reflection on some of the key findings of this research. This is not to say that we believe that citizens should or could be the ultimate arbiters in deciding whether and, if so, how the integration of ADM into decision-making procedures can incorporate ethical, moral or legal conceptions of fairness. Indeed, as the results of our empirical inquiry illustrate, people's judgements are often clouded by a range of important misconceptions about technology. Nevertheless, as we will show, understanding the attitudes of citizens towards ADM offers important insights regarding the conditions under which they would find it acceptable to take part in ADM processes, as well as their assumptions and expectations about how these systems should or do work.

5.1. *What makes ADM fair?*

In response to RQ1, we found that while the responses which considered ADM to be the fairer decision-maker outnumbered those that considered humans to be fairer, almost one third of the respondents were unable – or unwilling – to choose. This suggests that while much remains open and unclear with regard to people's attitudes towards ADM, at least from their perspective, ADM might enhance justice and contribute to fairer decision-making.[62] However, the responses also demonstrated that, for people to accept automated systems as decision-makers, more is needed. One central reason why people believed ADM to be fairer lay in the idea of a *systematic execution of the rules that society has agreed on, much in the Rawlsian sense that fair decisions reflect the agreement of society on the rules that decisions need to follow to be fair*.[63] An important element of this procedural interpretation of fairness in decision-making is the objectivity and alleged immunity of algorithmic systems to manipulation.

Nevertheless, *the infallible execution of rules was, at the same time, also a central concern of those who were critical* of ADM. A central quality that other respondents valued as an ingredient of fairer decision-making in humans was the ability to consider the broader context and, where necessary, make an exception to the rule. Closely related to this point was another crucial, although more controversial human quality of decision-makers, namely the role of emotions and the ability to show empathy, as part of a fair decision-making process. An inherently human trait, emotions could be as much a reason for preferring human decision-makers as for preferring ADM, depending, of course, on the kind of decision at stake. Interestingly, respondents who cited this as an important aspect of decision-making were more likely to associate the lack of emotion in ADM as a condition of ADM being the fairer decision-maker.

5.2. *Potential and limits to the datafication of justice*

There is a broader, more philosophical question that is perhaps implicit in the answers of the respondents, *namely, where to draw the limits to the datafication of justice*. In many of the responses, there was little doubt that machines can process larger amounts of information faster and more objectively than humans. However, as a number of respondents also remarked, reality is not always black and white, or easily captured in models. There is a limit to generalizability and the modelling of reality. Context does matter. Fairness in the statistical sense does not automatically translate into justice, at least not at an individual level.

Taking this further, we would like to argue that when discussing fairness in ADM it is necessary to distinguish more explicitly between a *narrow perspective on fairness in algorithmic systems* (i.e. how can we design algorithms that are able to achieve a fair balance between different, relevant decision factors, and to what extent are these algorithms able to sufficiently take into account the broader context?) and a *broader societal perspective on automated fairness* (i.e. even if automated decisions can be fair in the statistical sense, is that the sort of fairness that we, as a society, should strive for?) To date, much of the discussion of fairness in ADM has centred around the first aspect – how to make ADM systems fairer. There has been comparatively less debate on the second aspect – the question of whether there are certain situations in which we should resist the trend to automatization, or perhaps even ban ADM. For example, in more complex situations that do not lend themselves very well to modelling and statistical generalization; decisions that are particularly context-dependent; or where taking into account individual circumstances minimizes the risk of *unfair decisions*. Particularly in the legal realm, individual subjective elements such as intent, as well as objective factors, such as the existence of an emergency situation, a personal relationship or a situation of dependence, should be factored into the ultimate decision and may be difficult to model.[64]

Are there *situations in which we, as a society, decide that individuality, compassion and emotion should outweigh procedural fairness, now and in the future?* These questions are highly relevant to the legal/ethical debate about where to set limits to ADM. Arguably, there are situations in which *ADM is simply not socially acceptable or just,*[65] *even if, in theory, it would be possible to design fair ADM systems for these situations*. For example, there could be situations in which there is no clear 'right' and 'wrong', but where the ultimate decision is more a moral issue than a decision based on which rule applies. There might also be situations in which it is not so much fairness, but also, and even more so, human vulnerability and dignity that are at stake.

