Article

Patterns

Perception of fairness in algorithmic decisions: Future developers' perspective

Highlights

- Appropriate factors used in the decision-making does not ensure perceived fairness
- Trust to the system is strongly affected by the user's perception of fairness
- Algorithmic fairness is defined as the use of objective factors
- Sensitive attributes can be the most likely cause of unfairness

Authors

Styliani Kleanthous, Maria Kasinidou, Pınar Barlas, Jahna Otterbacher

Correspondence

styliani.kleanthous@ouc.ac.cy

In brief

The importance of this work lays primarily in that, while others have looked into how the end users and/or the general public perceive elements of fairness, accountability, rransparency, and ethics, it is important to understand how the people who are involved in the development of algorithmic decision-making systems perceive the above concepts. We approach this through a study with an international sample of students coming from fields adjacent to computing in order to examine their perceptions on the above topics.

Kleanthous et al., 2022, Patterns *3*, 100380
January 14, 2022 © 2021 The Author(s).
https://doi.org/10.1016/j.patter.2021.100380

CellPress

Patterns

CellPress
OPEN ACCESS

Article

Perception of fairness in algorithmic decisions: Future developers' perspective

Styliani Kleanthous,[1,2,3,*] Maria Kasinidou,[1] Pınar Barlas,[2] and Jahna Otterbacher[1,2]
[1]Cyprus Center for Algorithmic Transparency, Open University of Cyprus, Faculty of Pure & Applied Sciences, 33 Yiannou Kranidioti Avenue, 2220 Latsia, Nicosia, Cyprus
[2]Transparency in Algorithms Group, CYENS Centre of Excellence, 23 Dimarchias Square, 1016 Nicosia, Cyprus
[3]Lead contact
*Correspondence: styliani.kleanthous@ouc.ac.cy
https://doi.org/10.1016/j.patter.2021.100380

THE BIGGER PICTURE Fairness, accountability, transparency, and ethics (FATE) in algorithmic systems is gaining a lot of attention lately. With the continuous advancement of machine learning and artificial intelligence, research and tech companies are coming across incidents where algorithmic systems are making non-objective decisions that may reproduce and/or amplify social stereotypes and inequalities. There is a great effort by the research community on developing frameworks of fairness and algorithmic models to alleviate biases; however, we first need to understand how people perceive the complex construct of algorithmic fairness. In this work, we investigate how young and future developers perceive these concepts. Our results can inform future research on (1) understanding perceptions of algorithmic FATE, (2) highlighting the needs for systematic training and education on FATE, and (3) raising awareness among young developers on the potential impact that the systems they are developing have in society.

SUMMARY

In this work, we investigate how students in fields adjacent to algorithms development perceive fairness, accountability, transparency, and ethics in 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 using scenarios, in addition to defining algorithmic fairness and providing their view on possible causes of unfairness, transparency approaches, and accountability. The findings indicate that "agreeing" with a decision does not mean that the person "deserves the outcome," perceiving the factors used in the decision-making as "appropriate" does not make the decision of the system "fair," and perceiving a system's decision as "not fair" is affecting the participants' "trust" in the system. Furthermore, fairness is most likely to be defined as the use of "objective factors," and participants identify the use of "sensitive attributes" as the most likely cause of unfairness.

INTRODUCTION

Algorithmic systems are increasingly gaining an important role in decision-making, deciding on the posts and news we will see on social media[1,2] contributing to regulated areas, such as health,[3] prison releases[4] and job hiring,[5] even moderating education[6] and many more. The use of algorithmic decision-making has prospects to make decision-making more efficient and reliable.[7,8] It is no longer a worry, but rather a reality, that algorithms are making non-"objective" decisions that may reproduce and/or amplify social stereotypes and inequalities.[9] This type of behavior can be witnessed in many domains: gender discrimination has been detected in resume search engines;[10] auto-complete search terms can produce suggested terms that could be viewed as racist, sexist, or homophobic[11]; image search results are gender-biased depending on the search term used[12] and racially biased toward Black individuals.[13,14] Recent literature has given us a lot of examples to demonstrate that algorithmic (un)fairness is a real, complex construct[15,16] that affects the way we live our lives.

There is a great effort by the emerging research community on developing frameworks of fairness[17,18] and algorithmic models to alleviate biases;[19–21] however, due to the complexity of fairness as a concept, there is a need to understand how algorithmic fairness[22–28] and interlinked concepts, such as transparency[29–32] and accountability,[33,34] are perceived by people.

