"I agree with the decision, but they didn't deserve this"; Future Developers' Perception of Fairness in Algorithmic Decisions

Maria Kasinidou
Cyprus Center for Algorithmic Transparency,
Open University of Cyprus
Nicosia, Cyprus
maria.kasinidou@ouc.ac.cy

Styliani Kleanthous*
Cyprus Center for Algorithmic Transparency,
Open University of Cyprus
Nicosia, Cyprus
styliani.kleanthous@ouc.ac.cy

Pınar Barlas
Research Centre on Interactive Media,
Smart Systems and Emerging Technologies
Nicosia, Cyprus
p.barlas@rise.org.cy

Jahna Otterbacher*
Cyprus Center for Algorithmic Transparency,
Open University of Cyprus
Nicosia, Cyprus
jahna.otterbacher@ouc.ac.cy

ABSTRACT

While professionals are increasingly relying on algorithmic systems for making a decision, on some occasions, algorithmic decisions may be perceived as biased or not just. Prior work has looked into the perception of algorithmic decision-making from the user's point of view. In this work, we investigate how students in fields adjacent to algorithm development perceive algorithmic decision-making. Participants (N=99) were asked to rate their agreement with statements regarding six constructs that are related to facets of fairness and justice in algorithmic decision-making in three separate scenarios. Two of the three scenarios were independent of each other, while the third scenario presented three different outcomes of the same algorithmic system, demonstrating perception changes triggered by different outputs. Quantitative analysis indicates that *a*) 'agreeing' with a decision does not mean the person 'deserves the outcome', *b*) perceiving the factors used in the decision-making as 'appropriate' does not make the decision of the system 'fair' and *c*) perceiving a system's decision as 'not fair' is affecting the participants' 'trust' in the system. In addition, participants found proportional distribution of benefits more fair than other approaches. Qualitative analysis provides further insights into that information the participants find essential to judge and understand an algorithmic decision-making system's fairness. Finally, the level of academic education has a role to play in the perception of fairness and justice in algorithmic decision-making.

CCS CONCEPTS

• **Human-centered computing** → *Empirical studies in HCI*.

KEYWORDS

algorithmic fairness, algorithmic transparency, algorithmic accountability, algorithmic decision-making

ACM Reference Format:
Maria Kasinidou, Styliani Kleanthous, Pınar Barlas, and Jahna Otterbacher. 2021. "I agree with the decision, but they didn't deserve this"; Future Developers' Perception of Fairness in Algorithmic Decisions . In *ACM Conference on Fairness, Accountability, and Transparency (FAccT '21), March 1–10, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3442188.3445931

1 INTRODUCTION

Algorithmic decision-making is widely used for contributing to decisions affecting people's lives. Job hiring [36], healthcare [27], education [6], finance [25] and criminal justice[1] [2, 9], are just a few of the examples where algorithms are taking on what previously were human decision-making tasks. An algorithm is even deciding on the posts and news people will see on social media [40, 44]. While the use of algorithmic decision-making has prospects to make decision-making more efficient and reliable [11, 28], concerns have been raised about the fairness and justice of such decisions.

Algorithmic decision-making systems do not always behave as they should, making decisions that may discriminate against certain groups of people. There are many examples in different domains that show the misbehavior of these systems: gender discrimination has been detected in a recruitment system for reviewing and ranking applicants' resumes [12] and in resume search engines [8]; auto-complete search terms can produce suggested terms which could be viewed as racist, sexist, or homophobic [3]; image search results are gender-biased depending on the search term used [37] and racially-biased towards Black individuals [1, 29].

There has been an increasing focus in the research community from various disciplines on promoting and understanding fairness in algorithmic decision-making. While much effort has been devoted to developing frameworks of fairness [9, 16] and algorithmic models to alleviate biases [30, 31, 50], there is a need to understand how algorithmic fairness is perceived by people [5, 20, 24, 32, 38, 48, 49].

*Also with the Research Centre on Interactive Media, Smart Systems and Emerging Technologies, Nicosia, Cyprus.

[1] www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

FAccT '21, March 1–10, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8309-7/21/03…$15.00
https://doi.org/10.1145/3442188.3445931

While related work has looked into how the end users and/or the general public perceive algorithmic fairness, it is important to understand how the people who are developing – or will soon be involved in developing – algorithmic decision-making systems perceive algorithmic fairness. To our knowledge, perception of fairness in algorithmic decision-making of students' from fields adjacent to computing has not been studied previously.

To explore how future developers perceive algorithmic fairness, we conducted an online survey with students in fields adjacent to algorithm development, who will potentially be involved in the development of an algorithmic decision-making system. We presented participants with three scenarios of algorithmic decision-making systems describing different contexts and asked them to indicate their agreement regarding six statements related to the fairness and justice constructs [10, 32]. We then analyzed their responses in order to understand the interplay between these constructs in relation to the scenarios. Two of the three scenarios, their contents independent of each other, were used to trigger the participants' judgement on the use of particular factors used for decision-making. In the third scenario, participants were presented with the description of a single system and three different cases of algorithmic decision. The purpose of this scenario was to examine whether participants' perception was affected when presented with different outputs.

Our findings indicate that even when participants 'agreed' with the decision made by the algorithm they did not believe that the person in the scenario 'deserved the outcome'. Moreover, even when factors used in the decision-making were perceived as 'appropriate,' the overall process followed by the system was perceived as 'non-fair' by the participants. In addition, systems that were perceived as 'not fair' affected participants' 'trust' in the system. In the third scenario, participants show a preference for the proportional ('ratio') decision, as compared to the other two decisions (giving all the money to one candidate, splitting the money equally). Our results suggest that the level of education can change participants' understating of the process, their agreement with the decision and appropriateness of certain factors used by the algorithm. Finally, qualitative analysis shows that future developers, in order to judge fairness of a given algorithmic system find it essential to know more information about the *Factors* used and the *Process* followed in the decision-making, and whether *Sensitive Attributes* (e.g. age, race, gender) were used in the decision. Overall, our findings note the complexity of understanding perceptions of fairness in algorithmic decision-making even from the developer's perspective.

