'It's Reducing a Human Being to a Percentage'; Perceptions of Justice in Algorithmic Decisions

Reuben Binns[1], Max Van Kleek[1], Michael Veale[2], Ulrik Lyngs[1], Jun Zhao[1] and Nigel Shadbolt[1]
[1]Dept. of Computer Science, University of Oxford, United Kingdom
[2]Dept. of Science, Technology, Engineering & Public Policy, University College London, United Kingdom
{reuben.binns, emax, ulrik.lyngs, jun.zhao, nigel.shadbolt}@cs.ox.ac.uk, {m.veale}@ucl.ac.uk

ABSTRACT

Data-driven decision-making consequential to individuals raises important questions of accountability and justice. Indeed, European law provides individuals limited rights to 'meaningful information about the logic' behind significant, autonomous decisions such as loan approvals, insurance quotes, and CV filtering. We undertake three experimental studies examining people's perceptions of justice in algorithmic decision-making under different scenarios and explanation styles. Dimensions of justice previously observed in response to human decision-making appear similarly engaged in response to algorithmic decisions. Qualitative analysis identified several concerns and heuristics involved in justice perceptions including arbitrariness, generalisation, and (in)dignity. Quantitative analysis indicates that explanation styles primarily matter to justice perceptions only when subjects are exposed to multiple different styles—under repeated exposure of one style, scenario effects obscure any explanation effects. Our results suggests there may be no 'best' approach to explaining algorithmic decisions, and that reflection on their automated nature both implicates and mitigates justice dimensions.

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous; K.4.1 Computers and Society: Public Policy Issues

Author Keywords

Algorithmic decision-making;
explanation; justice; fairness; machine learning; transparency

INTRODUCTION

Important decisions—about the allocation of jobs, loans, or insurance—are increasingly based on the predictions and classifications of computer models. By training such models on large datasets of previous customers or employees, organisations seek to more accurately estimate the risks and rewards of different actions and act accordingly—sometimes even automatically. Such practices are commonly referred to as 'algorithmic decision-making', and have caused both celebration and consternation [67, 43, 56].

However, such decisions have significant consequences for those individuals directly affected by them [31]. While human decision-makers may exhibit errors of judgement and biases, we can at least demand that they rationalise their decisions and thereby hold them accountable. A significant body of research in legal and organisational psychology suggests that people do not only care about whether the outcome of a decision benefits them, but also whether it meets standards of justice [71, 40, 11]. In particular, information about decision-making processes plays an important role in justifying the decisions that are made. In their basic form, algorithmic decision-making systems do not naturally provide information relevant to justice judgements, and therefore, threaten the capacity for accountability inherent in human decision-making [6]. In relying on such systems, we therefore run the risk of important decisions being made unaccountably.

The challenge to justice has long been recognised in contexts where data-driven decision-making is used in government administration [64]. The increasing use of predictive systems by both private and public bodies has prompted regulation aimed at rendering them more transparent. The European Union's General Data Protection Regulation (GDPR) (and similarly, the preceding 1995 Directive) requires organisations deploying certain systems to provide affected individuals with 'meaningful information about the logic' behind their outputs [81]. Similarly, the United States' Equal Credit Opportunity Act (ECOA) requires lenders to provide 'statements of reasons' when applications are denied loans [32].

These legal requirements give rise to HCI challenges concerning how the outputs of such systems should be communicated to affected individuals and other stakeholders. These challenges are not entirely new. The ability to provide information about how a system derives its predictions or classifications has long been recognised in the AI, expert systems, and HCI communities [8, 50, 16]. However, recent work has addressed the novel challenges of complex machine learning models, whose logic and outputs are otherwise harder to explain. So-called 'model-agnostic' approaches aim to explain the outputs of any classifier, regardless of the machine learning algorithm used to train it [62].

The potential for, as well as the limitations of, such approaches to address these regulatory requirements have been noted [20, 65, 82]. By implementing these explanation facilities within their algorithmic decision-making systems, organisations may be able to meet their obligation to provide 'meaningful information' about how their decisions are made. However, evaluation of the human factors of such systems has primarily tested their utility in supporting *expert* decision-making (for instance, in clinical

diagnostic settings [7]). The kind of meaningful information required by experts is likely to differ from that required by individuals affected by automated credit scores, insurance pricing or hiring decisions, for the purposes of accountability as envisaged by the relevant regulatory regimes.

This study aims to provide a preliminary exploration of the ways in which a range of novel explanation approaches might serve regulatory goals of rendering algorithmic decision-making more fair, accountable, and transparent. In particular, what effects do *explanations* have on people's perceptions of algorithmic decisions with respect to these regulatory aims? Do different explanation approaches meet these aims in different ways? To explore these questions, we conducted a series of studies, both lab-based and online, examining people's responses to a series of scenarios involving algorithmic decisions accompanied by different explanation styles inspired by recent work on fairness, accountability and transparency in machine learning.

BACKGROUND
The questions raised above are connected to multiple distinct areas of research, including: the psychology of justice perceptions regarding decision-making; the design of interpretable models in machine learning and HCI; and interdisciplinary work at the intersection of machine learning, law and social sciences.

Interpreting intelligent systems
The ability to interpret a system's outputs has long been a core theme of research into intelligent systems. As a general principle, systems should be able account for their own operation in ways that help users understand how their tasks are being accomplished [16, 66]. Early knowledge representation and reasoning systems typically aimed to meet this requirement by producing explanations for their outputs, for instance by presenting derivations from rules in rule-based architecture [69, 8], or graphical representations of nodes and edges in the case of Bayesian networks [37]. Explanation facilities could be further augmented by explicitly modeling the knowledge engineering and design process and revealing information about them to end-users [50]. Inspired by different theoretical models of argumentation, various explanation types have been explored and tested, with the style of explanation provided affecting the extent to which end-users understand the decisions of expert systems [46, 39].

Various research suggests that users—again, primarily experts—may in certain cases systematically over-rely on the outputs of intelligent systems [26, 19], while in others they may systematically distrust systems, despite reassurances as to their accuracy [57, 24]. However, the provision of explanations can affect levels of trust and acceptance of algorithmic decisions. In some cases, explanation increases trust and reliance [33, 85, 1, 30, 60], but in others an explanation may have the opposite effect if the level of detail it contains is deemed insufficient [7].

The increasing popularity of machine learning models as decision making tools presents novel challenges for explanation, as it may not be as straightforward to extract the embedded 'knowledge' behind their outputs [73]. This problem is particularly pronounced for complex models, like many-layered neural networks, but even simple decision tree models can quickly become uninterpretable if they have too many branches. As users form mental models [52]

of machine learning systems over time, such long-term exposure may lead to greater understanding [76] and greater trust and reliance [47]. But in the absence of transparency from platform providers, end-users may develop folk theories of algorithms, leading to a diverse range of strategies and countermeasures with varying degrees of effectiveness [22, 14].

Recent work on interpretable machine learning has focused on providing primarily *pedagogical* [73], *local* explanations for the outputs of ML models: pedagogical in the sense that the explanations teach something about how the model works rather than attempting to represent it directly, and 'local' in that they focus on how a specific output was derived. Approaches such as Local Interpretable Model-Agnostic Explanations (LIME) [62] and Quantitative Input Influence (QII) [13] enable an end-user to see a list of features which contributed to the output, along with the strength and direction of the contribution to the outcome class.[1] Case-based explanations provide examples from the model's training data which most closely resemble the output in question [12, 54], while 'demographic' explanations provide statistics on the outcome classes for relevant demographic variables [2].