Issues such as these should also inform law and policymaking around ADM. For example, to date, the ability to object to ADM (e.g. in Art. 22 GDPR) is very restricted, and only possible for fully automated processing, decisions with a legal effect and subject to a number of limitations. Our findings question the limited scope of this provision and prompt the question of whether there should be additional grounds that entitle people to object to automated decisions, for example arguments related to dignity, a lack of trust in such systems rendering a balanced decision, or moral objections to the very idea of being subject to ADM.

Looking to the future, for legislators and policymakers this could mean that, as important as it is to set out regulatory guidance on what fair ADM is, or how people might have the ability to challenge automated decisions, there might also be a role for law makers in determining the situations in which ADM is societally unacceptable in any form,[66] and accordingly

[62] Prins, 'Digital Justice'.

[63] Rawls, 'Justice as Fairness', 176.

[64] La Diega, Guido Notto. 'Against the Dehumanisation of Decision-Making – Algorithmic Decisions at the Crossroads of Intellectual Property, Data Protection, and Freedom of Information'. JIPITEC 9 no. 1 (2018).

[65] Taylor, 'What Is Data Justice?'

[66] See for example the French prohibition of semi or fully automated decision making procedures with the objective to evaluate aspects of personality, Art. 10(1), Loi n ° 78-17 du 6 janvier 1978 as amended by Loi n °2018-493 du 20 juin 2018, discussed in Malgieri, Gianclaudio. 'Automated decision-making in the EU Member States: The right to explanation and other 'suitable safeguards' in the national legislations'. Computer Law and Security Review 35, no. 5 (2019). See also Council of Europe, Recommen-

cannot be used in exceptions, or rely on consent or contract, or be justified by law (Art. 22 (2) GDPR).

5.3. *ADM fairness is not necessarily justice*

Perhaps one of the key insights from the survey was that even in situations in which AI is considered to be able to make fairer decisions, this does not automatically mean that people are willing to accept ADM. This observation is particularly important in relation to the notion of procedural fairness in the Rawlsian sense, in which it is not sufficient for citizens to agree that certain decisions are fair or not, which would be literally impossible, but that they are willing to *accept* the judgements made, providing they were made following the procedural or other rules agreed on by a society.[67] Our research shows that *whether people are willing to do so also depends on the extent to which other factors are present, including: respect for human dignity, an ability to express and understand emotions, and a human touch.*[68]

In other words, fairness in decision-making not only concerns the outcome of a decision but also has an *inter-relational component*. Even if machines are better at mastering the data, do we want to follow their lead? In our research we found that that for users to accept a decision as *just* does not depend on fairness alone, or the lack of emotions in the decision-maker, but also upon the emotional response of those subjected to a decision, and whether they feel that their case has been adequately considered.

Consequently, and with respect to future work, formalizing fairness in ADM and ensuring the necessary checks and balances[69] remains one challenge, but another, and no less important, challenge is designing and implementing ADM systems in a way that humans feel that their dignity is respected and that they are treated appropriately as humans, not numbers. In concrete terms, this means that in the legal and policy debate around ADM and its integration into societal processes, it is important to focus not only on fairness in ADM itself, but the way it is implemented and affects inter-human relationships. This primarily concerns the development of new professional ethics regarding the use of ADM systems and the way these systems are integrated into professional routines and how they deal with recipients or subjects of a decision. In a similar vein, more research is needed to ascertain whether the current safeguards in Art. 22 (3), which are intended to guarantee a right to human intervention – to express one's point of view and to contest the decision – have a positive effect on perceived justice, or whether there is a need for additional safeguards.