2 BACKGROUND

With the increasing use of algorithms for making or supporting managerial decisions, those that humans used to make, researchers need to understand how not only users but also *developers* perceive these algorithms. Regardless of an algorithm's performance in terms of correctness and accuracy, the perception people have of these algorithms can influence their adoption. There is a growing body of work looking into perception of algorithmic decision-making, fairness perception and trust from the end-user point of view. However, an equally important and interesting aspect is how developers of such algorithms perceive algorithmic decision-making and in extent fairness of algorithmic decisions.

2.1 Fairness in Algorithmic Decision Making

Fairness is a complex construct consisting of several parameters that are considered by individuals when they are trying to define fairness in different contexts. There is a lot of work on fairness and justice[2] constructs in the psychology literature (see [10] for a comprehensive review). With regards to algorithmic fairness in decision-making, recent work in HCI and FAccT has looked into perceived fairness, only to find that people understand fairness differently and according to the context [26] where the system is operating. Perceived fairness is multi-dimensional and there is a lack of consensus on which features are perceived as unfair by different people [20]. Thus, usually scenario-based studies are employed in order to provide study participants a framing when asked to define fairness [5, 32, 33, 49].

Studies revealed contradicting results as to what is perceived as fair. People perceive algorithmic decision-making as less fair than human decision-making even when the decision requires 'human skills' [32] or more fair in other contexts, like school admissions [35]. Algorithms and systems should consider social and altruistic behavior in order to be considered as fair, elements that may be difficult to incorporate in mathematical modelling [8]. People tend to rate models as unfair when they consider them biased (and vice versa), and prefer human decision-making even if they consider the algorithmic model as fair or unbiased [23]. Accuracy was rated as more important than equality in [42], with demographic parity best represented people's understanding of fairness. Certain attributes are not considered fair when used in defining the outcome of a system in a certain context [41], suggesting that the use of features and attributes upon which decisions are made are context-dependent as well as output-dependent and can be perceived as fair or unfair accordingly [19, 32].

As users are becoming more aware of the concept of algorithmic fairness, they are starting to worry about potential biases in the decision as well as in the data or the algorithm interaction [7]. They seek more information about how different factors weighted in the decision and whether an algorithm uses sensitive attributes (such as race or gender). Variables — such as the computer literacy and favourable outcome, as well as development procedures of the system – have also been proven to correlate with the perception of algorithmic fairness [47]. In particular, people rate the algorithm as more fair when the decision is in their favour, irrespective of whether it appears to be biased towards certain social groups. In the same vein, Pierson showed that there are gender differences in perceptions of algorithmic fairness, while demographic differences contribute to the variability of opinions on fairness [38].

Education and training on algorithmic fairness appears to have an effect on students, whose perception of fairness changed after an hour-long lecture and discussion on algorithmic fairness [38]. However, in order for algorithms to become more fair, developers need to be educated and become aware of the potential biases and discrimination that can occur due to the algorithms they develop. They need to be in a position to understand the source of bias and perform the right steps to overcome it. The challenge developers have to face, though, is that with the use of machine learning (ML) approaches becoming more and more popular as a driving tool

[2] These terms are usually used interchangeably in the literature.

in algorithmic decision-making, it can get even more difficult to trace the source of bias in the system. Hutchinson and Mitchell, in their review of 50 years of work on (un)fairness [26] suggest that current and future work on ML should be informed by prior work and findings, rather than trying to generally define a 'fair' model. They urge ML researchers to look into questions that are deeper, define criteria that are context- and use-dependent, and question whether all subgroup dimensions (e.g. gender, age) can be served in one model or if a different approach might be required. Thus, in this study, we are interested in examining how future developers perceive certain constructs linked to fairness in algorithmic systems and the interplay of those in specific scenarios.

2.2 Explanations in Decision Making

Opacity [18] of algorithmic decision-making and whether different transparency approaches might enhance the perceptions of fairness of those systems [13], has also been a cause for discussion in the recent literature. With ML models being exploited for predicting sensitive individual information, classifying individuals in categories and providing decisions that were previously taken by humans [21], there is a need for interpreting those models in a way that the user would understand.

In light of the recently drafted General Data Protection Regulation (GDPR), people who interact with systems that involve automated decision-making have a right to obtain 'meaningful explanations of the logic involved'. Hence, there is a need for methods and approaches that will allow the user to understand the output of these opaque and automated processes in context [15]. At the same time, the user should be able to understand the potential consequences of applying the decision in the real world [21]. Hence, explanations should be informative and easy to be interpreted by the person they are created for. Edwards and Veale [17] argue that pedagogical approaches to explanation – explanations that teach how the model works – might be more promising for the general public than decompositional approaches – breaking the model down with the risk of trading secrets and intellectual property breach. Furthermore, ML tools, such as debiasing or transparency systems, will also need to take into consideration the contextual challenges early on [46].

Explaining a system's decision though, is not trivial. Different level of explanation is needed according to the audience and the purpose especially for black-box models [17]. According to [22] local explanations focus on explaining a particular output; global explanations explain how a set of outputs emerges from a particular input; and counterfactual explanations attempt to help the user understand how their input could change the output of the system by resembling everyday human conversation. Studies with users however are inconclusive as to what type and level of explanation they prefer. Binns et al. [5] ran an experiment using different explanation styles (input influence, sensitivity, case-based, demographics). They found significant differences in justice perception between different explanation styles. Particularly, case-based explanations – presenting a case from the model's training data, which is most similar to the decision – affected the judgments of justice negatively compared to sensitivity-based explanations – explaining how much the value of a variable used in the model affects the output. However, when people were exposed to the same explanation style in different scenarios, they observed none of the above. Rader et al. [39] found that explanations, in any form, help to create awareness of how the system works and understand potential bias in the system's output, but offer little in evaluating the correctness of the output. Explanations in group recommendations have been proven to improve the perception of fairness when all or the majority of group members' preferences are taken into account [45], emphasizing how fairness is subjective to each individual person.

The challenge of dealing with, and explaining, potentially harmful outputs has been demonstrated in several occasions. Take as an example the Google photos incident, where a Black American and his friend were mistakenly labeled by the system as 'gorillas'[3]. After a two-year effort at Google to 'solve' the problem, the final solution was just a work-around of removing the label from their lexicon. This demonstrates the difficulties that companies like Google, and in extend their developers, face in understanding and explaining possible unwanted decisions of their own ML-based systems. Holstein et al. [24] provide some important insights on how developers are struggling to find a balance between fairness in their systems and providing a product for their companies. They are calling for procedures, processes and training on concepts related to fairness, accountability, transparency and ethics for developers who are already in the business.