While these different explanation *styles* suggest promising ways to explain the outputs of any classifier, explanation quality—or 'interpretability'—does not have a formal definition or a standard evaluation methodology shared amongst machine learning researchers [15, 41]. Explanation serves different functions in different contexts, and its evaluation is therefore context- and purpose-dependent. To this end, various human subjects experiments have been undertaken, exploring whether particular explanations result in better performance on a given end-task, in comparison to a no-explanation baseline. Contexts studied include aiding medical experts in diagnosis, increasing test performance of users of personalised learning environments, or enabling network security analysts to correctly identify incidents (e.g. [34, 83, 21]).

The use of various explanation styles has been been widely explored in the field of recommender systems. Tintarev and Mastho identify seven purposes for recommender system explanations, namely: transparency, scrutability, trust, effectiveness, persuasiveness, efficiency and satisfaction [74]. Explanations can also be used to *justify* or *describe* [80]. While early recommender system explanation approaches provided a uniform explanation style for single-source collaborative filtering [30], more recent work explores how to derive explanations for recommender systems based on 'hybrid' multiple sources [36] and matrix factorisation [61]. However, a reported need for explanation may not always correspond to differences in behaviour or performance; in a study of news recommender systems, end-users expressed a desire for explanations, but the number of news items they opened did not change when provided with reasons for their recommendations [70].

The use of local, pedagogical explanation facilities has not yet been tested in the contexts affected by the aforementioned emerging regulatory requirements, such as the automated evaluation of people for loans, hiring, and insurance. While various tools designed to help decision-makers identify and

---

[1]Layer-wise Relevance Propagation/Deep Taylor Expansion [45], decomposes the innards of deep learning systems to provide similar results to LIME (without directionality), and is also relevant here.

correct discrimination in data mining exist, these are aimed at data scientists implementing systems rather providing information to the individual decision-subjects affected by them [3, 59, 29, 23, 28]. The potential for local, pedagogical explanation systems to provide justice-related information, fulfilling the policy goals of transparency, accountability and fairness, has recently been noted by computer scientists and law scholars [20, 65, 82]. It has been suggested that organisations might rely upon these explanation facilities to fulfill legal duties to provide meaningful information about the logic of specific system outputs to affected individuals.

Perceptions of justice regarding decision-making

In order to explore how such explanation facilities might serve the aims of justice in algorithmic decision-making, a clearer understanding of the psychological aspects involved in justice perceptions is required. To this end, it is worth considering the extant literature on perceptions of justice regarding *human* decision-making. Decisions made about people with significant effects, such as court decisions, hiring, firing and promotion in a workplace, or the allocation of financial products, are often contentious and require a higher burden of accountability than other decisions; they should meet the standards of 'bureaucratic justice' [42]. Early work in this area emphasised that perceived levels of justice of a decision outcome are separate from purely self-serving rationalisations of a decision outcome; an individual might be negatively affected by a decision whilst still thinking it is just [71, 40].

According to Colquitt and others, justice perceptions can be broken down into several aspects or antecedents [10]. *Procedural* justice concerns the processes, logic and deliberation behind a decision. *Distributive* justice concerns the allocation of positive and negative outcomes in a decision context and whether they are distributed equitably or deservedly amongst the affected population given their circumstances, performance or contributions. *Interactional* justice concerns the extent to which the affected individual is treated with dignity and respect by the decision-makers. Finally, *informational* justice pertains to the information and explanations provided for decisions; are they candid, thorough, and tailored to individual needs? These aspects of justice are distinct but correlated [10]. Receiving a thorough explanation (informational justice) is important in helping people to assess whether the decision-making procedure is just (procedural justice) [10]. In turn, decisions perceived to be procedurally just are more likely to be perceived as distributively just [79]. Such findings provide a rationale for giving individuals rights to information about significant decisions; requiring decision-makers to explain their decisions should promote informational justice, which should help people assess procedural justice and distributive justice in turn.

This rationale might plausibly apply to the regulatory regime around algorithmic decision-making. Requiring organisations to explain the logic behind their algorithmic decision-making systems (informational justice) enables affected individuals to assess whether the logic of the system is just (procedural justice), which in turn might moderate their assessments of fairness of the decision outcomes (distributive justice). In so far as these notions of justice capture the aims of the regulatory requirements, they may provide appropriate ways of measuring the adequacy of different explanation systems for these purposes.

Key questions and contributions

This work builds upon and contributes to these disparate areas of research. Work on perceptions of justice reveals much about the role of explanations in ethical assessments of human decisions, but the applicability of these findings to algorithmic decisions and the range of proposed explanation systems proposed for them remains largely unexplored. A notable exception is Kizilcec, who studied the effect of levels of transparency on perceptions of procedural justice [35]. However the nuances of the different explanation approaches recently developed in interpretable ML have not yet been explored in relation to perceptions of justice. We also do not yet know much the extent to which findings on correlations between perceptions of justice from human decision-making contexts hold in the context of algorithmic decisions.

Furthermore, as research into practically achievable explanations for machine learning systems accelerates, researchers are turning to consider which explanation styles are more desirable, rather than simply possible to implement [41, 15]. This itself is a loaded question: an explanation style that artificially inflates certain notions of justice by playing on cognitive biases might suffer from the same anti-paternalist critiques levied against proponents of 'nudge' philosophy [44]. Understanding how explanation styles might influence and interact with justice perceptions is therefore an important preliminary to further development of ML explanation tools.

With these issues in mind, our key questions are:

1. How do explanations for algorithmic decisions affect justice perceptions regarding algorithmic decisions? In particular, do the positive correlations observed between informational, procedural and distributive justice in human decision-making settings also hold in algorithmic decision-making settings?

2. How do *different styles* of explanation affect such justice perceptions?

STUDY DESIGN AND METHODOLOGY

In investigating these questions, we combine several methodological aspects of the related work from justice perceptions and intelligent systems research. Many of the previously mentioned machine learning model explanation systems have not been subjected to user evaluations, and those which have often only test people's ability to predict a model's outputs, or answer questions about its logic (e.g. [39]). For present purposes, we are not concerned about how faithfully an explanation facility imparts actual model logic to a user.[2] Rather, we are interested in how the provision of various kinds of information about a model's outputs might affect justice perceptions as measured in the psychological literature. Much of this research is based on field surveys of individuals with personal experience of being affected by decisions (e.g. [79]), although controlled experiments involving simulated decision-making and scenario-based methods have also been used [78].

Given our focus on the effects of information provision and explanation styles on perceptions of fairness, we conducted a set of experimental studies to elicit people's responses to a range of

_____

[2]Although we acknowledge this is an important challenge if explanations are to be more than just comforting stories, as cautioned by Lipton [41].

plausible algorithmic decision scenarios and associated explanations. While this may lack the ecological validity of field surveys, it gives us the ability to engage people with a wide range of novel explanation styles covered in 2.1), and have yet to be implemented alongside real machine learning systems currently deployed 'in the wild' to any significant degree. By testing out proposed explanation styles with fictional scenarios, we hope that designers can gain insight into user concerns, and perhaps create more meaningful explanation interfaces to meet the upcoming legal requirements.

Scenarios and application contexts
We selected 5 application contexts in which to situate the fictional scenarios. The main criteria for context selection were: a relatively common interaction for the target population (a large proportion has experienced or will experience it at some point in their life); likely to involve machine learning models for algorithmic decision-making (at present or in the near future); have significant economic or practical effects on the decision-subject; and would possibly fall under the aforementioned GDPR regulation on automated decisions.[3] After consulting relevant literature on how ML systems are being applied in each context (e.g. [43, 38, 43, 55, 72]), we settled on the following contexts:

1. Applying for a personal financial loan;

2. Applying for a promotion at work;

3. Car insurance premiums dynamically priced based on personal details and driving behaviour measured by a telematic sensor (as described in e.g. [77]);

4. Passengers on over-booked airline flights being selected for re-routing;

5. Freezing of bank account due to activity suspected as associated with money laundering;

For each context, we created fictional cases in which an individual has had a decision made about them automatically. These were inspired by now-classic algorithmic 'war stories' [20] (see, e.g. [56, 58]). Scenarios and cases were adapted where necessary to fit the context from which the in-person participants would be recruited (a city in the United Kingdom). Because explanations are most likely to be requested in response to a decision which has a negative outcome for the individual, and considerations of procedural justice have been found to be more strongly felt in such cases [4], each scenario ended in a negative outcome for the individual.

