Last author version of

Berendt, B. & Preibusch, S. (2017). Toward accountable discrimination-aware data mining: The importance of keeping the human in the loop – and under the looking-glass. *Big Data*, 5(2). DOI: 10.1089/big.2016.0055. © Mary Ann Liebert, Inc.

**Toward accountable discrimination-aware data mining:
The importance of keeping the human in the loop – and under the looking-glass**

Bettina Berendt¹ and Sören Preibusch²

¹ Department of Computer Science, KU Leuven, Leuven, Belgium;

bettina.berendt@cs.kuleuven.be (corresponding author)

² formerly Microsoft Research, Cambridge, UK;

mail@soeren-preibusch.de

Abstract

“Big Data” and data-mined inferences are affecting more and more of our lives, and concerns about their possible discriminatory effects are growing. Methods for discrimination-aware data mining and fairness-aware data mining aim at keeping decision processes supported by information technology free from unjust grounds. However, these formal approaches alone are not sufficient to solve the problem. In the present article, we describe reasons why discrimination with data can and typically does arise through the combined effects of human and machine-based reasoning, and argue that this requires a deeper understanding of the human side of decision making with data mining. We describe results from a large-scale human-subjects experiment that investigated such decision making, analysing the reasoning that participants reported during their task to assess whether a loan request should or would be granted. We derive data protection by design strategies for making decision making discrimination-aware in an accountable way, grounding these requirements in the accountability principle of the EU General Data Protection Regulation, and outline how their implementations can integrate algorithmic, behavioural, and user-interface factors.

Keywords: Discrimination discovery and prevention, Discrimination-aware and fairness-aware data mining, Accountability, Data protection, Evaluation, User studies

1 Introduction

“There’s software used across the country to predict future criminals. And it’s biased against blacks”: This (sub)title of a recent article¹ summarises the concerns that more and more citizens, authorities, and also the computer scientists who develop predictive methods have about computerised decisions: their potential to create unlawful or unwanted discrimination.

Discriminatory phenomena linked to software are not limited to the justice system, but have also been found in domains as diverse as university admissions², search-engine results and advertisements³ or waiting times for Uber taxis⁴. Substantial biases can also arise in employment-related decisions⁵. Effects are also not limited to discrimination based on ethnicity, but relate to gender^{6,7,2} or poverty⁸.

Freedom from discrimination is recognised as a right throughout jurisdictions. The European Union (EU) has explicitly included the protection against discrimination based on data processing into the General Data Protection Regulation (GDPR), its new data protection law to come into effect in 2018. As stated in Recital 71, data controllers should “implement technical and organisational measures” to prevent that processing has, or results in “discriminatory effects on natural persons on the basis of racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation”. The challenge for data scientists is how to best support this goal.

Computer scientists have proposed formal measures of (*non-*)*discrimination* or *fairness*^{9,10,11} a and modifications to data and/or algorithms to avoid or minimise discrimination (from the seminal work of Pedreschi et al.¹² to a rich body of literature described in recent surveys^{13,14}). We will refer to these methods as DADM (discrimination-aware data mining).

However, in many decision situations, software-generated rankings and predictions are only one component in decision making. While “the software” may be “biased”, it is not the software (“it”) that discriminates. Rather, it is the ensembles of algorithms, data, software implementing the operations of these algorithms and used on these data, humans, and larger dynamic systems that leads to discriminatory outcomes. Research that investigates and unpacks these complex ensembles is only just beginning. The present paper is part of a larger research effort at such investigation drawing on an interdisciplinary methodology.

In the first, empirical, part of this paper we present a user study that takes a detailed look at human decision processes in semi-automated decision making, as they relate to discrimination. Decision processes differ by domain, frequency and severity; it is therefore imperative to focus on concrete decision problems. We focus on a routine task that nonetheless can significantly affect individuals: the granting (or not) of a loan in retail banking. We analyse so-far-unexplored data from an experimental study with 215 participants¹⁵ who were given a fictitious data-mining

a The term “non-discrimination” tends to be used in EU and “fairness” in US contexts. In EU law including data protection law, “fairness” carries the connotations of “a fair deal”, “according to expectations”, etc., and hereby concerns the relationship between the data subject and the data controller rather than a comparison between the way in which different data subjects are treated. Where we use the term “fairness”, it carries the US meaning.

engine's results and asked to determine whether a portrayed loan applicant should or would receive a loan or not, keeping in mind that unlawful discrimination needs to be avoided. Results show argumentation patterns that focus on discrimination or on real-world knowledge, the influence of task and human-computer interaction (HCI) design on argumentation, and differences in argumentation between high-performing and low-performing participants.

In the second, conceptual, part of the paper, we widen the view to consider further ways in which humans are involved in shaping the outcomes of systems involving data mining, and consider limitations on formalising fairness or non-discrimination. We then turn to the notion of *accountability* which is currently intensively debated in connection with decisions involving computers. Informally speaking, this means that it should be possible to “hold someone/something to account” for decisions and outcomes¹⁶. Thus, accountability is a property that complements the (utopian?) idea that it could be possible to “fix” machine biases: an accountable data controller should be able to and support the inspection, questioning, and ideally also rectification of machine and other biases. This will often involve transparency, i.e. providing the basis of decisions. We then discuss the principle of accountability set forth in the new European General Data Protection Regulation (GDPR) and explain why we consider this definition particularly useful for the current problem. Drawing on the results of our empirical study and our conceptual analysis, we then ask how mining software can help make decision makers accountable with respect to discriminatory biases.

The twofold contribution and organisation of the paper are as follows: First, we present results from a user study that provide a detailed look at human decision processes in semi-automated decision making, as they relate to discrimination (Section 3; this builds on work summarised in Section 2). Second, we highlight key aspects of the role of humans in the discriminatory outcomes from (semi-)automated decision making in order to derive two *data protection by design* strategies to support accountable data processing, as defined by the GDPR and applicable also beyond it (Sections 4 and 5). Section 6 concludes with design recommendations, limitations, and future work. The relevant literature comes from different areas and is woven into the text to enhance readability.

To contextualise this content, we show, in Figure 1, a simplified overview of key actors, entities, and processes involved in decision making. Consider first the downward-pointing arrows. We focus on the human decision makers who *make* decisions by *performing* a decision process that *yields* some decision result. The experimental design assumes that there is a software tool that has mined data and computed analysis results, and keeping these constant, we investigate different ways in which a tool *presents* these results (Section 2.1) and study the resulting *decision processes* (Section 3). We argue throughout the paper that most decision making involves humans in this way, and we move from the HCI-centric perspective on such involvement to a more general (e.g. organisational and societal) one, see in particular Sections 4 and 5.1. This is in contrast to classical DADM, which focusses on the dashed line describing cases in which software tools alone can make decisions, and/or regards only algorithms and their outcomes.

We also discuss a number of the upward-pointing arrows: that decision results may not only impact data subjects directly, but also *become* data feeding into future mining and decision making (Sections 4.1 and 4.2), and relevant aspects of how the law *regulates* data processing (Sections 4.3, 4.4 and 5.1). We then move back to a software- and HCI-centric perspective and ask how these

legal requirements can *inform the design* of processing, focusing on the software tools (Sections 5.2 and 5.3). We briefly mention some of the other entities and processes (the influence of organisational context); the remaining ones are beyond the scope of the present paper.

The paper complements an earlier publication¹⁵, whose conceptual part introduced the notions of constraint-oriented and exploratory DADM (summarised in Section 2.1) and discussed the relative merits of these two forms of *computing and presenting* analysis results in the light of how laws conceptualise discrimination (not surveyed here), and whose empirical part focused on the *decision results* (summarised in Section 2.2).

FIG. 1. Actors, entities, and processes involved in data-based decision-making.

2 Computer-assisted (non-)discriminatory decisions with data mining? A user study

How do people make decisions with data mining? We investigated this question by focusing on a task discussed and modelled throughout the literature on discrimination and data mining: retail banking. This formed our first experimental setting: a bank clerk tasked with preparing a loan decision, and given the instruction to not discriminate against applicants. We also considered a second setting, which we had explored in a prior user study¹⁷: that of a person working in an anti-discrimination agency looking into loan applications to check for possible occurrences of discrimination. Again, we assumed that this person would work with the results of data mining as a basis for their judgements.