Thus, it is important for us in this work to understand how future developers deal with certain explanations provided in the scenarios they were given. Furthermore, we look into how different decisions change their perception of fairness, and whether the context and the output also have an impact on their perceptions.

3 METHODOLOGY

In order to understand how future developers perceive fairness in algorithmic decision-making, we conducted an online survey that ran between September 2019 and May 2020.

3.1 Scenarios

Participants were presented with three scenarios where algorithms made decisions that influenced humans. We selected contexts that our target population is familiar with. Two of the three scenarios were used to trigger the participants' judgement on the use of particular factors (e.g. demographics) considered for decision-making and explanations of the decision given. In the third scenario, three different decisions were presented with the purpose of examining whether participants' perception changes according to different outcomes.

- **Scenario 1:** A car insurance company's premiums dynamically-priced, based on personal details and driving behaviour. This scenario was adopted from Binns et al. [5].
- **Scenario 2:** Passengers on over-booked airline flights being automatically selected for re-routing:

  "Airline X is using a system for automatically selecting and rerouting passengers on overbooked flights based on the passenger's marital status, number of children the

[3] https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/

passenger has, whether they are part of a group booking, and their age and gender.
Based on the above information the system decided to reroute Frank, who was single, traveling alone, was 55 years of age, male instead of Lisa, who was single, traveling alone, 35 years old, female."

- **Scenario 3:** Applying for a personal financial loan. This scenario was adapted from Saxena et al. [41].
  "There are two candidates - Person A and Person B, they are identical in every way, except their race and loan repayment rates. Both of them have applied for a $50,000 loan to start a business, and the loan officer only has a $50,000."
  1. ***Case A**: "Taking into consideration the Gender, Race and Individual loan repayment rate, the system decided to split the money 50/50 between the two candidates giving $25,000 to Person A and $25,000 to Person B."*
  2. ***Case B**: "Taking into consideration the Gender, Race and Individual loan repayment rate, the system decided to give Person A $31,818, which is proportional to that person's payback rate of 70%, and give Person B $18,181, which is proportional to that person's payback."*
  3. ***Case C**: "Taking into consideration the Gender, Race and Individual loan repayment rate, the system decided to give all the money to Person A."*

For each scenario, participants were asked to rate their agreement in five statements according to [10] in addition to 'Trust'. A 5-point Likert scale, ranging from '1 - Strongly Disagree' to '5 - Strongly Agree', was employed for each of the six statements:

S1 Agreement: "I agree with the decision"
S2 Understanding: "I understand the process by which the decision was made"
S3 Appropriateness of factors: "The factors considered in the decision were appropriate"
S4 Fair process: "The decision-making process was fair"
S5 Deserved outcome: "The individual deserved this outcome given their circumstances or behaviour"
S6 Trust: "I would trust this system's decision more than a human's decision"

Participants were also asked to explain using free-text (Q1) *"Was the information provided in the above scenario sufficient?"* Participants free-text responses were coded as 'Yes', 'No', 'Unsure' and thematically analysed [43]. Finally, participants self-reported (*Yes/No/Other* write-in) whether they have taken (Q2) *"any training/course on Fairness, Accountability and Transparency in algorithmic systems"* and (Q3) assessed their knowledge on Fairness in algorithmic decision-making systems using a 5-point Likert scale (1, Not at all - 5, Very Knowledgeable).

3.2 Participants

We recruited respondents using snowball sampling. We emailed the survey to colleagues at other universities all over the world inviting them to pass the survey on to their students. We also shared the survey on our social media accounts, where the authors have a lot of computing-related students as connections. We recruited 100 undergraduate and postgraduate students from the fields related to Computer Science. One participant was removed due to providing non-serious answers, thus 99 respondents were considered. Participation was voluntary and all participants provided us with written, informed consent for their data to be used. The study received ethical clearance from the national ethics committee of the country where the authors' institution is operating.[4]

60.6% of our respondents were male, with 47.5% in the age group of 18-24, 35.4% between 25-32, 10.1% between 33-40, and 7.1% above 40 years old. Most of the participants (68.7%) identified themselves as a postgraduate student, and 54% of that group were Master's students. The rest of the participants were self-identified as undergraduate students, of them 58.1% being in their third or fourth year and 41.9% being in their first or second year of studies. The majority of the participants are enrolled in the following degree programs: 49% in Computer Science, 27% in Information Systems, 8% in Data Science, 7% in Machine Learning/Artificial Intelligence, 4% in Human-Computer Interaction/Human-Robot Interaction, 2% in Computer Science with Mathematics, and 5% in other programs. The majority of participants are studying at institutions in Europe 45.4% and the UK 40.4%, 7% in the USA, 4% in Israel, and 3% in China, Brazil and Australia.

4 QUANTITATIVE FINDINGS

Table 1: Descriptive Statistics for the variables used in the analysis

	Mean	Std. Deviation
Agreement	2.7232	.62316
Understanding	3.4798	.87531
Appropriateness	2.6040	.75267
Fair	2.6242	.69062
Deserved	2.5838	.62966
Trust	2.4788	.79581

Quantitative analysis was employed in order to understand participants' perception of each individual construct for Scenario 1 and Scenario 2, and to examine whether their perception changes if they are presented with the same scenario but a different algorithmic decision (Scenario 3).

Do constructs (Agreement, Understanding, Appropriateness, Fair process, Deserved Outcome and Trust) correlate across different algorithmic decision-making scenarios? Based on the literature on perceived fairness and justice [10] and recent work on perceptions of algorithmic justice [5], we expected to find correlations between all constructs. To examine this we calculated Pearson correlations. Descriptive statistics for all of the variables used in the Pearson correlations are available in Table 1. Although we were expecting that all constructs will correlate, similar to [5], we were surprised to see that understanding of the process followed correlates with appropriateness of the factors, and understanding of the process with deserved outcome (see Table 2).

[4] We do not explicitly mention the name of the committee for anonymity purposes. We will revisit upon acceptance.