Explanation styles
In order to derive a set of explanation styles to test, we reviewed a range of research, including both technical work on interpretable machine learning models, and legal discussions on the requirements of existing and forthcoming regulations on transparency of algorithmic decisions (e.g. [81, 65, 20]. Our aim was to find explanation styles which could plausibly meet or exceed the regulatory requirements regarding transparency of automated decisions, in particular the requirement that organisations provide 'meaningful information about the logic involved' in an automated decision (GDPR art 15(h)). We aimed to condense the many

_____
[3]The hurdle that a decision must be 'based *solely* on automated processing' is here read less restrictively than legal analysis has indicated [81, 20].

different proposals into a smaller set of explanation styles based on the kinds of information they would likely present to the end-user. We included only explanation methods which were model-agnostic; i.e. are applicable to any kind of learning algorithm, such as LIME [62], rather than methods which only work for particular ML models (e.g. those restricted to random forests [68] or neural networks [45]). This review resulted in 4 categories of promising explanation style:

1. **Input Influence**: Presents a list of input variables alongside a quantitative measure of their 'influence', positive or negative, on a decision (e.g. [13, 86])

2. **Sensitivity**: For each input variable used in a decision, sensitivity analysis shows how much the value of that variable would have to differ in order to change the output class (not to be confused with the notion of sensitivity used in ML evaluation) (e.g. [63, 65]).

3. **Case-based**: Presents a case from the model's training data which is most similar to the decision being explained (e.g. [18])

4. **Demographic**: Presents aggregate statistics on the outcome classes for people in the same demographic categories as the decision-subject, such as age, gender, income level or occupation (e.g. [2])

These explanation styles were tested and refined in a small-scale informal design phase. Each explanation system imparts different information, which is presented in various ways in the literature, including text [18], graphs [62] and bullet points [86]. In order to control for these differences in representation, we chose to use purely textual explanations. We informally tested each explanation style by varying the wording and information dimensions. We found that people generally did not pay attention to explanations independent of scenario descriptions. Through iterative testing we found that the use of dialog boxes focused attention on explanations (see Figure 2). These were deployed and investigated in depth in an in-person lab study and two online experiments. All phases of each study were approved by relevant University ethical review procedures.

Justice constructs
The justice constructs used in all three studies were based on those developed in the psychology of justice research, summarised by Colquitt et al [11]. Participants were asked to rate their agreement with five statements on a 5-point Likert scale (see statements at the bottom of Figure 1). Statements were drawn and adapted from previous studies of human decision-making to fit scenarios in which the decision was being made by an automated algorithmic system. Preliminary testing found that certain questions from Colquitt's scale, particularly those relating to *interactional* justice such as 'Have you been able to express your views and feelings during those procedures?' were difficult to interpret. This was partly due to the hypothetical nature of the scenarios, and partly due to being too human-specific to apply to a computer-based interaction. The remaining questions pertained to informational, procedural and distributive justice. Statement 2 (*understanding*) aimed to capture informational justice [27]; statements 3 and 4 (*appropriateness of factors, fair process*) pertained to procedural justice; and statement 5 (*deserved outcome*) addressed distributive

Figure 1. Example of a scenario (car insurance), case description (Sarah), explanation (Case-based), and questions

dimensions [11]. We also found that people tended to interpret these justice construct statements as a proxy for the simpler question of whether or not they *agreed* with the decision; to avoid this, an additional statement about agreement (1, *agreement*) was also included prior to the others.

Phase 1: Lab study

The first study consisted of in-person semi-structured interviews focused around a series of fictionalised cases as described above. The purpose of this study was to gain in-depth insights into how people interpret, evaluate and reason about algorithmic decision-making in the range of contexts and explanation styles mentioned above.

19 participants were recruited through paper flyers placed in various permitted locations in a small city in the UK (including cafés and restaurants, churches, libraries, shops, and museums), and through official social media accounts of the research lab and personal accounts of the researchers. Participants were then invited to the lab where the study occurred, which was set up with a standard PC keyboard and mouse, and a 27-inch display with a screen resolution of 1920x1200.

Participants were presented with 3 cases for each of the 5 contexts, (15 cases in total) via a web-based interface. Cases were presented with a brief description of the individual, the decision outcome, the explanation (except in the control condition) and an illustrative image depicting the context. The cases were used to prompt reflection on various aspects of the decision, and any explanatory information provided, using a concurrent think-aloud process [51].

Participants were asked to consider their agreement with the five measures described in section 3.3 and Figure 1. Agreement was indicated on a 5-point Likert scale, ranging from 'strongly agree' to 'strongly disagree'. Participants were asked to verbalise their responses to each case and explain their thought processes for each response. Each case was presented with one of the 4 explanation styles described above, or a control condition in which no explanation was provided. Participants were assigned pseudonymous identifiers to be associated with data generated during the experiment.

Audio recordings taken from the interviews and think-aloud processes were transcribed, segmented and labelled by each case and explanation style, and any personally identifying information was redacted. Thematic analysis was undertaken to identify common themes in participants' interpretation, evaluation and reasoning in response to cases. Thematic codes were independently developed by three coders, after which a single set was jointly agreed upon. The data were then re-coded using the new set of codes.

Phase 2: Online studies

Following on from the lab study, two online studies were conducted in order to generate quantitative data to test the following hypotheses:

1. Do different explanation styles result in differences in perceived levels of justice?

2. Do the positive correlations observed between informational, procedural and distributive justice in human decision-making settings also hold in algorithmic decision-making settings?

Both online studies consisted of pared-down versions of the protocols used in phase 1. Participants were recruited via the study platform Prolific Academic,[4] and filtered to include only individuals over 18 based in the UK, in order to maintain similarity with the participants recruited in phase 1.

A between-subjects design was used in order to test differences between responses under different explanation conditions. On the basis of described responses to the various explanation styles in the lab study, it was hypothesised that the ability to make direct comparisons between explanation styles might affect justice perceptions. We therefore also devised a follow-up within-subjects experiment, to test whether exposure to multiple explanation styles for a single case would have different effects than repeated exposure to a single explanation style.

Between-subjects study
325 participants were randomly assigned to 1 of 5 conditions (n = 65 in each group). Each group was presented with a selection of 12 cases used in experiment 1, accompanied by one of the 4 different explanation styles or a control condition featuring no explanation.[5] Participants were asked to indicate their agreement with the 5 measures described in phase 1 using the same Likert scale.

Within-subjects study
The within-subjects study involved 65 participants. Only the loan and insurance cases were included in this design, as the other

---

[4] www.prolific.ac
[5] The airline context was excluded at this stage, because experiment 1 had revealed that participants almost universally judged the cases to be very unfair and undeserved, thus preventing any useful conclusions to be drawn from this context.