Our goal was to find out whether a sanitisation of the data mining algorithm would be sufficient and necessary for avoiding discrimination, to investigate the influence of setting or task as well as HCI design (sanitising vs. flagging discrimination), and to study the reasoning and decision processes when working with data mining results.

A first analysis of data from this experiment was presented in the companion article¹⁵. To keep the present paper self-contained, in this Section we give an overview of the experiment's design and method, as well as main findings, and refer the reader to that earlier paper for details.

2.1 Design, materials and method

We created experimental conditions that differed along two dimensions:

- Setting (see Fig. 2)
 - *Bank*: participant assumes the role of a bank clerk
 - *ADA*: participant assumes the role of an anti-discrimination agency
- Mining form (see Fig. 3)
 - *cDADM*: the standard, algorithm-focused approach of discrimination-aware and fairness-aware data mining, in which constraints define which discrimination to avoid and rules are sanitised.
 - *eDADM*: an approach involving exploratory data analysis, which invites the user to explore the discriminatory components in rules to reach the desired decision outcomes. These components are highlighted visually.
 - *DM*: non-DADM data mining as a control condition.

This resulted in 2 (settings) * 3 (mining forms), i.e. 6 experimental conditions. The settings were introduced to participants via instructions about how to use data-mining results for reaching decisions and instructions to avoid discrimination in the process (Fig. 2).

a) Bank: You work in a bank, and your responsibility is to prepare loan decisions for your manager: Based on an applicant's data, you propose to either grant or deny a specific loan request. The bank's policy is to draw on data analyses of past loan data.

b) ADA: You work in a citizen-advice / company-watch center, and your responsibility is to prepare decisions for your manager: Based on an applicant's data, you predict likely outcomes. Your manager will use your predictions to derive "alerts" as to which cases to follow up. Various citizens have turned to the center for help: they want a loan from the bank, but are not sure whether they will get it. The center has access to data analyses of past loan data, and it uses these analyses both to advise individuals and to monitor patterns of lending.

c) Bank, ADA: Thus, you will receive

(a) data describing the requested loan and the applicant, and

(b) statistical rules that argue for or against granting the loan, given specific data from (a).

Based on the information from (a) and the decision support from (b), you will

- propose a decision for your manager and
- motivate that decision.

d) Note that banks are not allowed to discriminate against applicants based on gender, marital status, nationality, or age. It is important for {**Bank**: the bank | **ADA**: the center to monitor} that decisions that discriminate based on these features not be taken -- even if data from the past suggest it.

If any of your answers (the {**Bank**: recommendation to grant or deny the loan | **ADA**: conclusion as to the likelihood of the loan being granted or denied}, or your motivations) need explanation with regard to possible discrimination, please note this in the free-form answer box.

FIG. 2. The overall task descriptions for the settings Bank (a, c) and ADA (b, c), and the instructions against discrimination in the two settings (d). ADA, antidiscrimination agency.

215 participants were recruited over Amazon Mechanical Turk.^b They received USD 6.00 for full participation and up to USD 1.50 as an additional performance-dependent payoff (bonus). To reduce cultural confounds, we recruited only US participants. We also heeded factors that have been observed to drastically reduce the occurrence of cheating and increase data quality^{18,19}. We included attention-check questions whose cross-evaluation can help identify users who checked answer options randomly. All participants obtained a check score of at least 50 % of the possible maximum. Further analyses of our results gave no indication of cheaters either.

The participants, randomly and approximately equally distributed over the 6 conditions, were asked to consider a series of loan requests. They were given features of the request and the applicant, and provided with decision-supporting rules of a data-mining tool that was fictitious but based on the principles of the mining form. Bank participants were asked to decide whether to grant the loan or not, and to motivate their decision. ADA participants were asked to conclude whether they considered it likely that the loan would be granted or not, and to motivate their conclusion.

The textual loan request description contained an equal number of applicant and request details for each example, with the applicant details comprising features considered legitimate and others considered discriminatory in the decision context (e.g., "Dabiku is a Kenyan national. [...] She has been employed as a manager for the past 10 years. She now asks for a loan of \$10,000 for 24 months to set up her own business. [...]"). The rules given as tool output referred to these and other features. The tool output consisted of visualisations of decision rules in an intentionally minimalistic way that follows the basic logic of the rule miners that inspired DADM. However, it

^b Participants gave informed consent to the collection and processing of their answers via the study's Web interface.

does not perform the last step of calculating the scores that makes the miner decide between two classes (“yes”/ “no”). This calculation was left as a task for the user. These visualisations were given in three versions to implement the three mining forms. Figure 3 shows an example. The “foreign worker” and “gender” attributes are highlighted in the dark red rules in eDADM , but do not appear in cDADM, where sanitisation has led to different rule contents. The discriminatory content of rules can only be explored when it is there (in eDADM and not in cDADM). In principle, DM also allows for exploration, but due to its undifferentiated visual display, it does not actively encourage it.

FIG. 3. The tool interfaces for (top left) cDADM, (top right) eDADM, and (bottom) DM. The visualization is identical between cDADM and DM, and the risk factors are identical between DM and eDADM. eDADM highlights rules with discriminatory features in red (second and fourth bar in the example). Identical visualizations were used for the Bank and ADA settings. DADM, discrimination-aware data mining.

Participants solved three practice tasks with feedback and six assessed tasks. An exit questionnaire completed the study. First, we asked for impressions about the task and the tool. Subsequently, participants were asked for some basic demographics. Three quarters of participants stated that they had “applied for a loan at least once in [their] life” (73 %, validated by a reverse-coded question), with 50 % of these having at least once been denied a loan. Also, 50 % of all participants reported that they had “experienced discrimination in [their] own life”, and 59 % reported having experienced it around them. These tended to occur together ($r=.60$ overall, with 83% of those reporting “own” experience also reporting “other” experience, and 72% vice versa).

We applied some simplifications when operationalizing the constructs in order to test the formalisations of discrimination employed in today’s DADM, maximise experimental control, and

make the tasks feasible for participants. First, we applied a simplified definition of “discrimination”: we restricted the specified attributes to four (gender, marital status, nationality, and age), and declared any use of these attributes towards the decision as illegitimate discrimination, without exceptions. All other attributes were legitimate (for examples, see Fig. 3). Second, we wanted to avoid obtaining results confounded by the choice of any specific data mining algorithm. We therefore decided to implement only the key difference between cDADM and eDADM: whether to hide/remove or to highlight discrimination-indexed features in rules.

2.2. Results and discussion: Overview of decision outcomes

We addressed the accuracy and actionability of the conclusions and the reasoning process. Decision quality was measured by the number of correct decisions and motivations, i.e. those that corresponded to the outcome of taking into account all legitimate features present both in the scenario and in the rules, their numerical weights in the rules, and nothing else. The results suggest that both constraint-oriented and exploratory DADM support correct conclusions as well as reasoning and do so better than data mining without DADM support. They also showed that the “Bank” task focused on preventing discrimination was supported better by the sanitisation approach embedded in the cDADM interface, whereas the “ADA” task focused on detecting discrimination was supported better by the eDADM interface that focused on exploration and highlighting possible discrimination. We also found that features of the loan applicant that the data-mining engine filtered out, or even non-present features, were in some cases “re-inserted” by the decision maker in motivating their choice. In the example of Figure 3, re-insertion occurs for instance when, given the sanitised rules, a participant refers to the applicant being a foreign worker. Thus, discriminatory reasoning may persist even when algorithms and results are sanitised. Details of the results including the sizes and statistics of the effects are contained in the companion article¹⁵.

3. Results and discussion: Decision process

The analysis summarised in the previous section focused on the *task* and the *decision outcomes*. As an indication of how people reasoned towards their decision, we analysed the multiple-choice motivations regarding which actual and potential features of applicant and request were taken into account. In the experiment, participants were also encouraged to describe their reasoning in textual free form. Here, we focus on decision *process* as described in these comments, analyse their contents, and investigate possible determinants (setting, mining form, selected demographics) as well as the relation with decision outcomes. This will help to better understand a feature of decisions that is key for accountability: how decision makers explain and justify their conduct. In Section 5, we will situate the explanation and justification of conduct in the context of a comprehensive definition of accountability, and discuss implications of our results.