Table 2: Pearson Correlations for the six constructs of justice

		Agreement	Understanding	Appropriateness	Fair	Deserved	Trust
Agreement	Pearson Correlation	1	.413 **	.682 **	.765	.795 **	.574 **
	Sig. (2-tailed)		.000	.000	.000	.000	.000
Understanding	Pearson Correlation	.413 **	1	.401 **	.365 **	.319	.222 **
	Sig. (2-tailed)	.000		.000	.000	.001	.027
Appropriateness	Pearson Correlation	.682 **	.401 **	1	.678 **	.691 **	.535
	Sig. (2-tailed)	.000	.000		.000	.000	.000
Fair	Pearson Correlation	.765 **	.365 **	.678 **	1 **	.792 **	.703 **
	Sig. (2-tailed)	.000	.000	.000		.000	.000
Deserved	Pearson Correlation	.795 **	.319 **	.691 **	.792 **	1 **	.685 **
	Sig. (2-tailed)	.000	.001	.000	.000		.000
Trust	Pearson Correlation	.574 **	.222 *	.535 **	.703 **	.685 *	1 **
	Sig. (2-tailed)	.000	.027	.000	.000	.000	

4.1 Perception and Interplay of Constructs

To examine a number of hypotheses regarding participants’ perception of the Fairness constructs in Scenarios 1 & 2, we run a series of Wilcoxon signed ranked tests.

People who agreed with the decision also believe that the person in the scenario deserved the outcome. We were expecting that the people who indicated agreement with the system’s decision would also believe that the person in the scenario deserved the outcome. Surprisingly we found significant statistical differences in their opinions (Scenario 1 : z=2.70, p=0.007; Scenario 2: z=4.043, p<0.001). In Scenario 1 there was a considerable number of participants (37.4%) who selected options 4 and 5 on the Agreement scale, while 49.5% selected options 1 and 2 on the Deserved scale, indicating that they agreed with the decision but the person in the scenario did not deserve the outcome. In scenario 2 fewer participants but still a considerable number (18.2%) selected options 4 and 5 on the Agreement Scale indicating they agree with the decision, while 61.6% selected options 1 and 2 on the Deserved scale.

People who found the factors used in the decision making process appropriate will also think that the decision making process is fair. The results show significant differences between the responses of the participants in Scenario 1 (z=-3.193, p<0.001) with participants in their majority (42.4% selected 4 and 5 on the scale for S3) reporting that the factors used in the decision-making were appropriate, however, they do not believe that the decision-making process was fair (47.5% selected 1 and 2 on the scale for S4).

In Scenario 2 we do not have a statistical significant difference between the two scales, where participants in their majority (52.5%) agree that the factors used in the decision making processes were not appropriate, and 56.6% believe that the decision making process was not fair. Qualitative results (see below) show that in Scenario 2, participants felt that the use of gender and age as factors to determine the decision were not appropriate, which explains this result.

People who indicated the the decision making process was not fair would not trust this system’s decision more than a human’s decision. For both Scenario 1 and Scenario 2 we did not get any significant differences between the two scales. Specifically, 47.5% of the participants in Scenario 1 and 56.6% of the participants in Scenario 2 believe the decision making process was not fair, and 40.4% of the participants in Scenario 1 and 51.6% of the participants in Scenario 2 would not trust the system’s decision more than a human’s.

Next, we wanted to examine whether the different decisions in Scenario 3 (Case A, Case B and Case C) affected participants’ perception of the above constructs.

Does the participants’ perception of Agreement, Understanding, Appropriateness, Fair Process, Deserved Outcome and Trust change according to the decision of the system (given the same scenario)? To compare the responses in Scenario 3, we followed a within-subject analysis using ANOVA repeated measures followed by a Bonferroni post-hoc test. There were significant differences for (Agreement (F(2,196)=29.272, p<0.001); Appropriateness (F(2,196)=17.646, p<0.001); Fairness (F(2,196)=30.437, p<0.001); Deserved Outcome (F(2,196)=28.751, p<0.001) and Trust (F(2,196)=9.992, p<0.001)) in responses provided by the participants. Bonferroni post-hoc tests showed that participants perceived the decision in Case B (proportional outcome) as the most just, while the decision on Case C as the least.

Similarly, comparing their responses in question Q1 in all three cases in Scenario 3, we observed significant statistical differences (F(2,196)=15.556, p< 0.001) with the post-hoc test revealing that participants felt that the information provided in Case B was perceived as sufficient. 44.4% indicated sufficient information provided in Case B compared to 28.3% in Case A and 21.1% in Case C.

In all scenarios, we did not find any differences in the participants responses between self-reported gender groups. Previous training and self reported knowledge on topics related to algorithmic decision making did not have an impact on the responses of participants in our sample.

4.2 Differences between Undergraduate and Postgraduate Participants

Since this study ran with undergraduate and postgraduate students in fields adjacent to algorithmic development it is natural to examine whether the participants' level of education made a difference in their responses. A series of Mann-Whitney U tests were run to determine if there were differences between the two groups. Distributions of the engagement scores for undergraduates and postgraduates were similar, as assessed by visual inspection in all cases.

Firstly, we wanted to examine whether there is a difference between undergraduates and postgraduates in understanding the process by which the decision was made. Scenario 1 was the only scenario where statistical significant difference in understanding were found. Median engagement score was moderately statistically significantly higher in postgraduates than in undergraduates, (U = 1331, z = 2.07, p = 0.038), indicating that postgraduates understood the process that the system is following in making a decision better compared to undergraduates. There was no significant difference between the two groups with respect to the other parameters.

Since we had indications from the previous analysis where the different cases in Scenario 3 perceived differently, we wanted to see whether there is a difference between undergraduates and postgraduates in the perception of sufficiency of information provided. In Case A and Case C we did not find any significant differences between the two groups. In Case B median engagement score was statistically significantly lower in undergraduates (0.5) than in postgraduates (1.00), U = 764, z = -2.48, p = 0.013, indicating that undergraduates find the information provided less sufficient in this case compared to postgraduates.

Furthermore, we examined whether there is a difference in the agreement with the decision between undergraduates and postgraduates in our sample. In Case A and Case B we did not find any statistically significant differences between the two groups. For Case C, median engagement score was statistically significantly lower in undergraduates (1.00) than in postgraduates (2.00), U = 813.5, z = -2.043, p = 0.041, indicating that undergraduates agreed less with the decision of the system compared to the postgraduates.

Following the same line of thought, we examined whether there is a difference in the perception of appropriateness of the factors considered for the system's decision between undergraduates and postgraduates. Similar to above, in their responses regarding Case A and B we did not have any significant differences. In Case C median engagement score was statistically significantly lower in undergraduates (1.00) than in postgraduates (2.00), U = 795.500, z = -2.185, p = 0.029, indicating that undergraduates considered the factors used in the system for making the decision less appropriate compared to the postgraduates.