Input influence–based explanation ☐	Demographic–based explanation ☐	Case–based explanation ☐	Sensitivity–based explanation ☐
Our predictive model assessed your personal information and driving behaviour in order to predict your chances of having an accident. The more +s or -s, the more positively or negatively that factor impacted your predicted chance of accidents. Unimportant factors are indicated. > Your age (---) > Driving experience (---) > Level of adherence to speed limit (-) > Number of trips taken at night (++) > Miles per month (+)	> 29% of female drivers qualified for the cheapest tier. > 31% of drivers in your age group [30–39] qualified for the cheapest tier. > 35% of drivers with 17 years of experience qualified for the cheapest tier. > 15% of drivers who have been on one accident which was not their fault qualified for the cheapest tier. > 26% of drivers who regularly travel at night qualified for the cheapest tier. > 21% of drivers who exceed the speed limit once ever two months qualified for the cheapest tier	This decision was based on thousands of similar cases from the past. For example, a similar case to yours is a previous customer, Claire. She was 38 years old with 18 years of driving experience, drove 850 miles per month, occasionally exceeded the speed limit, and 25% of her trips took place at night. Claire was involved in one accident in the following year.	> If 10% or less of your driving took place at night, you would have qualified for the cheapest tier. > If your average miles per month were 700 or less, you would have qualified for the cheapest tier.
ok!	**ok!**	**ok!**	**ok!**

Figure 2. Illustrative examples of the different explanation types presented to participants: input influence, sensitivity, demographic and case based.

scenarios could not be adapted to fit into a realistic hypothetical scenario in which four different explanations would be presented for a single case. Participants were therefore presented only with cases drawn from the Loan and Insurance contexts, in which the individual has received negative decision outcomes from 4 different lenders/insurers. Each decision was explained using a different style (in a random order), allowing participants to directly compare multiple styles applied to an individual subject.

Data from both studies were analysed with Spearman's rho correlations to understand if justice constructs correlate in the same way as reported in the aforementioned psychology of justice literature, and an ANOVA with Tukey's post-hoc paired tests to understand how explanation styles relate to each other in light of different justice constructs. Details on both methods are presented further below alongside the analysis.

RESULTS

Participant information

The lab study featured 19 participants, 11 male, 8 female, with an average age of 28.8 (range = 21-60, sd = 10.2). They had a range of educational attainment, comprising: high-school / A-levels (4), Bachelors (7), Masters (7) and PhD (1). In terms of familiarity with the decision-making contexts featured in the study, 6 had experienced applying for a loan, 6 had sought a car insurance deal, 5 had experience of promotion at work, 8 had had their bank accounts frozen, and 1 had been bumped off a flight. Average time to complete the study was 34 minutes (sd = 6.56).

For the between-subject online study, 64% were female, with an average age of 37.6 (r = 17-69, sd = 11.4); 48% were in full-time employment, 26% were unemployed, 21% were in part-time employment (3% listed 'other'). Education levels comprised high school (16%), A-level (27%), undergraduate degree (34%), Masters (16%), and Doctorate (0.4%). 25% had experienced applying for a loan, 34% had sought a car insurance deal, 26% had experience of promotion at work, 14% had had their bank accounts frozen, and 2% had been bumped off a flight. Average time to complete the study was 7.5 minutes (sd = 3.26).

For the within-subject online study, 64% were female, with an average age of 39.77 (range = 24-70, sd = 11.8); 66% were in full-time employment, 14% were unemployed, 19% were in part-time employment (1% listed 'other'). Education levels comprised high school (20%), A-level (18%), undergraduate

degree (42%), Masters (17%), and Doctorate (3%). 26% had experienced applying for a loan, 30% had sought a car insurance deal, 32% had experience of promotion at work, and 9% had had their bank accounts frozen. Average time to complete the study was 8.1 minutes (sd = 3.59).

Qualitative Results: Themes and reflections

The think-aloud responses from the lab study resulted in five major themes: the lack of human touch; interpretations of the system's reasoning; the use of statistical inference; the degree of actionability in an explanation; important aspects which were seen to be unaccounted for by the system; and the meaning and relevance of moral concepts. Many of these themes occurred across different explanation styles and scenario types; although some reflections were particularly concentrated around certain explanation styles and scenarios, which are noted below.

The (lack of) human touch
An initial reaction from many participants concerned the indignity or 'weirdness' of algorithmic decision-making compared to human decision-making. A series of comments reflected upon the ways in which the very use of an algorithmic decision-making system could be impersonal and dehumanising for the subject of the decision. In response to a case in which an employee had been automatically rejected for a promotion, participant *DC* remarked:

> '[sarcastically] This really makes you feel like a valued employee! When a machine or some kind of system makes a decision about someone based on some kind of data points, then it's impersonal' –DC

Others considered it problematic that with an automated system, 'there's no sense of negotiation' (*DR*), and no opportunity for 'human interaction' (*CS*). In contrast to this apparent lack of humanity (or perhaps to counteract it), some participants attempted to humanise or anthropomorphise the system [48], one participant for instance describing it as 'rude' (*OR*).

For the same participant, the specificity of the figures used in some explanations (particularly, *input influence* and *sensitivity*) lead her to conclude that the computer must be using arbitrary thresholds, which in itself seemed 'mean':

> 'Oh that's so mean! [...] I can't do the maths, but why is it so specific? Hmmm. I don't understand. I don't know why the cut-off is like that.' –OR

Such remarks suggest that if algorithmic decisions are held to norms of social behaviour, *interactional* justice may be a relevant dimension for evaluation.

Interpreting the system's 'reasoning'

Many of the reactions concerned attempts to decipher the logic of the algorithm. Some subjects made their assessments on the basis of whether the system approximated their own reasoning or knowledge:

> 'The computer [...] used the same sort of reasoning that I did I guess' –MP

> 'Yeah I understand this decision [denying a young male cheap car insurance], because young males are much more likely to have accidents' –PC

Others tried to articulate the system's reasoning as a series of rules and consequences:

> 'Breaking the speed limit is breaking the law, therefore you are proving that you are not capable of meeting the standards for the cheapest insurance, which is presumably: "thou shalt obey all the laws".' –AS

In some cases, while the individual premises seemed understandable, their relevance to the reasoning behind the overall decision was not:

> 'It seems like I can understand every one of them, but if put into the picture [...] each point is relevant, but I don't know [how] it puts into the big picture' –OR

Faced with seemingly inexplicable reasoning processes, some participants concluded that the system must be simply 'making it up' (*OR*) and therefore on the face of it unfair.

Acceptability of statistical inference

In addition to concerns about indignity and opaque reasoning, participants commonly reflected on the (un)acceptability of generalisation and statistical inference as a basis for prediction and decision-making in general. Issues raised in this regard were the importance of scientific rigour, including the use of sufficient sample sizes:

> 'I don't know how many previous customers they're basing it on...' –VI

> 'I'm gonna assume that it looked at more than just John and was like: 'You're like these people and these people didn't perform very well so we're not gonna give you a promotion" –MP

However, many other participants expressed strong objections to the use of statistical inferences to make judgements about people in itself:

> 'This is just simply reducing a human being to a percentage. It's not taking any of his actual ability, success or whatever into account. It's just saying 'tough luck', you can only expect to be successful 20 percent [of the time]' –UT

> 'This is saying your gender determines how you act, how you operate, and that's based on sheer probability.' –BF

> 'Perhaps it's unfair to make the decision by just comparing him to other people and then looking at the statistics, he isn't the same person.' [...] 'They [...] seem like [...] just random stats, not reasons for why you'd make a decision' –MP

Some emphasised that an individual who seems similar to past individuals may in fact have extenuating circumstances, particularly in response to *case-based* explanations:

> 'Just because Joel performed badly in the end doesn't mean that Jing [will] perform badly, because there could've been other influencing factors' –MP

In addition to comparisons between different individuals, participants also commented on the acceptability of judging a single individual's future behaviour on the basis of their past behaviour. However in this case, some participants were more comfortable with inductive inferences; *AS* agreed with the computer's decision to deny a loan applicant with existing debts because they had already 'proven not to be able to pay it back'.