3.1 Argumentation features

111 participants gave comments explaining their decisions and motivations. For each participant who had commented in any way, the comments on all the assessed tasks were concatenated. On average, these concatenated comments were 66 words (375.2 characters) long, ranging from 2 words (“too risky” for task 1) to a detailed account for each task of a total of 220 words.

After a first reading of all comments and a discussion about relevant features, we settled on the following features, to be coded in binary fashion (present or absent). Each involved the mentioning of specific information. Except in obvious cases (NAME, OTHER_FEATURES, WEIGHTS, META), we also list sample words that were used for coding instructions.

1. Who?

- I: The participant in a first-person narrative (e.g., “I”, “me”)
- YOU: The reader or a fictitious dialogue partner who is given advice (e.g., “you”)
- NAME: The name of the person described (the loan applicant)

2. What?

- DISCRIMINATION: discrimination as such (e.g., “discrimination”, “discriminatory”)
- DISC_FEATURES: discriminatory features of the applicant; specifically: age, gender, nationality or family status (including obvious synonyms or labels)
- OTHER_FEATURES: other features of the applicant or the application

3. Why?

- WEIGHTS: Numerical weights as given by the data-mining results, with an explicit comparison between positive and negative weights (e.g., “outweigh”, “more than”, “stronger”, “higher”)
- WORLD_KNOWLEDGE: Real-world knowledge beyond the data mining tool’s rules was referred to (e.g., “unreliable job”).
- CAUSAL: Reasoning, establishing a causal relationship (e.g., “because”, “since”, “so”)

4. META: A meta-comment on the task was given.

Two coders independently classified each comment with these features. A comparison showed that there was disagreement on 9% of the labels (103 of 1110 = 111 participants * 10 variables). Cohen’s kappa per feature indicated a reasonable (0.77) inter-rater agreement for WORLD_KNOWLEDGE and a good to very good one for the remaining features, ranging from 0.83 (CAUSAL) to 0.99 (NAME). In a discussion round, all cases of disagreement were resolved, and the resulting consensus matrix was used in further analyses.

The argumentation styles were well distributed over the comments, and most comments included several argumentation styles, as shown in Table 1. Ten comments were not labelled with any code; these consisted only of ‘no’s (=“I have no comments”) and/or of non-pertinent remarks.

	0	1	2	3	4	5	6	7
Comments with that many distinct features	10	17	30	22	17	12	1	2

Table 1. Distributions of argumentation features over comments

The most frequent types of reasoning involved OTHER_FEATURES of the loan application or applicant: 46.8% of the comments. This was to be expected given the feature-centric descriptions, the data-mining results and the questions, as well as the instruction to not make decisions based on discriminatory features. Many participants made it explicit that and how *they* had made a decision (I: 43.2%); only 3.6% gave advice to another decision maker (YOU). This may have been encouraged by the question that addressed the participant personally. 31.5% each mentioned DISCRIMINATION or DISC_FEATURES. The loan applicant him/herself receded behind his/her features: a NAME was only mentioned in 17.1% of comments.

31.5% of the comments explained a decision made in terms of WORLD_KNOWLEDGE, which was sometimes at odds with the results of the fictitious data-mining engine. 27% explained a decision in terms of these results, referring to the WEIGHTS and comparing them explicitly. 9.9% employed CAUSAL reasoning that indicated they diverged from purely following the (correlational) data-mining results, explaining their reasoning in detail. 19.8% gave META comments. Figure 4 shows examples.

- a)** The fact that the rule that **Dabiku is foreign**, though giving a favorable rating, should be irrelevant. The fact that the **purpose of the business** is combined with her **foreign status** should make that irrelevant as well. The loan should be granted. **Seeing that** the rule that he is a **foreign born worker** is giving him a unfavorable score, it should be irrelevant due to **discrimination**. Therefor, the score shows he should be granted the loan.
- Though the **scores are close**, the rules show favor in granting the loan. Loan should be granted.
- The score that is given due to her being from **Canada** should not be considered due to **discrimination**. Removing that score shows that she should indeed be granted the loan. Loan should be granted.
- Taking into account the scores for the rules given, Henk should be granted the loan.
- Taking out the score** for where she is from due to **discrimination**, which is bundled with the rule of her **not having any savings**, she still should not be granted the loan.
- b)** The averages suggest she has a higher chance of getting the loan than not getting the loan.
- Even though the column is [red] having a loan out always seemed favorable in other tasks, this I feel **pushed him over the edge** into getting a loan.
- The averages suggest they would get the loan, I wish I knew **how much the loan was going to be** though.
- His **debts with no current payment plan** are what set him back from getting a loan.
- The **averages seem close**, however since she is asking for the **loan for a new car** I think that would set her up for getting the loan, **as a car has value that the bank can take back** should she not make her payments.

FIG. 4. Sample comments from two participants in mining form eDADM. Settings and argumentation feature coding, in bold font: (a) Bank; NAME, DISC_FEATURES, OTHER_FEATURES, DISCRIMINATION, WEIGHTS. (b) ADA; I, OTHER_FEATURES, WEIGHTS, WORLD_KNOWLEDGE, CAUSAL. (Each paragraph refers to one task).

To identify possible patterns, we inspected the results manually and calculated pairwise Fisher’s Exact Test of independence^c. Due to the large number of comparisons and the ex-post nature of the comparisons, the significance results must be interpreted with care, but they still serve to identify relevant findings. Throughout Sections 3.2 – 3.4, we use an error level of $\alpha=.05$ and distinguish between

(**) significant after a Bonferroni correction for all tests in Tables 2 and 3, and

(*) significant without the Bonferroni correction.

The results (total numbers) are shown in Table 2, and co-occurrences are described in the text. All pairwise relationships not described in the text were non-significant.

	forming patterns									
	NAME	I	YOU	DISC_FEATURES	OTHER_FEATURES	DISCRIMINATION	CAUSAL	WEIGHTS	WORLD_KNOWLEDGE	M
total	19	48	4	35	52	35	11	30	35	22
NAME	10	1		8	13	8	5	6	7	3
I		2		20	26	19	7	15	19	8
YOU			2		3	2	0	2	2	2
DISC_FEATURES				14		21	3	11	4	4
OTHER_FEATURES					8		9	15	35	10
DISCRIMINATION						3		11	1	3
CAUSAL							5		7	2
WEIGHTS								7		5
WORLD_KNOWLEDGE									7	
M										7

Table 2. Argumentation features forming patterns

3.2 Argumentation patterns

To identify possible argumentation patterns, we considered co-occurrences of argumentation features. The results are shown in Table 2.

DISC_FEATURES and DISCRIMINATION were often mentioned together (**), as were OTHER_FEATURES and WORLD_KNOWLEDGE (**). *All* comments involving WORLD_KNOWLEDGE also involved OTHER_FEATURES, but not vice versa. These appeared to be characterizing two modes of reasoning, with DISCRIMINATION and WORLD_KNOWLEDGE being dis-associated with each other (**). The same relationship re-surfaced in dis-associative relationships between OTHER_FEATURES – DISCRIMINATION and DISC_FEATURES – WORLD_KNOWLEDGE (both *). CAUSAL reasoning tended to be associated with OTHER_FEATURES (*) and with WORLD_KNOWLEDGE, although the latter was non-significant.

^c using <http://www.langsrud.com/fisher.htm>

Names of loan applicants were accompanied less often by this person’s discriminatory features (in 8/19 cases). NAMES of loan applicants and CAUSAL reasoning tended to be dis-associated (*).

We group these findings into the two pattern groups. 44.5% of all comments were *discrimination-centric* (DISC_FEATURES, DISCRIMINATION or both), and 47.3% were *financial-logic-centric* (OTHER_FEATURES, WORLD_KNOWLEDGE or both). 13.6% of all comments included both styles of reasoning (DISC_FEATURES or DISCRIMINATION, coupled with OTHER_FEATURES and WORLD_KNOWLEDGE).