Finally, there is a marginal statistical difference between undergraduates and postgraduates in their indication of whether the decision-making process was fair in Case C. Median engagement score was statistically significantly lower in undergraduates (1.00) than in postgraduates (2.00), U = 813, z = -2.06, p = 0.039, indicating that undergraduates considered the decision-making process less fair compared to the postgraduates.

5 QUALITATIVE FINDINGS

For Q1, participants were asked whether they had sufficient information. The free-text responses that simply stated a 'yes'/'no' were excluded from the qualitative analysis. To analyse participants' free-text responses we used content analysis, by coding responses for the themes mentioned in relation to the concepts in question. Two researchers analyzed the responses independently to define emerging categories. We allowed multiple categories per answer. The categories identified by the two researchers were then compared, the disagreements discussed, and sometimes a dimension's definition amended to come to a final consensus.

5.1 Scenario 1

59 participants elaborated on their response to Q1 for Scenario 1, where six thematic areas emerged from their responses (Table 3). Most often, participants discussed **Missing Factors**: important factors about the situation that were not taken into consideration. These included *"context of the day of accident, time, [weather]"* (participant 75 - p75), *"road infrastructures"* (p46), *"[driver's] attitude [and] her family history"* (p73), and *"condition of the car"* (p91). Interestingly, some participants even mentioned the need to consider other factors even when they indicated they found the information sufficient.

17 of the 59 participants referred to the **Similar Cases** on which the prompt said the decision was based. Although the prompt explicitly stated twice that *"[the] decision was based on thousands of similar cases from the past"* and went on to give one similar case only as an example, participants often remarked that *"a single example is not enough to adequately explain decisions"* (p78). Some participants questioned the exact number of cases in the dataset (p24), seemingly arguing what others explicitly stated: *"If the data is quite large, I think the decision is trustful"* (p21).

The third most common theme was the decision-making **Process** with a total of 15 responses. Most participants wanted to know *"how much each factor contributed to the decision"* (p80), some specifically asking for *"additional explanation on how age, driving at night etc. affects the probability of having an accident"* (p68).

A few participants (9/49) wanted **Specific Information** which seemed to be missing from the scenario, such as the *"criteria"* (p85) or cost (p67) of the cheapest tier, as well as more examples of similar cases (p24). These participants did not ask about other factors missing from the scenario, but for the specific values of factors already mentioned.

The remaining themes received few responses. Three participants mentioned the need to think about the **Human/Company Policy** of the scenario, such as participant 73 who said that a human being would able talk to the driver and better understand driver's attitude. 7 responses fell under the catch-all **Other** category, which includes responses that do not mention the other themes or responses where the participant indicated they *"don't understand the question"* (p76).

5.2 Scenario 2

In Scenario 2, 52 participants elaborated on their answer, from which five thematic areas emerged (Table 4). The most often discussed theme was the **Process** of the decision-making, appearing

Table 3: Themes emerged in Scenario 1.

Theme	Description	#
Missing Factors	Not considering all the appropriate factors	23
Similar cases	Comparison with similar cases, data used to train the model	17
Process	Procedures followed by the model; features' weights	15
Specific information	Specific value of a factor missing from the given scenario	9
Human/Company policy	Deferring to humans, following company's policy	3
Other	[falls outside of the established themes]	7

Table 4: Themes emerged in Scenario 2.

Theme	Description	#
Process	Procedures followed by the model; features' weights	23
Factors	Consideration of irrelevant factors and/or missing important factors	15
Age	Consideration of Age in the decision	13
Gender	Consideration of Gender in the decision	12
Other	[falls outside of the established themes]	11

in almost half (23) of the responses. Similar to Scenario 1, most of the responses wondered about *"what makes certain features less preferable than others"* (p49). Other responses commented on specific elements, such as *"age should factor more into the algorithm"* (p64), even though the scenario description did not disclose how much each factor influenced the decision.

The second most common theme (with 15 responses) consisted of responses discussing the **Factors**. The vast majority of the responses asked about and offered other *"important factors"* (p46) that the system should consider in this context, such as health condition (p29, p92), reason of their flight (p28, p46, p66, p97), and disability status (p37, p78). Some participants argued that the factors mentioned in the prompt were *"irrelevant to the scenario"* (p77).

A number of responses specifically mentioned **Age** (13) and **Gender** (12) in their responses, with 8 responses mentioning both. While some participants disputed only the use of age (p58) and gender (p91) in such systems, some argued that neither should be used to make such decisions (p32, p56). Some referred to the law, specifying that the use of factors such as gender and age is *"illegal"* (p38) and *"breaks lots of (UK) laws"* (p62).

Interestingly, one participant (p56) discussed a personal experience similar to that of the scenario, and argued that the decision should be based on *"the time the checkin was made."* [*sic*]. 11 responses fell under the catch-all **Other** category as they did not mention any of the other themes.

5.3 Scenario 3

The three cases in Scenario 3 were analysed together to compare the effect of the different outcomes on participants' perceptions. In addition to five main themes that emerged, the responses in Cases B and C were also coded for whether the participant made references to their response to an earlier case (see Table 5). Case A had 56, Case B had 46, and Case C had 52 responses that were analyzed.

In Case A, the majority of the participants (23 out of 56) asked about **Specific Information** missing from the description of the given scenario; however, only 9 participants (out of 46) in Case B and 14 (out of 52) in Case C discussed this theme. In Cases A and C, most of the participants noted that they wanted to know the loan repayment rates of the individuals and how they differed (e.g. p54, Case A [p54/A]; p67/C). Interestingly, a few participants also wanted to know the specific loan repayment rate in Case B (where the rate for one applicant was explicitly stated and the other implied via ratios). The remaining responses for Cases A and C mainly focused on the race and gender of the applicants, while only one participant mentioned them for Case B (p17).