5.4. *Long-term implications for the possibility of fairness in ADM*

Interestingly, some of the respondents not only considered fairness in ADM, but also how our concept of fairness might change under the influence of ADM. The picture that these respondents painted was consistent and not particularly optimistic. As one of the respondents framed it: '*As a result, there is a risk of less willingness to think in a nuanced manner in society. And promoting "winners" and "losers", also [in] politics. With hardening [of positions]*' (R308). Another respondent found the prospects rather frightening, if not totalitarian: '*North-Koreanish*' (R502). Moreover, a third warned that: '*(…) based only on numbers, only cold and chilling decisions can be made*' (R539).

Thus far, much of the discussion around fairness in ADM has centred around the question of how concepts of fairness can be translated into algorithmic decision-making. More attention and research need to focus on the medium *to long-term societal implications, and the way ADM interacts with and impacts on our very notion of fairness*. Will what we consider 'fairness' today be the same in 20 years when the automation of decisions is far more common? How will the integration of ADM systems affect the way fairness (and justice) is operationalized in society? Considering the potential transformative impact of ADM on society, and on our very notions of fairness and justice, arguably the implementation of ADM into daily routines should be accompanied by a suitable monitoring framework that is able to signal undesired societal consequences and side effects.

5.5. *Fairness and contestation*

Another interesting aspect of human decision-making that has been brought to the fore, but which has not been very prominent in the discussion on the fairness of ADM thus far, is the human tendency to make mistakes and change opinions. As some of the responses highlighted, we accept that humans are not infallible, that they can make mistakes, and that we can correct them. To err is human. Importantly, leaving room for error and the correction of error or erroneous judgements is actually an aspect of procedural justice, and the reason why there is room for contestation and redress.

If we accept that negotiating uncertainty and grey zones in decision-making are part of what makes decisions ultimately fair and just, this begs the question of how much room ADM will leave to humans to differ in their opinions. Moreover, how much room for error should we grant ADM? Thus far, much of the discussion has been about responsible AI design, and fairness in ADM seems to suggest that there is already agreement on what fairness is, or that a fair decision is a value for which a system can be optimized, provided we are able to find the right definition of 'fairness'.

Fairness, however, is not a static value; it is the result of a *balancing process*. There is not just one possible fair outcome of any decision-making process, and perhaps in some situations there will be a number of possibilities, and the ability to weigh the different arguments and our right to be heard are at least as important, if not more, as other factors in ensuring that a decision is just and fair. *Thus, perhaps we should start conceiving algorithmic fairness as not so much concerned with an outcome but*

---

dation CM/Rec(2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems, adopted by the Committee of Ministers on 8 April 2020, Appendix, Section A, RN 15.

[67] Rawls, 'Fairness as Justice", 176.

[68] Reuben Binns et al., ''It's Reducing a Human Being to a Percentage'; Perceptions of Justice in Algorithmic Decisions', *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI* '18, 2018, 1–14, 10.1145/3173574.3173951.

[69] Prins, 'Digital Justice'.

a process – a process that leaves room for error, contradiction and competing arguments, precisely an ability to show resistance, as suggested by Morison and Harkens.[70]

Moreover, in addition to defining rules of algorithmic procedural fairness and the right of human beings to contest wrong decisions, perhaps we should start exploring whether part of making ADM more acceptable to people would also entail reducing the deterministic impact of ADM and creating more room for resistance and disagreement with an automated decision. Thus far, Art. 22 (3) GDPR, concerning such safeguards, has received relatively little attention regarding the details or concrete requirements for implementation by national regulators.[71] Findings such as those from our survey suggest that procedural safeguards and the ability to contest or otherwise exert human control over ADM could be an important factor contributing to the perceived justice and acceptability of ADM, and should therefore receive more attention, both in research as well as in the realms of law and policymaking.