3.3 The influence of mining form and setting on argumentation

We investigated the possible influence of the experimental treatments on argumentation, with results shown in Table 3 (left).

	setting		mining form				demographic features		
	bank	ADA	c DADM	e DADM	DM		own	other	loan denied
<i>total</i>	61	50	38	41	32	<i>total</i>	48	63	44
NAME	10	9	5	9	5	NAME	13	11	9
I	25	23	11	23	14	I	22	31	17
YOU	2	2	2	2	0	YOU	2	2	0
DISC_FEATURES	16	19	5	16	14	DISC_FEATURES	17	21	14
OTHER_FEATURES	29	23	22	16	14	OTHER_FEATURES	22	32	16
DISCRIMINATION	13	22	2	20	13	DISCRIMINATION	13	18	13
CAUSAL	8	3	3	5	3	CAUSAL	5	5	4
WEIGHTS	14	16	14	11	5	WEIGHTS	14	16	12
WORLD_KNOWLEDGE	23	12	13	11	11	WORLD_KNOWLEDGE	14	21	10
M	15	7	11	6	5	M	12	13	6

Table 3. Argumentation features in relation to (left) setting and mining form; (right) demographics

The mining form influenced the style of reasoning: in eDADM, about half of the commenting participants in the eDADM treatments referred to discrimination, more than in DM, which in turn was more than in cDADM (mining form – DISCRIMINATION: **). Discriminatory features were also mentioned more often in eDADM than in cDADM; but interestingly, they were mentioned most often in DM (mining form – DISC_FEATURES: *). This may be interpreted as showing that eDADM made participants more aware of the need to avoid or detect discrimination, but DM made it more necessary to pay attention to the detailed features (because these were not flagged by the interface). This attention to features did however not lead to better results; instead, the decision quality was best in eDADM¹⁵. The observation that discrimination and discriminatory features were

mentioned less often in cDADM is probably a result of the fact that no discriminating features were shown in the fictitious data-mining results in the cDADM treatment.

Loan applicants' names were also mentioned more often in eDADM than in DM and in turn in cDADM, which may indicate that eDADM makes it easier to relate to the applicant as a person rather than a case. However, this relationship was not statistically significant, and it may also be an effect of other factors.

The setting also had an influence: ADA made it more likely that discrimination was mentioned (setting – DISCRIMINATION: *), which could be interpreted to mean that the clerk in the ADA setting had one goal (to detect cases of discrimination), whereas the clerk in the Bank setting also had the business goals of their bank to take into account.

3.4 Demographic features and argumentation

We also investigated the relationship between selected demographic variables and argumentation. We focused on “own experience of discrimination”, “discrimination of others experienced” and “having had a loan denied in the past”, each of which were quite evenly distributed in the respondent population. The results are shown in Table 3 (right).

Participants with “own experience” tended to mention loan applicants' NAMES more often (*).

Some further relationships did not reach statistical significance, but we mention them as interesting descriptive statistics of our observations. Participants with “other experience” tended to express more “first-person messages” than seemingly objective messages (other experience – I: $p=.2$) and refer more seldom to the applicants. Having had a loan denied in the past may be associated with less attention to OTHER_FEATURES and WORLD_KNOWLEDGE reasoning (however, each only had $p=.1$) – styles of argumentation that these participants may have experienced in their own unsuccessful past attempts at getting a loan.

3.5 Argumentation and decision outcomes

Finally, we investigated whether patterns could be found linking certain argumentation features to outcomes. First, the participants who did make comments scored better than those who did not (average numbers of correctly solved tasks were 4.4 versus 3.9, $p<.05$ on a two-tailed t-test). This indicates that reflection per se may have a beneficial effect on making less discriminatory decisions, although we cannot be sure of the causal structure.

With regard to argumentation styles it is difficult to establish a cause-effect relationship, and in addition the high number of variables and the cardinal scale level of outcomes make statistical tests difficult to apply. We therefore limit ourselves to a descriptive search for possible patterns of argumentation by more or less successful decision makers. Figure 5 shows the result.

FIG. 5. Proportion of participants who argued with the help of each of the features.

The results indicate that successful solvers (5 or 6 correct) tended to focus on the numerical WEIGHTS and their high-level task to avoid or detect discrimination, and they paid attention to DISC_FEATURES. Less successful solvers (3 or 4 correct, and in particular 1 or 2), in contrast, were led by WORLD_KNOWLEDGE reasoning with OTHER_FEATURES and tended more than the others to consider the applicant as a person (NAME); they tended to focus less on discrimination as a topic or as features (DISCRIMINATION, DISC_FEATURES).

In sum, task as well as HCI choices had an impact on reasoning, and outcomes were linked to different foci in the comments. The statistically weak relationships between demographic background and comment contents should be investigated in future work.

The tasks and HCI choices we studied were designed to model prototypical conditions in which decision makers and their organisations have to (be able to) explain and justify their conduct as part of being accountable for it. Our results indicate that the tasks and HCI choices have an impact, via their influence on human computer-assisted decision making, both on decision outcomes (Section 2) and on explanations and justifications (Section 3) and should therefore be investigated carefully in technological and process decisions that aim at reducing or avoiding discrimination. The question arises whether there are further dimensions along which human influences and influenceability can contribute to less (or more) discrimination in connection with data mining. After a conceptual

investigation of this question in the following Section 4, we will derive, from an integrated view of these results, proposals for designing for accountability in Section 5.

4. The role of humans in generating discriminatory outcomes of data mining, and the limits of formalisation

In the previous sections, we have investigated the direct involvement of human decision makers in data-based decisions and investigated factors related to task and HCI factors. However, humans are involved in data mining at other levels too. These will be investigated in this section.

4.1 Data and data creation

The role of data and the potentially discriminatory outcomes (“disparate impact”) produced by the interaction of ‘harmless’ or ‘neutral’ algorithms and data fraught with biases have been shown by analysis⁵ and many examples, as illustrated in the Introduction. More often than not, it is humans generating these data, as when search-engine users create an association between black-sounding names and searches for criminal records (this is one possible reason³) or searching often for “women shouldn’t vote” (which then became an auto-completion proposal⁷), or Twitter users ‘training’ a bot to make racist jokes²⁰. In predictive policing applications, where algorithms are often praised for being more ‘objective’ than people, data on crimes are generated where police have searched for these crimes in the past, which can bias recommendations for future searches²¹.

The “sanitisation” of data is an approach alongside the “sanitisation” of algorithms (see the classification of approaches²²). Such data/result modifications are possible when one knows what to avoid (such as using gender or ethnicity as features, and outcomes such as loan decisions or insurance premiums as the target variable), or when indirect discrimination can be detected by correlational analysis of attributes (such as location of residence as a proxy for ethnicity or poverty, e.g.^{23,24,22,25}). Even when the problem is clear, the detection of discrimination is often not straightforward and requires a whole knowledge discovery (KDD) process²⁶.

However, it is impossible for an algorithm designer, or even a search-engine provider with access not only to algorithms, but also to rich data, to expect (and remedy a priori) all possible relations of, e.g., all words referring to a protected group on the one hand and all phrases denying this group their rights, ascribing them negative properties, etc. Also, even the automated detection of learned discriminatory associations would require a human-level artificial intelligence. Barring “associations” as such, on the other hand, would preclude adaptivity and personalisation of services. This trade-off remains an open problem.

4.2 Dynamics of data, decisions, and discrimination

Vicious circles have been recognised as major contributors to discrimination for a long time, from Myrdal's classic study on US racial inequalities²⁷ to current investigations of discrimination against Muslims in Europe²⁸. Myrdal described the "cumulative causation" in which prejudices of white populations and low living standards of black populations reinforce each other in a downward spiral: prejudice leads to institutional discriminatory processes, which lead to deteriorating standards of living, which feeds prejudice, and so on. Thus, decisions often arise from hard-to-break "entire networks of rules and practices [that disadvantage] less empowered groups" known as structural or institutional discrimination⁵. At any point in the chain, there may or may not be intent to discriminate – good intentions are immaterial for the outcome when viewed through the lens of disparate impact. Today, "data" about individuals assume part of the role of "living conditions".