Process was the second most common theme in Case A (19/56), and the most common theme discussed by the participants in Case C (20/52), but was mentioned in only 10 responses (out of 46) for Case B. These responses often noted that there was *"no information on the decision process"* (p68/A). Interestingly, for Case B, participants mentioned the proportional outcome as an indication of the calculation/reasoning of the algorithm; in contrast, with the other cases, the outcome was a reason to question the process leading to the decision. Some participants wondered about the influence of the different factors on the final decision, one remarking that *"Yes [I had sufficient information] but as long as the parameters are awful [the system] is biased"* (p20/B). Other participants specifically asked about the role of gender and race; in fact, many of the participants discussing Process also discussed Race/Gender (7/19 in Case A, 5/10 in Case B, 6/20 in Case C).

Race/Gender was the most common theme discussed in Case B (20/46), and a popular theme in Case A (17/56) and Case C (14/52). While some responses simply questioned the role of race/gender in the decision-making process, others argued that race and gender were not relevant to the decision (p69/B) and should not be taken into account (p71/C). Certain participants specifically said that the use of these features were *"illegal"* (p38/C).

Less often, participants made references to other, missing **Factors** to be considered by the system (8 in Case A, 11 in Case, 9 in Case C). Among the factors mentioned were the applicants' ability (p13/A), their job stability (p21/A), annual income (p46/A) or financial situation (p7/C), and the risks of the business they proposed (p6/C). One participant argued that the factors are *"not*

Table 5: Themes emerged in Scenario 3 (Case A, Case B and Case C).

Theme	Description	A	B	C
Specific information	Specific value of a factor missing from the given scenario	23	9	14
Process	Procedures followed by the model; features' weights	19	10	20
Race/Gender	Consideration of race and/or gender in the decision	17	20	14
Factors	Consideration of irrelevant factors and/or missing important factors	15	6	10
Other	[falls outside of the established themes]	8	11	9
Same as above	Same answer as the previous case(s)	–	15	13

sufficient" and that a human is needed to *"analyze the business proposal"* (p70/A).

Overall, 28 responses (8 in Case A, 11 in Case B, 9 in Case C) fell under the catch-all **Other** category, which includes responses that do fall under any of the other themes as well as responses where the participant indicated they *"don't really understand all the questions"* (p50/C).

6 DISCUSSION

We investigated the relationship between six constructs related to fairness and justice [5, 10] in algorithmic decision-making: Agreement with the decision, Understanding of the decision-making process, Appropriateness of factors considered, Fairness of the decision-making process, whether the individual Deserved the outcome, and Trust in the system's decision over a human's [32]. Although we expected all constructs to correlate with one another, we found that specifically, understanding of the process was highly correlated with factors used appropriately, as well as with that the individual(s) in the scenario deserved the outcome.

Similar to previous work [19, 32, 41] we found that **factors** are context- and output-dependent, something that is also obvious in our qualitative analysis. Surprisingly, in the scenarios describing a car insurance premium (Scenario 1) and an airline rerouting (Scenario 2) decisions, participants tended to both agree with the decision, and believe that the person in the scenario did not deserve the outcome, given their circumstances or behavior. Considering that participants often noted that some **important factors** were missing, this could imply that the participants find that while the calculations with the given information are accurate (agreeing with the outcome), the process needs to take into account other factors, and therefore does not actually calculate the most just decision (deserved outcome). Participants generally find that some important factors were not considered, highlighting the complexity of real life situations and the *context-dependent nature of algorithms*.

In both scenarios, participants' responses focused heavily on the **factors** that were involved in the decision making as well as the actual **decision-making process** followed. However, in Scenario 1 statistical evidences show that although the participants found the factors used in the decision-making process as 'appropriate', they did not believe that the decision-making process was 'fair'. In contrast, in Scenario 2, participants indicated in their majority that the factors used in the decision-making process were 'not appropriate' and hence the process was 'not fair'. Qualitative results confirm that the participants found the use of some factors – specifically age and gender – inappropriate in Scenario 2, so they were reluctant to believe the process was fair. Our findings confirmed our expectations and are in line with [14, 23] that people who believed the decision-making process was not fair would also not trust the system's decision more than a human's decision.

Looking closer at the way different outcomes can affect the perception of fairness and justice in algorithmic decision-making systems, our results for Scenario 3 showed that dividing resources proportionally (based on a factor considered relevant) was perceived as more fair than dividing the resources equally, which was still more fair than giving all resources to one individual over another. Our finding aligned with [41] where the 'ratio' decision was found to be more fair than the 'equal' decision, supporting thus Liu et al. calibrated fairness [34] instead of the treating similar people in a similar way approach [16].

The (lack of) information/explanations provided in the scenarios prompted participants to comment on those heavily. Sometimes participants did not necessarily think other factors were required, but instead needed **specific information** from a factor already mentioned in the scenario. For example, nearly half of the participants asked about the specific gender, race, and/or loan repayment rate of the individuals in Scenario 3, Case A. Fewer people, but still a notable amount asked for similar information in the other cases of Scenario 3 as well as in Scenario 1. Scenario 2, in contrast, disclosed the specific age, gender, and other details about the individuals in the scenario; accordingly, participants did not ask for further specific information about the individuals, but instead focused on the generalized, more abstract discussion of the use of gender and age as factors. This implies that participants can judge the process, the decision, and their fairness better if the specific information (e.g. gender) of the individuals involved are disclosed. Therefore, for developers to enable full judgement of an algorithmic process, it may not be enough to give information about the process in the abstract, but provide concrete details about the cases involved [4]. Another very common theme discussed was the **process or reasoning** of the algorithm, appearing often in every scenario. Participants often asked about the weight of the factors/features in general, or stated that one factor (often gender, race, or age) should/should not have more influence than the others. It seems that our sample, hence developers, are in favour of a decompositional approach to explanations which requires more specific information about the system, the process, the weights etc., rather than pedagogical approaches that might be more suitable for end users [17].

As was expected we have found differences related to the participants' level of academic education. Postgraduate students appeared to understand the decision-making process in Scenario 1 more than undergraduate students did, but questioned whether 'sufficient information' was provided. Postgraduates in their majority

50% replied that they were not sure whether the information provided for this scenario was sufficient, while undergraduates in their majority 46,9% replied positively. This shows the experience that postgraduates students have over undergraduates especially considering their comments: *"My answer is no,in my opinion, information should also include the health condition of certain drivers which were involved in the research"* (p7, postgraduate) or *"No, they lack psychosocial characteristics, such as driving attitude (eg: I like speed, when I can not respect the signals, etc) or the ability to react to stress (eg: if I had a bad day or I'm late change my driving style). Finally, at the level of behavior monitoring, driving routines could be established on a weekly basis, considering both the driving environment (eg motorways vs busy areas), and the timing."* (p28, postgraduate). Clearly, postgraduate students understood the system well enough to be able to challenge the factors, values/weights and the model overall.