Actionability

In judging the appropriateness of certain features, or the fairness of a decision, issues of personal responsibility and the possibility of acting differently were frequently raised. Some participants were happy to endorse negative decisions if they were able to identify a reasonable alternative course of action available to the participant which would resulted in a different outcome, as in *sensitivity*-based explanations:

> 'If he's not earning enough money for the company to feel they can give him a loan of 5,000 pounds, then that's fair. But they're also giving him an alternative in there; 'if you had an extra 2,000 pounds income, you'd have been successful', or 'if you'd asked to borrow [...] less, you'd have been successful'. Which is the right thing to do—reject him, but give him alternatives.' –MP

Where an individual was deemed at fault for risky or otherwise undesirable past behaviour, participant *AS* believed the decision was particularly justified:

> 'Drive more carefully [...] unless it is down to entirely unluckiness, you had a lot of accidents so you deserve it [being denied cheap insurance]' –AS

Conversely, where a judgement was made on the basis of circumstances which were outside the individual's control, or the course of action suggested was deemed unrealistic, the outcome was seen as undeserved:

> 'I don't think she could have done anything [...] she didn't make any poor decisions that led to this, so I somewhat disagree that she deserved it' –PC

> 'You don't choose where you're born, you don't choose your gender.' –BF

Several responses suggested that explanations ought to help guide future action, something which was particularly lacking in the *demographic* explanations:
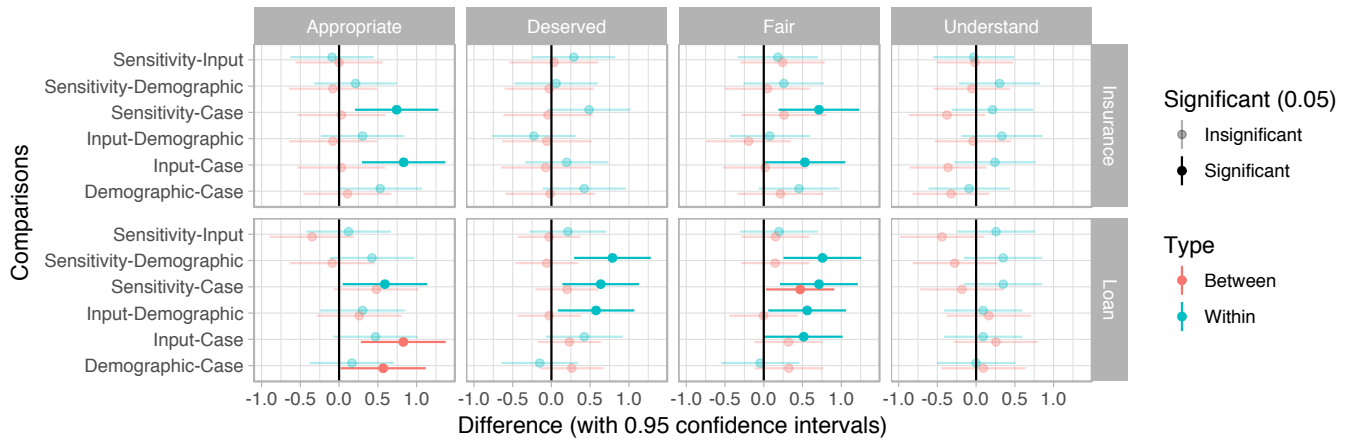
Figure 3. Tukey's post-hoc paired tests showing comparative differences between answers in different scenarios by question (horizontal facet), scenario (vertical facet), study design (color) and explanation type (y-axis). Insignificant results are displayed in lower transparency for convenience.

'What do you do with '1.5% of women have had their accounts frozen'? That gives you nothing, no information you can do anything with' –PC

Although participant *CS* was concerned that behaviour change in response to an algorithm could prove dangerous

'If you know it's on the back of an algorithm, it would incentivise people to work out how to game the algorithm, to find out what the algorithm is exactly doing' –CS

Unaccounted aspects
Many criticisms of the decisions came down to things that were missing or unaccounted for. For some, it was the absence of important information:

'It should also base it on his credit score as well' –VD

'I think the bank ought to know a bit more about you' –DR

'We have no idea why John wasn't able to pay off the loan, he may have suddenly been made suddenly terminally ill' –UT

In some cases, what was missing was not just information, but also complexity:

'If you just ask them more questions you could find out they're fine [...] or just make the algorithm more complicated' –DR

In others, the missing element was a sense of proportionality. Responding to an *input influence* explanation for denial of a loan, VI remarked:

' I don't know why [...] why he's got a minus 7 for existing debts. I think the information was relevant but I think it's kinda out of proportion' –VI

Meaning and relevance of moral concepts
The meaning and relevance of moral concepts like 'fairness' and 'desert' in the context of a computer system was occasionally questioned by certain participants. Participant *PC* struggled to evaluate the moral dimensions of computer-driven decisions:

'These last two questions ['Was the decision-making process fair?', 'Did the individual deserve this outcome?'] are hard to answer... You can't really make a judgement about a computer, if you tell it to do something and it just does it.' –PC

For *DR*, here was a sense that the very notion of 'fairness' may not be attributable to a system that is primarily designed to serve a different organisational goal like efficiency:

'Fair in this sense hasn't really come into it, this is like a cold decision [...] It's not fair at all really, but it's understandable from the perspective of the business' –DR

One participant argued that a truly random system could also be seen as *more* 'fair', at least for certain definitions of fairness:

'I suppose the fair thing would be to just do it completely randomly [...] like a Harvey Dent definition of fair—from Batman, you know, Two Face?—"You know what's fair? Chaos is fair!"' –DR

Quantitative results
To recap from the methodology, 325 participants from the *Prolific Academic* platform were presented with different scenarios and different explanation styles, and asked to respond to questions that related to them on 5-point Likert scales.

Do justice correlations from human decision-making settings hold in relation to algorithmic decisions?
According to the literature on justice perceptions above, we should expect to find correlations between *fair process* and *deserved outcome*, *understanding* and *fair process*, *appropriateness of factors* and *deserved outcome*, and *appropriateness of factors* and *fair process* [10]. We were not expecting strong links between *appropriateness of factors* and *understanding* (as with a high level of understanding, one might find the factors inappropriate) or *understanding* and *deserved outcome* (as we suspect that judgments about deserved outcomes to be more dependent on inputs than on processes). To examine this, we calculated Spearman's rho correlation coefficients (with 0.05 confidence level bootstrapped intervals). The results (see Table 1) partially confirmed our hypotheses, as the stronger correlations were

connected to the construct connections we believed would exist. Surprisingly, we also found significant positive correlations between the two construct connections we were not expecting.

Do different explanation styles result in differences in perceived levels of justice?

In order to examine contexts in which a decision subject is faced with a *single* explanation style per decision, we conducted a between-subjects study. To examine contexts in which subjects are faced with *multiple* explanation styles per decision, we conducted a within-subject study (like the qualitative study), which featured multiple explanation styles per decision.

We used ANOVA tests, followed by Tukey's post-hoc paired tests [75], to analyse the relative effects of different explanation styles within different study designs.[6]

In the between-subjects study, explanation styles generally did not significantly affect justice perceptions, with the exception of fair process in loans only (loan $F(3, 258) = 2.71$, $p = 0.046$) and appropriate factors ($F(3, 258) = 5.35$, $p = 0.001$) showing significant between group differences.

In the within-subjects study—where multiple explanation styles were presented for the same decision—significant effects were observed across all questions and decision-making contexts. There was a significant effect of explanation styles on perceptions of fair process (loan scenario $F(3, 260) = 7.52$, $p < .001$; insurance scenario $F(3, 260) = 4.5$, $p = .004$), on the perception of appropriate use of factors (loan $F(3, 260) = 3.312$, $p = .02$; insurance $F(3, 260) = 6.44$, $p = .0003$), and on the perception of deserved outcome (only in the loans case, $F(3, 260) = 7.31$, $p = .0001$).