5.6. *Limitations and future research*

While this study makes an important contribution in bringing to light not only the current expectations and assumptions regarding the fairness of ADM – and the underlying reasons people have for considering AI or human decision-makers fairer – some limitations must be acknowledged. As we aimed to capture overall public perceptions on the issue, the survey question asking about perceptions of ADM fairness in comparison to human decision-makers was articulated in a broad manner and analysed using a combination of qualitative and quantitative content analysis techniques. While this approach helped gather multiple insights into the issue as a general societal trend, future research should extend these findings by evaluating how perceptions and assumptions about ADM as a fair decision-maker vary depending on the context and specific situations of professional decision making (e.g. in the area of justice, work, public service, education, etc.), as well as engaging with the rich, emerging literature on the issue of fairness.[72] Such research should also more actively explore the role of socio-demographic factors (e.g. socio-economic status, field of work, etc.), as well as previous experience with ADM systems, in shaping people's expectations and underlying assumptions about the fairness of these systems. Moreover, in order to elicit clear, intuitive answers and stimulate respondents to think about exactly the way how machine decision making compares to human decision making we chose to present them with two extremes: automated vs human decision making. In practice, many decision-making systems will adopt a hybrid approach in which humans and ADM systems will co-operate. Further research should be studying responses to such mixed systems, and explore under which conditions a 'human in the loop' contributes to perceptions of procedural fairness. Doing so will also help to advance our understanding of Art. 22 GDPR, which is the provision that deals with (fully) automated decision making and that stipulates, without much further detail, a right of users to request human intervention.

5.7. *Who is the fairest of them all?*

In conclusion, this study found that a greater number of respondents considered AI the fairer decision-maker, with some possibly harbouring an almost idealistic (and dangerous because misleading) belief in the potential and objectivity of ADM. This seems to reflect a belief in a machine heuristic,[73] also providing additional evidence about levels of algorithm appreciation in contemporary society.[74] Some respondents, however, seemed primarily to be disappointed with human decision-makers,[75] while another, substantial share of respondents were convinced that humans are and will remain the fairer decision-makers.

A more promising view, we would hope, might lie with another, not unsubstantial group: those who recognized that algorithmic versus human fairness is not an 'either/or question'. Instead, there are varying boundaries concerning where human fairness ends and machine fairness begins. Ultimately, the real question is how humans and machines can usefully cooperate to make decisions fairer. In the words of R272: '*I think that the strength lies in their combination. The data from the computer cannot do without human input. And people cannot live without data*'.

Declaration of Competing Interest

There are no competing interests to declare.

Acknowledgment

This research was supported by the Research Priority Areas Information & Communication in the Data Society (https://www.uva-icds.net/), and Communication and its Digital Communication Methods Lab (digicomlab.eu) at the University of Amsterdam. The funder had no influence on the research, research design or execution.

Appendix A – Content analysis

A.1. *Qualitative content analysis – themes*

The following themes emerged from a review by the first two authors (as outlined in the methods section):

1 Ability to make exceptions and take individuals into account

[70] Morison and Harkens, 'Re-Engineering Justice?', 4.
[71] Malgieri, 'Automated decision-making in the EU Member States'.
[72] Baleis et.al., 'Cognitive and Emotional Response to Fairness in AI – A Systematic Review'.
[73] Sundar, 'The MAIN Model'; Sundar and Kim, 'Machine Heuristic'.
[74] Logg, Minson, and Moore, 'Algorithm Appreciation'.
[75] With the current President of the USA, Donald Trump, being among the names of human decision-makers mentioned in these instances.