In the scenario for different loan distributions, Case B, where the loan repayment rate for the two individuals was disclosed was thought to have 'sufficient information' by postgraduates; in contrast, undergraduates tended to think more information should be provided. Similar to Scenario 1, this could be a result of the postgraduate students being more familiar with the type of decision-making systems using proportionality. Three more differences were observed between these groups in the case where all the loan amount was given to one individual. Postgraduates tended to (1) agree with the decision of the system, (2) find the factors used appropriate, and (3) find the decision process fair, more so than the undergraduates.

Clearly, education has a great role to play in affecting the development of fair algorithmic decision-making systems. According also to previous work [24] there is a need for incorporating seminars, modules and training courses in the computing related degrees as well as professional training courses for recently graduated practitioners. Pierson et al. [38] reported evidences of statistically significant changes in perception and attitudes of students towards algorithmic fairness and transparency after just an hour of lecture and discussion. Thus, in order for future algorithmic decision-making systems to be fair, we need to ensure that the people developing them are aware of concepts related to Fairness, Accountability, Transparency and Ethics in algorithmic systems. They also need to be aware that the systems they are developing have an impact (positive or negative) to the society.

Limitations. It is important to note that this study, as any empirical study, faced several limitations. The numbers in the thematic analysis indicate only whether a theme was mentioned or not within a response. Themes were mentioned in many ways, from information that should have been included in the scenario, to concepts (positively or negatively) affecting "fairness" in algorithmic systems, to personal subjective opinions.

In contrast to previous work [38], the participants' gender, and previous training/self-reported knowledge on algorithmic fairness did not appear to affect the individuals' perception of the constructs we were examining. This can be due to the limited number of participants in our study and should be taken into account when one interprets this result. Also, due to the initial goals of the survey design, our data does not include race information about the participants and therefore the dimension of race is missing from our analysis of the participants' perceptions.

Finally, the number of the participants (N=99) allowed us to run some quantitative analysis over the collected data; however, the reader should take into account that this number is relatively small, the subjects were students in degrees with varying distance from algorithmic development, and they were from a limited set of countries, so our findings may not be representative of the general public.

7 CONCLUDING REMARKS

Algorithmic decision-making systems are becoming very popular, prompting us to rely more and more on their decisions, with potentially serious consequences for the affected social groups. Developers have an important role to play when they are called to develop algorithms that will drive these decisions. Algorithmic fairness might be a first step in understanding how people perceive and assess the decisions and the explanations provided. Most importantly, we need to understand how developers perceive fairness in the systems they develop, which will potentially decide on behalf of a human, and in some occasions for matters with real social impact.

This paper provides some insights on how future developers perceive algorithmic fairness in algorithmic decision-making. It suggests that their level of academic education has a role to play in their understanding of the decision-making process, as well as their critical thinking on the factors and the decision-making process involved. Factors that are employed are context- and output-dependent, and appropriate factors might not presuppose the fairness of the decision-making process. Future developers in our sample were in favour of a 'ratio' decision rather than the others provided. We hope that this work will act as a starting point for understanding the concept of fairness from the developer's perspective instead of the user/person affected, in order to inform policies, procedures and guidelines for the respective industry.

8 ACKNOWLEDGMENTS

This project is partially funded by the European Union's Horizon 2020 research and innovation programme under grant agreements No. 739578 (RISE) and 810105 (CyCAT).

REFERENCES

[1] Antoine Allen. 2016. The 'three black teenagers' search shows it is society, not Google, that is racist. *The Guardian* 10 (2016).

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. ProPublica. Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica, 23 Mai 2016.

[3] Paul Baker and Amanda Potts. 2013. 'Why do white people have thin lips?'Google and the perpetuation of stereotypes via auto-complete search forms. *Critical discourse studies* 10, 2 (2013), 187–204.

[4] Reuben Binns. 2020. On the Apparent Conflict between Individual and Group Fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 514–524. https://doi.org/10.1145/3351095.3372864

[5] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. "It's Reducing a Human Being to a Percentage": Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173951

[6] Nigel Bosch, Sidney K. D'Mello, Ryan S. Baker, Jaclyn Ocumpaugh, Valerie Shute, Matthew Ventura, Lubin Wang, and Weinan Zhao. 2016. Detecting Student Emotions in Computer-Enabled Classrooms. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (New York, New York, USA) *(IJCAI'16)*. AAAI Press, 4125–4129.

[7] Anna Brown, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. 2019. Toward Algorithmic Accountability in Public Services: A Qualitative Study of Affected Community Perspectives on Algorithmic Decision-Making in Child Welfare Services. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300271

[8] Irene Y. Chen, Fredrik D. Johansson, and David Sontag. 2018. Why is My Classifier Discriminatory?. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (Montréal, Canada) *(NIPS'18)*. Curran Associates Inc., Red Hook, NY, USA, 3543–3554.

[9] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.

[10] Jason A Colquitt and Jessica B Rodell. 2015. Measuring justice and fairness. (2015).

[11] Bo Cowgill. 2018. Bias and productivity in humans and algorithms: Theory and evidence from resume screening. *Columbia Business School, Columbia University* 29 (2018).

[12] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. *San Fransico, CA: Reuters. Retrieved on October* 9 (2018), 2018.

[13] Nicholas Diakopoulos. 2016. Accountability in algorithmic decision making. *Commun. ACM* 59, 2 (2016), 56–62.

[14] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.

[15] Paul Dourish. 1997. Accounting for system behaviour: Representation, reflection and resourceful action. *Computers and design in context* (1997), 145–170.

[16] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (Cambridge, Massachusetts) *(ITCS '12)*. Association for Computing Machinery, New York, NY, USA, 214–226. https://doi.org/10.1145/2090236.2090255

[17] 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 L. & Tech. Rev.* 16 (2017), 18.

[18] Motahhare Eslami, Kristen Vaccaro, Min Kyung Lee, Amit Elazari Bar On, Eric Gilbert, and Karrie Karahalios. 2019. User Attitudes towards Algorithmic Opacity and Transparency in Online Reviewing Platforms. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300724

[19] Ben Green and Lily Hu. 2018. The myth in the methodology: Towards a recontextualization of fairness in machine learning. In *Proceedings of the machine learning: the debates workshop*. Stockholm, Sweden.