Tukey's post-hoc paired tests (see Figure 3) showed that case-based explanations result in lower perceptions of appropriateness, fair process perception, and (in the loans case) deservedness, consistently compared to sensitivity based styles and occasionally compared to other styles. This is an effect primarily observed, like most effects in the quantitative part of our study, in the *within* subject study design, indicating that the act of comparison in a particular scenario is important for these differences to become apparent. Case-based explanations seem to have the most consistent negative impact on justice perceptions when presented alongside alternative explanation styles.

Q1	Q2	lower	rho	upper
Fair	Deserved	0.63	0.69	0.74
Fair	Appropriate	0.53	0.60	0.66
Appropriate	Deserved	0.43	0.51	0.58
Fair	Understand	0.32	0.40	0.49
Appropriate	Understand	0.31	0.40	0.48
Understand	Deserved	0.28	0.36	0.44

Table 1. Spearman's rho correlations for justice constructs with bootstrapped (1000 rep) confidence interval at 0.02 level.

DISCUSSION

Faced with a range of algorithmic decision-making scenarios, people appear conflicted about how to respond to questions of justice.

---

[6]Some readers might have reservations over parametric tests being applied to ordinal data: we point to recent literature that has emphasised [53] that Likert data are robust to parametric assumptions, and that 'parametric methods can be utilized without concern for "getting the wrong answer"'.

For some participants, the *very idea* of an algorithmic system making an important decision on the basis of past data seemed unfair;

> 'She's been a victim of this computer system that has to generalise based on, like, somebody else ... what it should be looking for is ability to pay back the loan, not characteristics of others who couldn't pay back the loan' -WP

Whereas for others, delegation of such responsibility to a computer, and the authority of statistical models, actually removed the relevance of questions of fairness altogether; the system 'just does what it is supposed to', and as long as its inferences are accurate, it could be seen as 'statistically fair' (*PC*).

In this sense, while algorithmic decision-making implicates dimensions of justice, it may also mitigate them, in so far as people resist imputing morality to a computer system [25, 5]. On the other hand, people frequently interact with computers as if they were social actors [49]; why stop when justice is involved? One participant described how she might have it both ways, simultaneously entertaining dual standards of justice, one juridical, the other computational;

> 'I would say I agree he deserved this outcome given the way this computer model has judged his behaviour, whether he actually deserved this outcome remains to be seen. If we were in a court of law, I would argue we don't know his circumstances, but given this computer model and the way it works it's deserved.' -WP

Such attempts to delineate moral culpability between *systems* and *people*, or between different standards of fairness or evidence, may reflect a more general psychological need to see the world as morally coherent [9]. The introduction of algorithmic systems seems likely to shake up previously stable categories and concepts, requiring adjustment and incorporation into one's moral world-view.

In terms of the value of providing explanations, and the relative importance of different explanation styles, the picture is complicated. On the one hand, people strongly engaged with the details of each explanation when discussing each case. This was borne out in the within-subjects experiment, where significant differences were observed in justice perceptions between different explanation styles—in particular that case-based explanation styles impacted negatively on justice-related judgements especially compared to sensitivity-based explanation styles. And yet, in the between-subjects design where individuals were exposed to only one explanation style repeated across multiple cases, these explanation effects largely disappeared.

One possible explanation is that when people repeatedly see the same explanation style, they become habituated and pay more attention to the features of the cases rather than any specific features of the explanation provided. When exposed to multiple explanation styles regarding the same case, the differences may become more salient. While one-off interactions with individual systems seem most likely, both scenarios—comparison of multiple styles, and exposure to a single style—may occur in different real world contexts. For instance, people may realistically face multiple explanations for similar algorithmic decisions when applying for a loan, mortgage or credit. Various explanation tools already exist

within the credit scoring industry, and a savvy consumer might well test them all out when preparing for a significant application.

The relative consistency in justice-related judgements between different explanation styles might also be explained by strongly-held intuitions about the variables used in a decision. While different explanation styles represented the input variables in different ways, the case descriptions all featured the same set of inputs. If people have strong intuitions about the fairness of using particular features in a model (regardless of their weights), as has been observed in a previous study by Grgic et al [28], this may account for such consistency. Another important factor may be the extent to which experimental design allows interpersonal comparison. Research on justice judgements in non-automated contexts has found that individuals pay more attention to other people's outcomes than they do to information about the procedure behind the decision that affects them [78, 84].

Finally, the observed correlations between different justice constructs fall largely in line with those observed within the psychology of justice literature amassed since the 1970s. Procedural and distributive dimensions had a relatively high correlation coefficient of 0.69; for comparison, a 2001 meta-analysis of 45 studies in a variety of organisational and institutional settings reported a correlation coefficient of 0.48 for the same constructs [10].

LIMITATIONS AND FUTURE WORK

While the combination of different methods employed in these three studies goes some way to addressing the problems associated with any single approach, there are a number of limitations which are important to mention.

Threats to validity

First, the samples were not representative of the general population. Participants for the in-person lab study were all drawn from an affluent and academically dominated city, and from the UK where experiences of certain decision contexts (e.g. financial loans) likely differ to other countries (e.g. those with greater or lesser dependence on private debt). Participants in the online studies were not gender balanced, with almost double the number of females to males. Second, the scenarios considered were hypothetical, not affecting the participants directly, and therefore lacked the first-person consequences and significance of a real world decision, as well as the possibility of alternative courses of action such as requesting a human review or appealing the decision. The explanations were also not drawn from real machine learning model outputs due to our desire to understand explanations in the context of rejections and not to introduce confounding elements—but in practice, this might introduce additional complicated factors. Finally, comparisons were made between four different explanation styles, but further confirmatory work may be needed to address the specific reasons for differences between explanation styles. In particular, the explanation styles proposed in the interpretable machine learning literature vary in the information they contain. Despite our efforts to standardise them, differences in the length and format of the text required to impart an explanation between explanation conditions may have been a confounding factor.

Future work

A number of opportunities for future work arise from this discussion and limitations. One would be to consider the application of other possible measures and methods from the psychology of justice research, such as interactional justice. Indeed, some participants' remarks (e.g. about the computer being 'rude') suggest that there may be useful notions of interactional justice even when humans are not directly involved in communicating the decision. Another would be to examine how to design systems designed to make ML outputs interpretable in different ways to multiple end-users for different ends; not only the data scientist attempting to detect and mitigate discriminatory effects (e.g. [3]), but also other stakeholders such as third party auditors. This is part of a wider need to consider how to re-engineer better user experiences for multiple stakeholders interacting with machine learning systems in general [17].

CONCLUSION

Algorithmic decisions are likely to become increasingly relied on for a range of decisions with potentially important repercussions for those affected. Understanding how people assess the fairness of such decisions, and how explanations might help, is therefore of increasing significance. Despite repeated calls for more transparency over how such decisions are made, there is still much to learn about what people want and need to know about algorithms in order to hold them accountable to justice. As lawmakers legislate for mandatory provision of information to decision-subjects, human-computer interaction research has much to offer in how such information should be extracted, presented and delivered.

This paper suggests that people do consider justice-related aspects of algorithmic decision-making systems, much as they do for manual decision-making processes. However, depending on how and when they are deployed, explanations may or may not help individuals to evaluate the fairness of such decisions. The algorithmic nature of these systems results in an array of novel considerations which are not captured by traditional research on perceptions of justice. Conversely, creators of ML explanation systems have not typically designed them with the information needs of those individuals facing significant personal consequences of model outputs in mind. It is our hope that such concerns will instigate renewed focus on this range of important use cases for algorithmic explanations, and more broadly for HCI research to support the pursuit of justice as algorithmic decision-making systems take hold in a wide array of high-stakes domains.

REFERENCES

1. Stavros Antifakos, Nicky Kern, Bernt Schiele, and Adrian Schwaninger. 2005. Towards improving trust in context-aware systems by displaying system confidence. In *Proceedings of the 7th International Conference on Human Computer Interaction with Mobile Devices & Services.* ACM, 9–14.