2 Distrust in humans' ability to make fair decisions
3 Acceptability of ADM – even if you accept it as fairer, which are the conditions to accept its decisions
4 AI better in calculating, data-driven
5 Limits of data and datafication
6 AI rule-driven, consistency and consequent application of fairness
7 Potential for manipulation
8 AI cannot detect lies
9 AI programmed by humans
10 Conditions of fairness
11 Fairness depends on the context
12 The role of emotion
13 Division of tasks AI and humans
14 Human touch
15 What is fairness?
16 Human dignity / 'I don't want to have my life decided by a computer'
17 Unique human qualities
18 Humans make more human decisions
19 Humans are inherently biased
20 AI inherently unbiased
21 Ability to see the broader context
22 Predictability
23 Humans as decision-makers
24 Not everything is black or white
25 Possibility to give input / appeal
26 Ability to err
27 Role of education
28 Totalitarian / technocratic
29 Transparency & black box
30 Explainability
31 Ability to compare/contrast arguments

A.2. *Quantitative content analysis – categories*

The themes that emerged in the qualitative content analysis were subsequently reviewed and further combined into categories within a codebook for the quantitative content analysis stage. After four rounds of coder training and double-coding of the responses (R1: 50 responses, R2: 100, R3: 150, R4: 100), ten categories – including a general question about who is fairer? – were frequent and reliable enough to be reported in this paper. They are outlined below. The complete codebook can be provided upon request.

A.2.1. *Who is fairer?*

This question covered whether the respondent considers a Human or an AI as being fairer in decision maker.

Number	Code	Code name
1	HUMAN	if the Human is considered as fairer
2	AI	if the AI is considered fairer
3	MIXED	if the respondent said both are fair (or unfair), depending on the circumstance
4	UNCLEAR	if the respondent did not provide a clear answer about fairness[1]
5	EQUAL	if the respondent clearly indicates that both are equally fair (or unfair) regardless of the circumstance
6	COMBINED	if the respondent clearly indicates that both should be combined in order to be fair

[1]This category also included instances in which the respondent indicated that they did not know or could not provide an answer for who was fairer (humans or AI).

Intercoder reliability results: Round 1 - 84% agreement, Krippendorff's $\alpha = 0.78$; round 2: 88% agreement, $\alpha = 0.83$.

A.2.2. *Reasons for fairness*

This question focused on the reasons that respondents provide to justify their responses regarding who is fairer regarding the following dimensions:

Dimension	Definition
Decision-making process	Aspects about the decision itself, and how it should be taken. This includes for example what the decision-maker should consider, whether emotions are (or are not) part of the considerations, how arguments should be evaluated etc.
Characteristics of AI or Humans as decision-makers	Statements that highlight certain characteristics of AI or humans. The codes under this category may or may not contain an explicit value judgement about the characteristic.
Considerations about ADM and Fairness	Statements that reflect upon the implementation of automated decision-making at a *societal* or *structural* level. They do not cover individual aspects of a decision (or how a decision should be made), but rather the implications of implementing ADM, or how ADM should be implemented. They also include statements that reflect upon the concept of fairness in general, including whether fairness is an absolute or relative concept etc.

Each response could belong to multiple dimensions (and categories within each dimension). Each statement made by the respondent was evaluated and coded accordingly. The reliability scores provided are for the fourth round of coding.

A.2.2.1. *Decision-making process*

Code	Definition	Intercoder Reliability
Comparison of arguments	Response highlights that a decision-maker must have (or misses) the ability to compare and/or contrast multiple arguments and take multiple factors into account.	Agreement: 100% $\alpha = 0.1$

(Continued on next column)
(Continued on next page)

Code	Definition	Intercoder Reliability
Emotion	Response mentions emotion or feelings (includes empathy) as a source of fairness or of bias. *Note: this code is also relevant for emotion (or lack thereof) mentioned as an AI or human characteristic.*	Agreement: 95% $\alpha = 0.810$
Manipulation	Answer suggests that the decision-maker (human or AI) can be manipulated when making a decision	Agreement: 97%, $\alpha = 0.754$
Need for human touch	Response indicates that 'eye contact', or physical/human presence is important for the decision. This also includes statements that highlight that human(e) aspects, including human control, are needed in the decision-making process.	Agreement: 94% $\alpha = 0.668$