[20] Nina Grgić-Hlača, Elissa M. Redmiles, Krishna P. Gummadi, and Adrian Weller. 2018. Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction. In *Proceedings of the 2018 World Wide Web Conference* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 903–912. https://doi.org/10.1145/3178876.3186138

[21] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.* 51, 5, Article 93 (Aug. 2018), 42 pages. https://doi.org/10.1145/3236009

[22] Leif Hancox-Li. 2020. Robustness in Machine Learning Explanations: Does It Matter?. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 640–647. https://doi.org/10.1145/3351095.3372836

[23] Galen Harrison, Julia Hanson, Christine Jacinto, Julio Ramirez, and Blase Ur. 2020. An Empirical Study on the Perceived Fairness of Realistic, Imperfect Machine Learning Models. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 392–402. https://doi.org/10.1145/3351095.3372831

[24] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. 2019. Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3290605.3300830

[25] Nicolas Huck. 2019. Large data sets and machine learning: Applications to statistical arbitrage. *European Journal of Operational Research* 278, 1 (2019), 330–342.

[26] Ben Hutchinson and Margaret Mitchell. 2019. 50 Years of Test (Un)Fairness: Lessons for Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 49–58. https://doi.org/10.1145/3287560.3287600

[27] Fei Jiang, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, Yilong Wang, Qiang Dong, Haipeng Shen, and Yongjun Wang. 2017. Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology* 2, 4 (2017), 230–243.

[28] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human decisions and machine predictions. *The quarterly journal of economics* 133, 1 (2018), 237–293.

[29] Kyriakos Kyriakou, Styliani Kleanthous, Jahna Otterbacher, and George A Papadopoulos. 2020. Emotion-based Stereotypes in Image Analysis Services. In *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization*. 252–259.

[30] Preethi Lahoti, Krishna P Gummadi, and Gerhard Weikum. 2019. ifair: Learning individually fair data representations for algorithmic decision making. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. IEEE, 1334–1345.

[31] Preethi Lahoti, Krishna P Gummadi, and Gerhard Weikum. 2019. Operationalizing individual fairness with pairwise fair representations. *Proceedings of the VLDB Endowment* 13, 4 (2019), 506–518.

[32] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1 (2018), 2053951718756684. https://doi.org/10.1177/2053951718756684 arXiv:https://doi.org/10.1177/2053951718756684

[33] Min Kyung Lee and Su Baykal. 2017. Algorithmic Mediation in Group Decisions: Fairness Perceptions of Algorithmically Mediated vs. Discussion-Based Social Division. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 1035–1048. https://doi.org/10.1145/2998181.2998230

[34] Yang Liu, Goran Radanovic, Christos Dimitrakakis, Debmalya Mandal, and David C Parkes. 2017. Calibrated fairness in bandits. *arXiv preprint arXiv:1707.01875* (2017).

[35] Frank Marcinkowski, Kimon Kieslich, Christopher Starke, and Marco Lünich. 2020. Implications of AI (Un-)Fairness in Higher Education Admissions: The Effects of Perceived AI (Un-)Fairness on Exit, Voice and Organizational Reputation. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 122–130. https://doi.org/10.1145/3351095.3372867

[36] Dena F Mujtaba and Nihar R Mahapatra. 2019. Ethical considerations in ai-based recruitment. In *2019 IEEE International Symposium on Technology and Society (ISTAS)*. IEEE, 1–7.

[37] Jahna Otterbacher, Jo Bates, and Paul Clough. 2017. Competent Men and Warm Women: Gender Stereotypes and Backlash in Image Search Results. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 6620–6631. https://doi.org/10.1145/3025453.3025727

[38] Emma Pierson. 2017. Demographics and discussion influence views on algorithmic fairness. arXiv:1712.09124 [cs.CY]

[39] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as Mechanisms for Supporting Algorithmic Transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173677

[40] Emilee Rader and Rebecca Gray. 2015. Understanding User Beliefs About Algorithmic Curation in the Facebook News Feed. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 173–182. https://doi.org/10.1145/2702123.2702174

[41] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C. Parkes, and Yang Liu. 2019. How Do Fairness Definitions Fare? Examining Public Attitudes Towards Algorithmic Definitions of Fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (Honolulu, HI, USA) *(AIES '19)*. Association for Computing Machinery, New York, NY, USA, 99–106. https://doi.org/10.1145/3306618.3314248

[42] Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019. Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Anchorage, AK, USA) *(KDD '19)*. Association for Computing Machinery, New York, NY, USA, 2459–2468. https://doi.org/10.1145/3292500.3330664

[43] David R Thomas. 2006. A general inductive approach for analyzing qualitative evaluation data. *American journal of evaluation* 27, 2 (2006), 237–246.

[44] Kjerstin Thorson, Kelley Cotter, Mel Medeiros, and Chankyung Pak. 2019. Algorithmic inference, political interest, and exposure to news and politics on Facebook. *Information, Communication & Society* 0, 0 (2019), 1–18. https://doi.org/10.1080/1369118X.2019.1642934 arXiv:https://doi.org/10.1080/1369118X.2019.1642934

[45] Thi Ngoc Trang Tran, Müslüm Atas, Alexander Felfernig, Viet Man Le, Ralph Samer, and Martin Stettinger. 2019. Towards Social Choice-Based Explanations in Group Recommender Systems. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization* (Larnaca, Cyprus) *(UMAP '19)*.

Association for Computing Machinery, New York, NY, USA, 13–21. https://doi.org/10.1145/3320435.3320437

[46] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3174014

[47] AJ Wang. 2018. Procedural Justice and Risk-Assessment Algorithms. *Available at SSRN 3170136* (2018).

[48] Ruotong Wang, F. Maxwell Harper, and Haiyi Zhu. 2020. Factors Influencing Perceived Fairness in Algorithmic Decision-Making: Algorithm Outcomes, Development Procedures, and Individual Differences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376813

[49] Allison Woodruff, Sarah E. Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A Qualitative Exploration of Perceptions of Algorithmic Fairness. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3174230

[50] Richard Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. 2013. Learning Fair Representations. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28* (Atlanta, GA, USA) *(ICML'13)*. JMLR.org, III–325–III–333.