One part of these dynamics is that discrimination solidifies into data as described in Section 4.1, but the dynamic nature of the process makes it much harder to detect this. For instance, "applicant has no collateral" may have been identified as a factor against granting a loan by an algorithm. Even if one detects that it is correlated with membership in a protected group, it appears to be a perfectly legitimate factor in the banking domain and thus should not be blocked by algorithmic safeguards against indirect discrimination; it becomes "explainable discrimination"^{29,30,31}. At the micro-level of the individual loan decision, the resulting differential treatment appears "rational"³². However, at the macro-level one may find discriminatory dynamics.

Consider two examples. First, there is evidence that people with bad credit scores (presumably because of reasons such as missing collateral), i.e. people who are likely to not get loans from the traditional banking sector, turn to payday loans. These not only impoverish applicants further by higher interest rates, the data feature "has had a payday loan" also negatively impacts the individual's future credit score³³. Similar dynamics may play out, second, in the domain of research funding, in which evidence of discrimination against female researchers was found²⁶. As a result of such funding decisions, female researchers will have less research funding – a data item which often negatively impacts future funding decisions.

Complicating matters further, dynamics can create new discrimination grounds, intersectionality and pluridimensional discrimination, phenomena that generally defy the algorithmic detection and prevention of discrimination¹⁵.

Cumulative causation can also work towards improvement, via virtuous circles, and this thesis has influenced major policy decisions such as the desegregation of schools in the US. Vicious and virtuous circles are also consistent with European explanations of and concepts against discrimination^{28, d}. A necessary prerequisite for such attempts at remediation is to notice and understand the dynamics of cumulative causation. This first should draw on the huge body of knowledge from the social sciences³². However, the study of manifestations of discrimination in data has only just begun (if only because "Big Data" itself is a relatively recent phenomenon). Computer science can contribute tools for exploratory and interactive data analysis, and the understanding that phenomena may need repeated and iterative analysis.

d Myrdal's optimism as well as his belief in the strength of the moral basis needed to overcome racial inequality in the US, on the other hand, have been more controversial^{34,35}.

4.3 Decision making is rarely fully automated

Even if we differentiate between the roles of data and algorithms and account for their interactions: which of the decisions referred to in the introduction, and similar ones, are actually made in a fully automated way? The serving of search-engine results may be one example, although even here decisions to censor certain content must be “manually” programmed, for example as exceptions. However, most real-life decision making is not purely algorithmic, but computer-*assisted*.

The claim of less-than-full automation is, first, *descriptive*. It follows from the observation that discrimination can not only result from the application of “the data-mining algorithm” to “the data” (e.g., through the reproduction of biased associations, see Section 4.1), but also from the definition and selection of features, target variables, and labels, as well as the collection of data⁵. In other words, discrimination can result from decisions along the whole process of knowledge discovery. In domains such as the justice sector, automated risk assessments regularly are – and must remain – input to human decisions. But what about domains such as banking and insurance?

EU data protection law requires forms of involving humans in decisions in a *normative* way in Article 22 GDPR: “The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.” In the – permitted – exceptions to this general rule that are relevant to our discussion (decisions necessary for contract or based on data subject consent), “the data controller shall implement suitable measures ... [giving the data subject] the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision”. A similar combination of prohibition, exceptions, and requirements of safeguards, is already contained in current law, Article 15 of the EU Directive 95/46/EC^e.

The reasoning towards this provision reflected, long before “Big Data”, the same concerns about unbridled fully automated decision making that we summarised in Sections 1 and 4.1. Further arguments concerned legal or ethical responsibility and the roles of data subject and data controller: the former should remain a participant in a decision concerning him or her, the latter should not abdicate responsibility³⁶). Whether and what protection against out-of-control algorithmic power this actually offers has been controversial³⁶.

Investigating uses of Article 15 (using the example of Germany, since a full exploration was beyond the scope of this paper), we saw that only two court judgments^f have actually referred to § 6a *Bundesdatenschutzgesetz* (BDSG), the implementation of Article 15 in the Directive – and that they found it inapplicable. The complaints concerned the expected main use case of § 6a: loan decisions. In both cases, the question was whether scores produced by the major German credit scoring agency constitute an example of “automated individual decision making” and are therefore contestable under § 6a. The courts ruled in both cases^{37,38}: The loan decision procedure is not fully automated, because it is not a unitary decision but one composed of parts. The score is only an

^e There are some differences between the two Articles, irrelevant for the current argument and thus omitted for brevity.

^f As determined by a search on the pertinent legal portal, juris.de

“expression of opinion” (not a “fact”^g). In a KDD sense, this means that a bank deploys human discretion to use the scores in some way – which confirms the *descriptive* claim made above.

Another important type of data processing does not fall under § 6a BDSG: the compilation of candidate lists for job interviews³⁹ – the significant individual decision is the decision to hire someone or not, the shortlist is but a preparation for this decision. It should be obvious after Section 4.2 that such interpretations become dubious when a large proportion of job applications are never seen by a human (an estimate about the US is 72%⁴⁰) and people get continuously “redlighted”, i.e. de-selected, based on (maybe the same) undisclosed criteria. However, for the present purposes they serve to highlight the multitude of ways in which humans are and remain involved in decisions involving data.

Article 22 and its predecessors do not prohibit the use of automated individual decisions. Instead, Article 22 entitles individuals *to hold data processors to account* for such decisions and demand a human intervention: data subjects have the right to ask for and obtain an explanation of the decision, to enter into a conversation in the sense of putting their own point of view forward, and to contest the decision. Under the GDPR, data subjects are also entitled to ask for and obtain an explanation of the logic behind processing (Article 15 1 (h)).

4.4 Limitations of formal measures of discrimination

The Aristotelean principle “treat equal things equally and unequal things unequally” stands behind non-discrimination laws. If one had a formal measure of equality, a test would be able to determine automatically whether this principle was followed. However, it is extremely difficult in practice to find such measures. In addition, satisfying one formal notion of fairness may mean violating another⁴¹.

Intended to regulate human societies, laws are not formal rules, but require interpretation in the respective contexts. This can present a challenge for data-based modelling. First, equal feature values (of individuals and of treatments) suggest equality. However, a *failure* to apply differential treatment may constitute discrimination. For example, requiring all employees to enter the workplace building by a staircase can discriminate against people in wheelchairs. Conversely, differential treatment may constitute no discrimination when it is justified by a legitimate aim and the means of achieving that aim are appropriate and necessary (“proportional”).

Data as representations of what individuals *are* may fail for a second reason. Consider a recent court case that ruled that it had been discriminatory to treat an unmarried homosexual couple equally to unmarried heterosexual couples (the non-EU partner was denied a residence permit on family grounds)⁴². The court argued that a heterosexual couple *can* get married (and thus change its feature value relevant to this domain), whereas at the time and place under consideration a homosexual couple was not able to do so. The judgement not only shows the importance also of

^g For this reason, the agency is also allowed to not disclose its formula (this was ruled in a different judgement), and the differential treatment that was the original motivation of the first court case does not constitute gender discrimination.

counterfactual reasoning (which may be amenable to AI reasoning^h). It also shows how non-discrimination interacts with other fundamental rights such as autonomy and agency.

It is difficult and maybe impossible to preview all such considerations and formalise them. Again, human intelligence can be brought to bear better through exploratory analysis tools. Such tools constitute an important component of the approach that we will sketch in the following section.

5. Towards accountability with respect to discrimination-aware decisions

Given the risks of data-based decision making and the importance of humans in the related processes, how can data subjects' rights be strengthened? In the "Bank" setting of the user study, the bank had the task of taking measures to ensure that their mining-based decisions do not violate anti-discrimination requirements. The "ADA" setting presumed another task of the bank: that of supplying the authority with documentation of its data processing (here: their data-mining rules, as in Fig. 3). This enabled the authority to perform its task of monitoring the bank's compliance; the ADA setting modelled this task. In addition to, and in preparation of, such external monitoring, internal monitoring is useful and could be mandated legally.