A.2.2.2. *AI characteristics*

Code	Definition	Intercoder Reliability
AI data- and calculation-driven	Response mentions that AI is fact-, data-based, better in performing calculations, and more efficient.	Agreement: 92% $\alpha = 0.623$
AI programmed by humans	Response mentions that AI or ADM is controlled or programmed by humans.	Agreement: 96% $\alpha = 0.797$

A.2.2.3. *Considerations about ADM and fairness*

Code	Definition	Intercoder Reliability
Fairness conditional to the type of decision	The statement indicates that the type of decision (or context) is important to determine whether AI or humans are fairness (e.g., how fairness is defined depends on the type of decision - and therefore humans or AI might be better). Statements that indicate that fairness depends on the type of decision, or that AI and/or humans could be fair depending on the decision or situation fall under this category.	Agreement: 93% $\alpha = 0.658$
Non-acceptance of ADM for principled reasons	Response indicates that the respondent will not accept ADM even if it is fair based on principles. These can be for example religion, worldviews of society, and aspects of human dignity.	Agreement: 99% $\alpha = 0.663$
What is fairness?	Response discusses the extent to which fairness exists, or how it can be defined. Also includes discussions of differences between being fair and being objective.	Agreement: 99% $\alpha = 0.663$

REFERENCES

Alarie B, Niblett A, Yoon AH. How artificial intelligence will affect the practice of law. Univ. Toronto Law J. 2018;68(supplement 1):106–24. doi:10.3138/utlj.2017-0052.

Aletras N, Tsarapatsanis D, Preoţiuc-Pietro D, Lampos V. Predicting judicial decisions of the European court of human rights: a natural language processing perspective. PeerJ Comput. Sci. 2016;2:e93.

AlgorithmWatch. Automating society: taking stock of automated decision-making in the EU. AlgorithmWatch 2019 https://algorithmwatch.org/wp-content/uploads/2019/01/Automating_Society_Report_2019.pdf.

Angelopoulos Ch. MTE v Hungary : A new ECtHR judgement on intermediary liability and freedom of expression. J. Intell. Prop. Law Pract. 2016;11(8):582–4. doi:10.1093/jiplp/jpw081.

Baleis, J., Keller, B., Starke, C., & Marcinkowski, F. Cognitive and emotional response to fairness in AI – a systematic review, Working Paper, https: https://www.phil-fak.uni-duesseldorf.de/fileadmin/Redaktion/Institute/Sozialwissenschaften/Kommunikations-_und_Medienwissenschaft/KMW_I/Working_Paper/Baleis_et_al.__2019__Literatur_Review.pdf

Barna, L., D. Juhász, and S. Márok. What makes a good judge'. Budapest: European Judicial Training Network Themis Competition, 2017. http://www.ejtn.eu/Documents/Team%20HU%20semi%20final%20D.pdf.

Barocas S, Selbst AD. Big data's disparate impact. Calif. L. Rev. 2016;104:671.

Bellamy R. The democratic qualities of courts: a critical analysis of three arguments. Representation 2013;49(3):333–46. doi:10.1080/00344893.2013.830485.

Binns R. Data protection impact assessments: a meta-regulatory approach. Int. Data Priv. Law 2017;7(1):22–35.

Binns R, Kleek MV, Veale M, Lyngs U, Zhao J, et al. It's reducing a human being to a percentage'; perceptions of justice in algorithmic decisions. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems – CHI; 2018. p. 1–14.

Blum L. Moral perception and particularity. Ethics 1991;101(4):701–25. doi:10.1086/293340.

Butterworth M. The ICO and artificial intelligence: the role of fairness in the GDPR framework. Comput. Law Secur. Rev. 2018;34(2):257–68. doi:10.1016/j.clsr.2018.01.004.

Citron DK. Technological due process. Washington Univ. Law Rev. 2008;6:1249–314.

Clifford D, Ausloos J. Data protection and the role of fairness. Yearb. Eur. Law 2018;37:130–87. doi:10.1093/yel/yey004.

Cranston R. What do courts do. Civ. Just. Q. 1986;5:124.