2. Liliana Ardissono, Anna Goy, Giovanna Petrone, Marino Segnan, and Pietro Torasso. 2003. Intrigue: personalized recommendation of tourist attractions for desktop and hand

held devices. *Applied Artificial Intelligence* 17, 8-9 (2003), 687–714.

3. Bettina Berendt and Sören Preibusch. 2014. Better decision support through exploratory discrimination-aware data mining: foundations and empirical evidence. *Artificial Intelligence and Law* 22, 2 (2014), 175–209.

4. Joel Brockner and Batia M Wiesenfeld. 1996. An integrative framework for explaining reactions to decisions: interactive effects of outcomes and procedures. *Psychological bulletin* 120, 2 (1996), 189.

5. Joanna J Bryson. 2010. Robots should be slaves. *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues* (2010), 63–74.

6. Jenna Burrell. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society* 3, 1 (2016), 2053951715622512.

7. Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *Healthcare Informatics (ICHI), 2015 International Conference on*. IEEE, 160–169.

8. William J Clancey. 1983. The epistemology of a rule-based expert system — a framework for explanation. *Artificial intelligence* 20, 3 (1983), 215–251.

9. Cory J Clark, Eric Evan Chen, and Peter H Ditto. 2015. Moral coherence processes: Constructing culpability and consequences. *Current Opinion in Psychology* 6 (2015), 123–128.

10. Jason A Colquitt, Donald E Conlon, Michael J Wesson, Christopher OLH Porter, and K Yee Ng. 2001. Justice at the Millennium: A Meta-Analytic Review of 25 Years of Organizational Justice Research. *Journal of Applied Psychology* 86, 3 (2001). DOI:`http://dx.doi.org/10.1037//0021-9010.86.3.425`

11. Jason A Colquitt and Jessica B Rodell. 2015. Measuring justice and fairness. In *Oxford Handbook of Justice in the Workplace*, Russell Cropanzano and Maureen L Ambrose (Eds.). Oxford University Press, Oxford, 187–202.

12. Pádraig Cunningham, Dónal Doyle, and John Loughrey. 2003. An evaluation of the usefulness of case-based explanation. *Case-Based Reasoning Research and Development* (2003), 1065–1065.

13. Anupam Datta, Shayak Sen, and Yair Zick. 2016. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *Security and Privacy (SP), 2016 IEEE Symposium on*. IEEE, 598–617.

14. Michael A DeVito, Darren Gergle, and Jeremy Birnholtz. 2017. Algorithms ruin everything: #RIPTwitter, Folk Theories, and Resistance to Algorithmic Change in Social Media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, CHI'17*. 3163–3174.

15. Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. (2017). `https://arxiv.org/abs/1702.08608`

16. Paul Dourish. 1997. Accounting for system behaviour: Representation, reflection and resourceful action. *Computers and Design in Context, MIT Press, Cambridge, MA, USA* (1997), 145–170.

17. Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. 2017. UX Design Innovation: Challenges for Working with Machine Learning as a Design Material. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, CHI'17*. ACM, 278–288.

18. Donal Doyle, Alexey Tsymbal, and Padraig Cunningham. 2003. *A Review of Explanation and Explanation in Case-Based Reasoning (TCD-CS-2003-41)*. Trinity College Dublin, Dublin.

19. Mary T Dzindolet, Scott A Peterson, Regina A Pomranky, Linda G Pierce, and Hall P Beck. 2003. The role of trust in automation reliance. *International Journal of Human-Computer Studies* 58, 6 (2003), 697–718. DOI:`http://dx.doi.org/10.1016/S1071-5819(03)00038-7`

20. Lilian Edwards and Michael Veale. 2017. Slave to the Algorithm? Why a 'Right to an Explanation' is Probably Not the Remedy You Are Looking For. *Duke Law & Technology Review* 16, 1 (2017), 18–84. DOI:`http://dx.doi.org/10.2139/ssrn.2972855`

21. Kate Ehrlich, Susanna E Kirk, John Patterson, Jamie C Rasmussen, Steven I Ross, and Daniel M Gruen. 2011. Taking advice from intelligent systems: The double-edged sword of explanations. In *Proceedings of the 16th International conference on Intelligent User Interfaces*. ACM, 125–134.

22. Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. 2016. First I "like" it, then I hide it: Folk theories of social feeds. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, CHI'16*. ACM, 2371–2382.

23. Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. 259–268. DOI:`http://dx.doi.org/10.1145/2783258.2783311`

24. Robert Fildes, Paul Goodwin, Michael Lawrence, and Konstantinos Nikolopoulos. 2009. Effective forecasting and judgmental adjustments: An empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting* 25, 1 (2009), 3–23.

25. Batya Friedman and Peter H Kahn. 1992. Human agency and responsible computing: Implications for computer system design. *Journal of Systems and Software* 17, 1 (1992), 7–14.

26. Kate Goddard, Abdul Roudsari, and Jeremy C Wyatt. 2011. Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators. *Journal of the American Medical Informatics Association* 19, 1 (2011), 121–127.

27. Jerald Greenberg and R Cropanzano. 1993. The social side of fairness: Interpersonal and informational classes of organizational justice. *Justice in the workplace: Approaching fairness in human resource management, Lawrence Erlbaum Associates, Hillsdale, NJ* (1993).

28. Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. 2016. The case for process fairness in learning: Feature selection for fair decision making. In *NIPS Symposium on Machine Learning and the Law, Barcelona, Spain, 8 December 2016*.

29. Sara Hajian and Josep Domingo-Ferrer. 2012. Direct and indirect discrimination prevention methods. In *Discrimination and Privacy in the Information Society*, Bart Custers, Toon Calders, Bart Schermer, and Tal Zarsky (Eds.). Springer, Berlin, Heidelberg, 241–254.

30. Jonathan L Herlocker, Joseph A Konstan, and John Riedl. 2000. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work*. ACM, 241–250.

31. Mireille Hildebrandt and Serge Gutwirth. 2008. *Profiling the European Citizen*. Springer.

32. David C Hsia. 1978. Credit scoring and the equal credit opportunity act. *Hastings LJ* 30 (1978), 371.

33. Hilary Johnson and Peter Johnson. 1993. Explanation facilities and interactive systems. In *Proceedings of the 1st International Conference on Intelligent User Interfaces*. ACM, 159–166.

34. Been Kim, Caleb M Chacha, and Julie A Shah. 2015. Inferring Robot Task Plans from Human Team Meetings: A Generative Modeling Approach with Logic-Based Prior. *Journal of Artificial Intelligence Research* 52 (2015), 361–398.

35. René F Kizilcec. 2016. How much information?: Effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2390–2395.

36. Pigi Kouki, James Schaffer, Jay Pujara, John OâĂŹDonovan, and Lise Getoor. 2017. User Preferences for Hybrid Explanations. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. ACM, 84–88.

37. Carmen Lacave and Francisco J Díez. 2002. A review of explanation methods for Bayesian networks. *The Knowledge Engineering Review* 17, 2 (2002), 107–127.

38. Stefan Lessmann, Bart Baesens, Hsin-Vonn Seow, and Lyn C Thomas. 2015. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research* 247, 1 (2015), 124–136.

39. Brian Y Lim, Anind K Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2119–2128.

40. E Allan Lind and Tom R Tyler. 1988. *The social psychology of procedural justice*. Springer Science & Business Media.

41. Zachary C Lipton. 2016. The Mythos of Model Interpretability. In *2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)*. New York. https://arxiv.org/abs/1606.03490

42. Jerry L Mashaw. 1985. *Bureaucratic justice: Managing social security disability claims*. Yale University Press.

43. Viktor Mayer-Schönberger and Kenneth Cukier. 2013. *Big Data: A Revolution That Will Transform How We Live, Work and Think*. John Murray, London.