5.1 Accountability as a social relationship and as a data protection principle

These two tasks instantiate the *principle of accountability* as defined by Articles 5 (2) and 24 of the GDPR on the responsibility of the data controller: "the controller shall implement appropriate technical and organisational measures to ensure and to be able to demonstrate that processing is performed in accordance with this Regulation". Accountability is also a requirement in other common data protection guidelines, in particular the OECD Guidelines⁴³ or the ISO/IEC 29100 standard⁴⁴, and in some national laws (e.g., Canada⁴⁵). However, the term's meaning remains contested and often vague^{46,45}.

Bovens⁴⁶ provided a clear operationalisation (and an in-depth discussion): "A relationship qualifies as a case of accountability when there is a relationship between an actor and a forum, in which the actor is obliged to explain and justify his conduct, the forum can pose questions, pass judgement, and the actor may face consequences."ⁱ For the present purposes in which *accountability* stands *in relation to a norm*, we propose to add that the conduct is *relative to the norm* and that the actor also, and in general before the other activities, has to *report on* the conduct.

EU data protection law has so far contained, implicitly, elements of accountability, such as the data controller (the actor) being responsible for informing data subjects about processing and

^h "Counterfactual fairness" is a recent research direction exploring similar issues, see https://www.turing.ac.uk/research_projects/counterfactual-fairness/.

ⁱ We believe that the consequences, which may be sanctions as well as rewards, constitute an essential component of accountability: the incentive to comply. It is for this reason that we believe true accountability can only be a property of people or organisations that in turn internally make people accountable, and not of algorithms.

correcting inaccuracies in data if requested. The possibility for individuals to hold the controller to account (Section 4.3) is another example in which the forum is the data subject. The motivation for including an explicit principle of accountability into the GDPR was the perception that such implicit requirements to be accountable, as well as other data protection requirements, were not put into practice sufficiently: while data processors were under meaningful obligations and may have, internally, put in place meaningful policies designed to meet these obligations, this did not ensure that they actually acted accordingly⁴⁷. The principle of accountability of the GDPR aims to overcome this problem. Here, data protection authorities are the forum, and they are generally likely to be in a stronger position to monitor and enforce compliance than data subjects. Through Article 5 the GDPR emphasises (a) responsibility (and therefore people) as an essential element of accountability, and (b) that transparency and accountability require each other, but are not identical.

5.2 Design strategies for accountability, and their limitation in the age of profiling

Requirements to comply with rules, and, maybe even more strongly, requirements to ensure and demonstrate that one complies with rules, are more effective when these considerations are built, from the start, into policies, procedures, practices and last not least products. To further strengthen data protection, the GDPR therefore requires data controllers to apply *data protection by design* (Article 25). Methodologies for designing software and systems in this way have been and are being developed actively (see for example a recent proposal and overview⁴⁸. Hoepman⁴⁹ has provided a useful analytical complement to such procedural methodologies: to use the *principles* of the GDPR as high-level design goals and the wealth of *existing PETs* (*privacy-enhancing technologies*) as possible implementations. This gulf is bridged by *design strategies* and *design patterns* as increasingly concrete steps.

Specifically for the principle of accountability, Hoepman formulates two *design strategies*:

1. ENFORCE: A privacy policy compatible with legal requirements should be in place and should be enforced.
2. DEMONSTRATE: The controller is required to demonstrate compliance with the privacy policy and legal requirements.

He then lists typical *design patterns* for the two strategies, such as access control and sticky policies for ENFORCE and logging and auditing for DEMONSTRATE, and gives sample PETs.

However, these strategies, patterns and PETs focus strongly on those threats that classical data protection is concerned with: the wrong people (for the wrong purposes, at the wrong time, ...) get access to data. This focusses on the data as they are and the knowledge about single individuals as encapsulated in a data point about this individual. With data increasingly being used to make *inferences*, the threat changes to also encompass inferred knowledge (e.g. “profiles”), which may (a) derive from data about other individuals, (b) be based on data that was anonymised in sophisticated ways and otherwise handled in compliant ways, and (c) that has harmful effects on groups, whereas Article 22 and others only protect individuals. The requirement to protect against harms including discrimination arising from profiling is therefore not covered adequately.

5.3 Design strategies for accountability with respect to discrimination-aware decisions

As Sections 2-4 have shown, non-discrimination can neither be enforced nor demonstrated by software means alone, and it is difficult, if not impossible, to prove the absence of discrimination. We therefore propose two new design strategies with a wider scope of corresponding design patterns:

PREVENT-D specifies that technical and organisational measures that avoid discrimination should be in place and be enforced. This design strategy is implemented by design patterns consisting of

- Formal and algorithmic means to ensure discrimination-/fairness-awareness (data mining used in a predictive fashion),
- HCI choices to support discrimination-aware decision making,
- Recursively, implementations of the design strategy DEMONSTRATE-ND, in order to ensure that (1) the data operated upon and (2) the data generated as a result encode as little discrimination as possible.

DEMONSTRATE-ND requires the controller to demonstrate that given the available knowledge, no discrimination (as specified by the agreed-upon measures and thresholds) can be detected. This design strategy is implemented by design patterns consisting of

- Formal measures of non-discrimination/fairness, applied in data mining used in a descriptive fashion,
- Tools and KDD processes for discrimination discovery and exploratory discrimination discovery,
- HCI choices to support the detection of possible discrimination.

6 Conclusions and future work

We conclude with recommendations for design, important limitations of our method, and ideas for future work. The discussion extends elements of Fig. 1: the ways in which laws – and also other norms and values – can *inform design*, how mining results should be *presented*, and which *decision makers'* behaviour we should investigate.

Recommendations for design. Data controllers must perform different tasks to avoid discrimination and comply with data protection law. Our study has shown that different tasks are better supported by different interface choices, that the tasks as well as the interfaces draw attention to different features of the problem, and that verbalizing argumentation per se, as well as different argumentation patterns, were associated with different outcomes. This suggests that sanitisation vs. flagging and highlighting should be employed in interfaces depending on the task, and nudges or requirements to verbalise used encourage reflection.

Interactive decision-support tools can support such reflection and help avoid that unconscious bias unwittingly re-enters decision making. We have made a first proposal for a visualisation tool that supports the multi-perspective exploration of a decision dataset for possible discrimination, and

demonstrated in a user study that this form of presentation and interaction can help human decision making¹⁷. However, design was informed only by general HCI considerations, and it was constant across tasks.

One design pattern, which implements the strategy DEMONSTRATE-ND and thereby also helps PREVENT-D, is an admissibility test of new features. When a human decision maker considers taking into account a feature that the data-mining engine did not use (e.g. the “unreliable job” referred to in Section 3.1), tools could check whether there is organisation-internal or -external knowledge that signals that this must not be used. This knowledge could for example be guidelines or statistics that show that using this new feature would amount to indirect discrimination because it is correlated with an explicitly protected ground, and use the available detection algorithms for such indirect discrimination. The knowledge could also be a reminder of the fact that a particular piece of information, such as health or pregnancy status, must not be asked. To enhance the detection capabilities, third-party datasets could be integrated into the analysis. This requires well-designed interactions between human intelligence, machine intelligence, and organisational measures that encourage the reflection needed in the first step (realising and making explicit the consideration of a certain feature), the analytics in the second step (and the enforcement of their use), and the action upon the results in the third step (discouraging/preventing the use of features found to be inadmissible, accountable documentation and management of the process). The flipside is that such searches could also help find better proxies to mask discrimination. Empirical evaluations will be needed to determine whether such processes are effective and efficient.

Extension by further notions of accountability. The strategies should be combined with others for data protection by design: Interaction with tools could be logged, thereby enhancing the accountability of the whole process (strong accountability⁵⁰). Logging could automatically generate raw forms of report of the consultation session that help demonstrate that possible problems were checked and policies and procedures complied with (the account to be given¹⁶). However, logging itself creates new personal data and thereby poses new challenges. Note that all these software approaches are *technical measures* that cannot alone ensure accountability (including responsibility). They need to be complemented by *organisational measures* to make the law inform the design of processing in Fig. 1.