Dawes RM, Faust D, Meehl PE. Clinical versus actuarial judgement. Science 1989;243(4899):1668–74. doi:10.1126/science.2648573.

Dietvorst BJ, Simmons JP, Massey C. Algorithm aversion: people erroneously avoid algorithms after seeing them Err. J. Exp. Psychol.: Gen. 2015;144(1):114–26. doi:10.1037/xge0000033.

Dietvorst BJ, Simmons JP, Massey C. Overcoming algorithm aversion: people will use imperfect algorithms if they can (even slightly) modify them. Management Science 2018;64(3):1155–70. doi:10.1287/mnsc.2016.2643.

Dijkstra JJ, Liebrand WBG, Timminga E. Persuasiveness of expert systems. Behav. Inf. Technol. 1998;17(3):155–63. doi:10.1080/014492998119526.

Domselaar IV. Moral quality in adjudication: on judicial virtues and civic friendship. Netherlands J. Legal Philos. 2015;1:24–46.

Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R. Fairness through awareness. Proceedings of the Third Innovations in Theoretical Computer Science Conference. ACM; 2012. p. 214–26.

Feldman A, Menounou E. What motivates the justices: utilizing automated text classification to determine supreme court justices' preferences'. Proceedings of the Annual Meeting of the Southern Political Science Association (SPSA), 2015.
Gale NK, Heath G, Cameron E, Rashid S, Redwood S. Using the framework method for the analysis of qualitative data in multi-disciplinary health research. BMC Med. Res. Methodol. 2013;13:117. doi:10.1186/1471-2288-13-117.
Goodman B, Flaxman S. European Union regulations on algorithmic decision-making and a 'right to explanation. AI Mag. 2017;38(3):50–7.
Heidari H, Loi M, Gummadi KP, Krause Andreas. A moral framework for understanding fair ML through economic models of equality of opportunity. Proceedings of the Conference on Fairness, Accountability, and Transparency – FAT*. ACM Press; 2019. p. 181–90.
Hoff KA, Bashir M. Trust in automation: integrating empirical evidence on factors that influence trust. Hum.Fact.: J. Hum. Fact. Ergonom. Society May 2015;57(3):407–34. doi:10.1177/0018720814547570.
Kamiran F, Calders T, Pechenizkiy M. Techniques for discrimination-free predictive models. Discrimination and Privacy in the Information Society. Springer; 2013. p. 223–39.
Katz DM. Quantitative legal prediction-or-how i learned to stop worrying and start preparing for the data-driven future of the legal services industry. Emory Law J. 2012;62:909.
Klijnsma JG. Contract Law as Fairness: A Rawlsian Perspective on the Position of SMEs in European Contract Law. Universiteit van Amsterdam; 2014.
Krippendorff K. Content Analysis: An Introduction to Its Methodology. 2nd ed. Thousand Oaks, Calif: Sage; 2004.
Kroll JA, Barocas S, Felten EW, Reidenberg JR, Robinson David G, et al. Accountable algorithms. Univ. Pennsylvania Law Rev. 2017;3:633–706.
Diega L, Notto G. Against the dehumanisation of decision-making – algorithmic decisions at the crossroads of intellectual property, data protection, and freedom of information. JIPITEC 2018;9(1) urn:nbn:de:0009-29-46778.
Lee JD, See KA. Trust in automation: designing for appropriate reliance. Hum. Fact. 2004;46(1):50–80.
Lee MK. Understanding perception of algorithmic decisions: fairness, trust, and emotion in response to algorithmic management. Big Data Soc. 2018;5(1) 2053951718756684. doi:10.1177/2053951718756684.
Logg J, Minson J, Moore DA. Algorithm Appreciation: People Prefer Algorithmic to Human judgement. Rochester, NY: Social Science Research Network; 2018 SSRN Scholarly Paper https://papers.ssrn.com/abstract=2941774 .