44. Gregory Mitchell. 2004. Libertarian Paternalism is an Cxymoron. *Nw. UL Rev.* 99 (2004), 1245.

45. Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. 2017. Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognition* 65 (2017), 211–222. DOI: http://dx.doi.org/10.1016/j.patcog.2016.11.008

46. Johanna D Moore and William R Swartout. 1991. A reactive approach to explanation: Taking the user's feedback into account. In *Natural Language Generation in Artificial Intelligence and Computational Linguistics*, Cecile Paris, William R. Swartout, and William C. Mann (Eds.). Springer, 3–48.

47. Bonnie M Muir. 1994. Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics* 37, 11 (1994), 1905–1922.

48. Clifford Nass, Jonathan Steuer, Ellen Tauber, and Heidi Reeder. 1993. Anthropomorphism, agency, and ethopoeia: computers as social actors. In *INTERACT'93 and CHI'93 Conference Companion on Human Factors in Computing Systems*. ACM, 111–112.

49. Clifford Nass, Jonathan Steuer, and Ellen R Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 72–78.

50. Robert Neches, William R. Swartout, and Johanna D. Moore. 1985. Enhanced maintenance and explanation of expert systems through explicit models of their development. *IEEE Transactions on Software Engineering* 11 (1985), 1337–1351.

51. Janni Nielsen, Torkil Clemmensen, and Carsten Yssing. 2002. Getting access to what goes on in people's heads?: reflections on the think-aloud technique. In *Proceedings of the Nordic Conference on Human-Computer Interaction*. ACM, 101–110.

52. Donald A Norman. 1983. Some observations on mental models. *Mental Models* 7, 112 (1983), 7–14.

53. Geoff Norman. 2010. Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education* 15, 5 (2010), 625–632. DOI: http://dx.doi.org/10.1007/s10459-010-9222-y

54. Conor Nugent and Pádraig Cunningham. 2005. A case-based explanation system for black-box systems. *Artificial Intelligence Review* 24, 2 (2005), 163–178.

55. Cathy O'Neil. 2017. *Weapons of math destruction: How big data increases inequality and threatens democracy.* Broadway Books.

56. Cathy O'Neill. 2016. How algorithms rule our working lives. *The Guardian* (2016).

57. Dilek Önkal, Paul Goodwin, Mary Thomson, Sinan Gönül, and Andrew Pollock. 2009. The relative influence of advice from human experts and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making* 22, 4 (2009), 390–409.

58. Frank Pasquale. 2015. Digital Star Chamber. *Aeon* (2015). `https://aeon.co/essays/judge-jury-and-executioner-the-unaccountable-algorithm`

59. Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-aware Data Mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '08)*. ACM, New York, NY, USA, 560–568. `DOI: http://dx.doi.org/10.1145/1401890.1401959`

60. Pearl Pu and Li Chen. 2006. Trust building with explanation interfaces. In *Proceedings of the 11th International Conference on Intelligent User Interfaces*. ACM, 93–100.

61. Bashir Rastegarpanah, Mark Crovella, and Krishna P Gummadi. 2017. Exploring Explanations for Matrix Factorization Recommender Systems. (2017).

62. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1135–1144.

63. Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. 2016. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems* (2016).

64. Paul Schwartz. 1991. Data processing and government administration: the failure of the American legal response to the computer. *Hastings LJ* 43 (1991), 1321.

65. Andrew D. Selbst and Solon Barocas. forthcoming. Regulating Inscrutable Systems. On file with authors. (forthcoming).

66. Ben Shneiderman and Pattie Maes. 1997. Direct manipulation vs. interface agents. *interactions* 4, 6 (1997), 42–61.

67. Tom Simonite. 2017. Machine Learning for Everyone. *MIT Tech Review* (2017). `https://www.technologyreview.com/s/600987/machine-learning-for-everyone/`

68. Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics* 8, 1 (2007), 1. `DOI: http://dx.doi.org/10.1186/1471-2105-8-25`

69. Randy L Teach and Edward H Shortliffe. 1981. An analysis of physician attitudes regarding computer-based clinical consultation systems. *Computers and Biomedical Research* 14, 6 (1981), 542–558.

70. Maartje ter Hoeve, Mathieu Heruer, Daan Odijk, Anne Schuth, and Maarten de Rijke. 2017. Do News Consumers Want Explanations for Personalized News Rankings?. In *FATREC Workshop on Responsible Recommendation Proceedings*.

71. John Thibaut, Laurens Walker, Stephen LaTour, and Pauline Houlden. 1973. Procedural justice as fairness. *Stan. L. Rev.* 26 (1973), 1271.

72. Lyn Thomas, Jonathan Crook, and David Edelman. 2017. *Credit scoring and its applications*. SIAM.

73. Alan B Tickle, Robert Andrews, Mostefa Golea, and Joachim Diederich. 1998. The truth will come to light: Directions and challenges in extracting the knowledge embedded within trained artificial neural networks. *IEEE Transactions on Neural Networks* 9, 6 (1998), 1057–1068.

74. Nava Tintarev and Judith Masthoff. 2015. Explaining recommendations: Design and evaluation. In *Recommender Systems Handbook*, F. Ricci, L. Rokach, B. Shapira, and P.B. Kantor (Eds.). Springer, 353–382.

75. John W Tukey. 1949. Comparing individual means in the analysis of variance. *Biometrics* (1949), 99–114. `DOI: http://dx.doi.org/10.2307/3001913`

76. Joe Tullio, Anind K Dey, Jason Chalecki, and James Fogarty. 2007. How it works: a field study of non-technical users interacting with an intelligent system. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 31–40.

77. Giovanni Vaia, Erran Carmel, William DeLone, Harald Trautsch, and Flavio Menichetti. 2012. Vehicle Telematics at an Italian Insurer: New Auto Insurance Products and a New Industry Ecosystem. *MIS Quarterly Executive* 11, 3 (2012).

78. Kees Van den Bos, E Allan Lind, Riël Vermunt, and Henk AM Wilke. 1997. How do I judge my outcome when I do not know the outcome of others? The psychology of the fair process effect. *Journal of Personality and Social Psychology* 72, 5 (1997), 1034.

79. Kees Van den Bos, Lynn Van der Velden, and Allan Lind. 2014. On the role of perceived procedural justice in citizens' reactions to government decisions and the handling of conflicts. *Utrecht Law Review* 10, 4 (2014). `DOI: http://dx.doi.org/10.18352/ulr.287`

80. Jesse Vig, Shilad Sen, and John Riedl. 2009. Tagsplanations: explaining recommendations using tags. In *Proceedings of the 14th International Conference on Intelligent User Interfaces*. ACM, 47–56.

81. Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. 2017. Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. *International Data Privacy Law* 7, 2 (2017), 76–99. `DOI: http://dx.doi.org/10.1093/idpl/ipx005`

82. Adrian Weller. 2017. Challenges for Transparency. In *2017 ICML Workshop on Human Interpretability in Machine Learning (WHI 2017)*. New York. `https://arxiv.org/abs/1708.01870`

83. Joseph Jay Williams, Juho Kim, Anna Rafferty, Samuel Maldonado, Krzysztof Z Gajos, Walter S Lasecki, and Neil Heffernan. 2016. Axis: Generating explanations at scale with learnersourcing and machine learning. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale*. ACM, 379–388.

84. Lan Xia, Kent B Monroe, and Jennifer L Cox. 2004. The price is unfair! A conceptual framework of price fairness perceptions. *Journal of Marketing* 68, 4 (2004), 1–15.

85. L Richard Ye and Paul E Johnson. 1995. The impact of explanation facilities on user acceptance of expert systems advice. *MIS Quarterly* (1995), 157–172.

86. Jiaming Zeng, Berk Ustun, and Cynthia Rudin. 2017. Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 180, 3 (2017), 689–722.