Future work should also investigate the relationships with further notions of accountability. It is impossible to survey this wide field here, so we will focus on links to resource collections. These include informationaccountability.org and algorithmic accountability (e.g. the summary⁵¹ of group discussions in the FATML (fatml.org) and “Data, responsibly” workshops’ context). These notions share many characteristics with the notion used here, such as auditability. Many of them do not make a strong distinction between accountability and transparency (such as the discussions around algorithmic transparency, epic.org/algorithmic-transparency/). In recent work, the question what and how much needs to be transparent, explained, and understandable is investigated in more detail^{52,53}.

Limitations of the user study. Besides the limitations of the user study design and operationalization¹⁵, parts of our interpretation hinges on our assumption of what the “correct” decisions were. We designed the tasks such that considering the numerical weights and the features in the rules carefully, automatically led to the “correct decisions”. Future work should investigate

the value of the decisions made by participants who were loath to use the sometimes counterintuitive rules and relied on background knowledge instead – thereby also opening new questions about how to deal with data-mining results.

Our interfaces supported decision makers by information on the features of their task. This is only one facet of individual decision making, and it ignores other relevant factors such as social influence or prior decisions. In future work, we aim to investigate a wider range of choice strategies, as well as associated HCI support⁵⁴.

The present paper focused on results from exploratory data analysis rather than experimental design. Future work should refine these by experimental controls or nudges. For example, it could be explored how instructions to verbalise reasoning, or measures that draw attention to the applicant-as-a-person and/or strengthening self-reflection of the decision-maker-as-a-person, influence the detection and prevention of discrimination.

Mechanical Turk. Our participants were recruited via MTurk, and we used established methods for ensuring study validity and data quality (see Section 2.1). Still, the question arises whether these participants' choices can be regarded as representative of the decision makers whose role they took: people working in banks or anti-discrimination agencies. To the best of our knowledge, no studies of this question exist. We therefore ask (a) to what extent MTurkers resemble study participants recruited in more traditional ways and (b) to what extent the latter resemble the modelled decision makers.

Summarising pertinent observations from a huge body of literature, we find: (a.1) The demographics of MTurkers are similar, but not equal to the general population with regard to age, gender, ethnic origin, political orientation, and education (for specifics and further references, especially US samples, see ^{19,55}; our sample showed a similar pattern¹⁵). On the other hand, they are more diverse than standard samples and exhibit very similar value-based motivations, including “racial resentment”⁵⁶. MTurkers are mainly motivated financially, but also by the contents of studies¹⁹; in this sense, they are more similar than standard participants to the working decision makers whose role they took here. Also, 36% of our sample reported having had professional experience working with data mining, finance, and/or banking. Of course, this does not make them *equal* to the employees with specific skills whose role they took. (a.2) MTurkers are comparable to standard participant populations with regard to decision-making biases^{57,19}. Regarding (b), a study of risk assessment in commercial bank lending⁵⁸ indicated that real bankers are subject to the same cognitive biases as general study populations, but that organisational influences, specifically profit motives, were stronger than cognitive factors when both are present. MTurkers, in turn, because of the remuneration and reputation system of the crowdworking platform, tend to pay more attention to the instructions they are given than standard participants⁵⁹. We speculate that to the extent that organisational factors are captured in business rules, and these are made explicit in instructions, this may lead to MTurkers behaving more similarly to professionals than standard participants. In sum, while laboratory and MTurk studies have unavoidable limitations in terms of external validity, these results indicate that recruitment through MTurk is adequate for studies such as ours.

Empirical studies in real-life settings. Still, future work should also study decision making by real-life professionals. It should be kept in mind, however, that this presents not only the challenges that professionals are expensive participants and that the experimental data-mining tool should be

realistic with regard to algorithms, data, and discrimination-aware modifications. It may also be difficult to carry out meaningful experimental studies; different study designs may be needed. First, the introduction of new AI is usually accompanied by other organisational changes such as staff reductions⁶⁰. Second, while statistics and data mining have a long tradition in fields such as banking and insurance, the use of such methods is less wide-spread in anti-discrimination agencies, NGOs, and similar organisations, and also by data protection authorities. Algorithmic accountability reporting⁶¹, in which the media act as Boven's forum⁴⁶, is a development in this direction. However, additional research is needed to identify the best algorithmic, HCI, and generally organisational design decisions for these different use cases. Third, real-life decisions are often not made at one point in time only. The law does not forbid fully automated decision making, but (in the EU) allows data subjects to request a human explanation and rectification. This points to another area of future research: how do humans reason about machine-based decisions ex post? Under which cognitive biases and organisational pressures do they revisit an already-made decision, and how can HCI assist?

Acknowledgements

We thank the participants of the Dagstuhl 2016 seminar "Data, responsibly ..." as well as those of the 2016 Summer School Big Data for Peace and Justice, and Patrick Berendt for valuable discussions.

Disclosure statement

No competing financial interests exist.

References

1. Angwin J., Larson J., Mattu S., Kirchner L. Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica, May 23, 2016. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
2. Lowry S., Macpherson, G. A blot on the profession, 296 BRIT. MED. J. 657, 657. 1988.
3. Sweeney L. Discrimination in Online Ad Delivery. Comm of the ACM, 56(5), 44-54. 2013.
4. Stark J., Diakopoulos N. Uber seems to offer better service in areas with more white people. That raises some tough questions. The Washington Post Wonkblog, March 10, 2016. <https://www.washingtonpost.com/news/wonk/wp/2016/03/10/uber-seems-to-offer-better-service-in-areas-with-more-white-people-that-raises-some-tough-questions/>
5. Barocas S., Selbst A.D. Big Data's Disparate Impact. 104 California Law Review 671. 2016. Available at SSRN: <http://ssrn.com/abstract=2477899>
6. Datta A., Tschantz M.C., Datta A. Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination. Proc. PETS 2015. 2015.

7. UN Women. UN Women ad series reveals widespread sexism. 2013.
<http://www.unwomen.org/en/news/stories/2013/10/women-should-ads>
8. Crawford K. Think Again: Big Data. Foreign Policy. May 10, 2013.
http://www.foreignpolicy.com/articles/2013/05/09/think_again_big_data
9. Pedreschi D., Ruggieri S., Turini F. A study of top-k measures for discrimination discovery. Proc. SAC '12, pp 126–131. ACM, New York, NY, USA. 2012.
10. Dwork C., Hardt M., Pitassi T., Reingold O., Zemel R. S. Fairness through Awareness. Proc. ITCS 2012, 214-226. 2012.
11. Zliobaite I. Measuring discrimination in algorithmic decision making. Data Mining and Knowledge Discovery. DOI: 10.1007/s10618-017-0506-1. 2017.
12. Pedreschi D., Ruggieri S., Turini F. Discrimination-aware data mining. Proc. KDD'08, pp 560–568. ACM. 2008.
13. Hajian S., Bonchi F., Castillo C. Algorithmic bias: from discrimination discovery to fairness-aware data mining. Tutorial at KDD 2016. 2016.
http://francescobonchi.com/algorithmic_bias_tutorial.html
14. Stoyanovich J., Abiteboul S., Miklau G. Data, Responsibly: fairness, neutrality and transparency in data analysis. Tutorial at EDBT 2016. 2016.
<http://abiteboul.com/PRESENTATION/16.DataResponsibly.EDBT.pdf>
15. Berendt B., Preibusch S. Better decision support through exploratory discrimination-aware data mining: foundations and empirical evidence. Artificial Intelligence and Law, 22 (2), 175-209. 2014.
https://people.cs.kuleuven.be/~bettina.berendt/Papers/berendt_preibusch_2014.pdf
16. Raab C. The Meaning of “Accountability” in the Information Privacy Context. In D. Guagnin, L. Hempel, C. Ilten, I. Kroener, D. Neyland, H. Postigo (Eds.). Managing Privacy through Accountability, London: Palgrave Macmillan. 2012.
17. Berendt B., Preibusch S. Exploring discrimination: A user-centric evaluation of discrimination-aware data mining. IEEE 12th Intern. Conference on Data Mining Workshops (pp. 344-351). IEEE Computer Science Press. 2012.
18. Eickhoff C., de Vries A.P. Increasing cheat robustness of crowdsourcing tasks. Inf Retr 16(2): 121–137. 2013.
19. Paolacci G., Chandler J. Inside the Turk: Understanding Mechanical Turk as a participant pool. Current Dimensions in Psychological Science, 23(3), 184-188. 2014.
20. The Verge. Twitter taught Microsoft's AI chatbot to be a racist asshole in less than 24 hours. Mar 24, 2016. <http://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist>
21. Lum K., Isaac W. To predict and serve? Significance, 13 (5), 14-19. 2016.
22. Hajian S., Domingo-Ferrer J. A methodology for direct and indirect discrimination prevention in data mining. IEEE Trans Knowl Data Eng 25(7):1445–1459. 2013.
23. Ruggieri S., Pedreschi D., Turini F. DCUBE: discrimination discovery in databases. In Proc. SIGMOD'10, pp 1127–1130. 2010.
24. Calders T., Verwer S. Three naive Bayes approaches for discrimination-free classification. Data Min Knowl Discov 21(2):277–292. 2010
25. Feldman M., Friedler S.A., Moeller J., Scheidegger C., Venkatasubramanian, S. Certifying and Removing Disparate Impact. Proc. KDD 2015: 259-268. 2015.