Malgieri G. 'Automated decision-making in the EU member states: the right to explanation and other 'suitable safeguards' in the national legislations'. Comput. Law Secur. Rev. 2019;35(5). doi:10.1016/j.clsr.2019.05.002.
McGinnis JO, Pearce RG. The great disruption: how machine intelligence will transform the role of lawyers in the delivery of legal services. Fordham Law Rev. 2014;82:3041.
Morison J, Harkens A. Re-engineering justice? Robot judges, computerised courts and (semi) automated legal decision-making. Legal Stud. 2019:1–18. doi:10.1017/lst.2019.5.
Olsaretti S. The idea of distributive justice. In: Olsaretti Serena, editor. The Oxford Handbook of Distributive Justice. Oxford University Press; 2018.
Oswald M. Algorithm-assisted decision-making in the public sector: framing the issues using administrative law rules governing discretionary power. Philos. Trans. R. Soc. A: Math. Phys. Eng. Sci. 2018;376(2128).
Poort J, Zuiderveen Borgesius FJ. 'Does everyone have a price? Understanding people's attitude towards online and offline price discriminationdoes everyone have a price? understanding people's attitude towards online and offline price discrimination'. Internet Policy Rev. 2019. doi:10.14763/2019.1.1383.
Prins C. Digital justice. Comput. Law Secur. Review 2018;34(4):920–3. doi:10.1016/j.clsr.2018.05.024.
Radeideh M. The Principle of Fair Trading in EC Law: Information and Consumer Choice in the Internal Market. Rijksuniversiteit Groningen; 2004.
Rawls J. Justice as fairness. Philos. Rev. 1958;67(2):164–94. doi:10.2307/2182612.
Selbst AD, Boyd D, Friedler SA, Venkatasubramanian S, Vertesi J. Fairness and abstraction in sociotechnical systems. Proceedings of the Conference on Fairness, Accountability, and Transparency – FAT* '19. ACM Press; 2019. p. 59–68.
Smith A. Public attitudes toward computer algorithms. Pew Res. Center 2018 http://www.pewinternet.org/2018/11/16/public-attitudes-toward-computer-algorithms/.
Sourdin T. Alternative Dispute Resolution. 5th ed. Pyrmont: Thomson Reuters; 2016.
Sourdin T, Cornes R. Do judges need to be human? The implications of technology for responsive judging. The Responsive Judge. Springer; 2018. p. 87–119.
Sundar SS. The MAIN model: a heuristic approach to understanding technology effects on credibility. Digital Med. Youth Credib. 2008;73100 http://www.marketingsociale.net/download/modello_MAIN.pdf.
Sundar SS, Kim J. Machine heuristic: when we trust computers more than humans with our personal information. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. ACM, 2019.
Sundar SS, Nass C. Source orientation in human-computer interaction programmer, networker, or independent social actor. Commun. Res. 2000;27(6):683–703. doi:10.1177/009365000027006001.
Taylor L. What is data justice? The case for connecting digital rights and freedoms globally. Big Data Soc. 2017;4(2):1–14. doi:10.1177/2053951717736335.
Veale M, Brass I. Administration by algorithm? Public management meets public sector machine learning. In: Yeung K, Lodge M, editors. Algorithmic Regulation. Oxford University Press; 2019.
Veale M, Kleek MV, Binns R. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. ACM; 2018. p. 1–440 14. CHI '18.
Wachter S, Mittelstadt B. A Right to reasonable inferences. Columb. Bus. Law Rev. 2019(2):494–620.
Whittaker M, Crawford K, Dobbe R, Fried G, Kaziunas E, Mathur V, et al. AI now report 2018. AI Now 2008 https://ainowinstitute.org/AI_Now_2018_Report.pdf.
Zarsky Tal. The trouble with algorithmic decisions: an analytic road map to examine efficiency and fairness in automated and opaque decision making. Sci. Technol. Hum. Val. 2016;41(1):118–32 Conference Name: ACM Woodstock conference. doi:10.1177/0162243915605575.