26. Romei A., Ruggieri S., Turini F. Discrimination discovery in scientific project evaluation: A case study. *Expert Systems with Applications*, 40(15), 6064-6079. 2013.
27. Myrdal G. *An American Dilemma: The Negro Problem and Modern Democracy*. New York: Harper & Bros. 1944.
28. Adda C.L., Laitin D.D., Valfort M.-A. *Why Muslim Integration Fails in Christian-Heritage Societies*. Cambridge, MA: Harvard University Press. 2016.
29. Luong B.T. Generalized discrimination discovery on semi-structured data supported by ontology. PhD thesis, IMT Institute for Advanced Studies, Lucca, Italy. 2011
30. Kamiran F., Zliobaite I., Calders T. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowl Inf Syst* 35(3):613–644. 2013.
31. Kamishima T., Akaho S., Asoh H., Sakuma J. Considerations on fairness-aware data mining. *IEEE 12th Intern. Conference on Data Mining Workshops* (pp. 344-351). IEEE Computer Science Press. pp 378–385. 2012.
32. Romei A., Ruggieri S. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*. 29(5), 582-638. 2014.
33. Insley J., Bachelor L. Payday loans can put credit rating at risk. *The Guardian*, 17 Nov. 2012. <https://www.theguardian.com/money/2012/nov/17/payday-loans-credit-rating>
34. Bell, D. Racial realism. *Connecticut Law Review*, 24, 363–379. 1992.
35. Morey M. Are Americans Really Champions of Racial Equality? Apr 12, 2015. <http://www.theatlantic.com/politics/archive/2015/04/are-americans-champions-of-racial-equality/389826/>
36. Bygrave L.A. Minding the Machine: Article 15 of the EC Data Protection Directive and Automated Profiling. *Computer Law & Security Report*, 17, 17–24. 2001.
37. OLG München, Urt. v. 12.03.2014, Az. 15 U 2395/13
38. OLG Frankfurt, Urt. V. 14.12.2015, Az. 1 U 128/15
39. ErfK/Franzen. *Erfurter Kommentar zum Arbeitsrecht* Rn. 2-2. 16. Auflage 2016. BDSG § 6a
40. O'Neill C. How algorithms rule our working lives. *The Guardian*. Sep 1, 2016. <https://www.theguardian.com/science/2016/sep/01/how-algorithms-rule-our-working-lives>
41. Kleinberg J., Mullainathan S., Raghavan, M. Inherent Trade-Offs in the Fair Determination of Risk Scores, In *Proc. 8th Conf. on Innovations in Theoretical Computer Science (ITCS)*. 2017.
42. ECHR/European Court of Human Rights. Judgment. *Taddeucci and McCall vs. Italy*. Jun 30, 2016.
43. OECD. Guidelines on the protection of privacy and transborder flows of personal data. 1980. http://www.oecd.org/document/18/0,2340,en_2649_34255_1815186_1_1_1_1,00.html
44. ISO/IEC 29100:2011. Information technology -- Security techniques -- Privacy framework. 2011. http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=45123
45. Van Alsenoy B., Coudert F., Jasmontaite L., Verdoodt V., Valcke P., Kindt K., Vedder A. *Cultures of Accountability*. 2015. <https://www.law.kuleuven.be/citip/en/news/item/coa-workshop-report.pdf>
46. Bovens M. Analysing and Assessing Public Accountability. A Conceptual Framework. *European Governance Papers (EUROGOV)* No. C-06-01. 2006. <http://www.connex-network.org/eurogov/pdf/egp-connex-C-06-01.pdf>

47. Article 29 Data Protection Working Party. Opinion 3/2010 on the principle of accountability (00062/10/EN WP 173). 2010.
http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2010/wp173_en.pdf
48. Crespo García A. et al. PRIPARE. Privacy- and Security-by design Methodology Handbook. 2016. <http://pripareproject.eu/wp-content/uploads/2013/11/PRIPARE-Methodology-Handbook-Final-Feb-24-2016.pdf>
49. Hoepman J.-H. Privacy design strategies – (extended abstract). Proc. ICT Systems Security and Privacy Protection - 29th IFIP TC 11 International Conference. pp. 446–459, 2014.
50. Butin D., Chicote M., Le Métayer D. Strong Accountability: Beyond Vague Promises. In *Reloading Data Protection: Multidisciplinary Insights and Contemporary Challenges*. pp.343-369. Springer. 2014.
51. Diakopoulos N., Friedler S. How to hold algorithms accountable. MIT Technology Review, 16 Nov 2016. <https://www.technologyreview.com/s/602933/how-to-hold-algorithms-accountable/>
52. Kroll J. A., Huey, J., Barocas, S., Felten, E.W., Reidenberg, J.R., Robinson, D.G., Yu, H. Accountable algorithms. *University of Pennsylvania Law Review* 165: 633-705. 2017.
<https://ssrn.com/abstract=2765268>
53. Lipton Z.C. The Mythos of Model Interpretability. ICML 2016 Workshop on Human Interpretability in Machine Learning (WHI 2016). 2016.
http://zacklipton.com/media/papers/mythos_model_interpretability_lipton2016.pdf
54. Jameson A., Berendt B., Gabrielli S., Cena F., Gena C., Vernerio F., Reinecke K. Choice Architecture for Human-Computer Interaction. *Foundations and Trends in Human-Computer Interaction*, 7(1-2), 1-235. 2014.
55. Huff C., Tingley D. "Who are these people?" Evaluating the demographic characteristics and political preferences of MTurk survey respondents. *Research and Politics*, July-September 2015, 1-12.
56. Clifford S., Jewell R.M., Waggoner P.D. Are samples drawn from Mechanical Turk valid for research on political ideology?. *Research and Politics*, October-December 2015. 1-9.
57. Goodman J., Cryder C., Cheema A. Data collection in a flat world: the strengths and weaknesses of Mechanical Turk samples. *J Behav Decis Mak*, 26, 213–224. 2012.
58. McNamara G., Bromiley P. Decision making in an organizational setting: Cognitive and organizational influences on risk assessment in commercial lending. *Academy of Management Journal*, 40(5), 1063-1088. 1997.
59. Berinski A.J., Huber G.A., Lenz G.S. Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis*, 20, 351-368. 2012.
60. McCurry J. Japanese company replaces office workers with artificial intelligence. *The Guardian*, 5 Jan 2017. <https://www.theguardian.com/technology/2017/jan/05/japanese-company-replaces-office-workers-artificial-intelligence-ai-fukoku-mutual-life-insurance>
61. Diakopoulos N. Algorithmic Accountability Reporting: On the Investigation of Black Boxes. *Tow Center for Digital Journalism*. 2013. http://www.nickdiakopoulos.com/wp-content/uploads/2011/07/Algorithmic-Accountability-Reporting_final.pdf