While related work has looked into how the end users and/or the general public perceive elements of fairness, accountability,

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

transparency, and ethics (FATE), it is important to understand how the people who are developing—or will soon be involved in developing—algorithmic decision-making systems perceive the above concepts. To our knowledge, this is a gap in the field, as we have not been able to find any studies that internationally sample students from fields adjacent to computing in order to examine their perceptions on the above topics.

To explore how future developers perceive FATE, 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.[25,35] Two of the three scenarios are context independent of each other and were used to trigger participants' judgment 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.

In addition, we ask participants: *to define algorithmic fairness*, what they see as *possible causes of (un)fairness*, what they would do in order to *make a system more transparent* to its users, and who would be *held accountable* if the system behaved unfairly to some parts of the population. The last three questions are explored in the context of a hypothetical recruitment system that the participants are asked to imagine they would help develop.

The findings of the first part 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 with 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 the appropriateness of certain factors used by the algorithm. Qualitative analysis shows that future developers, in order to judge the fairness of a given algorithmic system, will 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.

The qualitative part of our study shows that most of the young developers defined algorithmic fairness as the selection of *objective/appropriate factors*, while a notable portion defined fairness as the quality of the *outcome/decision* and the opposite of *biases/discrimination*. The use of *irrelevant and/or sensitive attributes*, such as gender, age, and race, identified as the most likely cause of unfairness. Regarding ways for increased transparency, the participants overwhelmingly suggested adding *explanations of the process and output*. Finally, our findings show that young developers understand that, in the case where a system behaves unfairly, the team who developed the system should be held accountable. Overall, our findings note the complexity of understanding perceptions of FATE in algorithmic decision-making even from the developer's perspective.

Fairness in algorithmic decision-making

Fairness is a complex construct, which is perceived differently by different actors and people understand it differently and according to the context where the system is operating.[36] Fazelpour and Lipton[16] follow a political philosophy lens to stress the complexity of building fairness in machine learning (ML) systems that have to operate in the complex sociotechnical environment. They emphasize the need for a human-centered perspective when formulating mitigation strategies in algorithmic fairness. Similarly, Leben[37] highlights the difficulty in building fairness in ML systems for all protected groups in society and the potential for discrimination in some cases. Looking at the psychology literature, fairness is usually studied relative to justice,[35] where at times the terms are used interchangeably. The diversity of people and contexts, where fairness and justice are studied, call for algorithmic fairness and its impact in society, to be investigated as to how these concepts are perceived by humans. Thus, usually scenario-based studies are employed in order to provide study participants a framing when asked to define fairness.[22,25,28,38]

Perceived fairness has been investigated by many, with usually contradicting results to be reported. In a work-management scenario, the public perceive algorithmic decision-making as less fair than human decision-making even when the decision requires "human skills."[25] In a similar manner, in the context of admission in higher education, university students found algorithmic decision-making more fair than human decision-making.[33] To be considered as fair, algorithms and systems need to consider social and altruistic behavior,[38] elements that may be difficult to incorporate in mathematical modeling.[10] 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.[39] Accuracy is considered more important than equality, while demographic parity (demographic parity seeks to equalize the percentage of people who are predicted to be positive across different groups) 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,[40] 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.[25,41]

The continuous exposure of people to algorithmic fairness makes them more aware of the potential biases in the decision as well as in the data or the algorithm interaction.[42] They seek more information about how different factors weighted in the decision and whether an algorithm uses sensitive attributes (such as race or gender). Different variables—such as the education level and favorable outcome, as well as development procedures of the system—have also been proven to affect the perception of algorithmic fairness.[27] In particular, people rate the algorithm as more fair when the decision is in their favor, irrespective of whether it appears to be biased toward 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.[26]

Table 1. Descriptive statistics for the variables used in the analysis

	Mean	SD
Agreement	2.7232	0.62316
Understanding	3.4798	0.87531
Appropriateness	2.6040	0.75267
Fair	2.6242	0.69062
Deserved	2.5838	0.62966
Trust	2.4788	0.79581

Education and training play an important role for the society in general in understanding algorithmic fairness. Students' perception of fairness changed after an hour-long lecture and discussion on algorithmic fairness.[26] Education level appears to be an important predictor for comprehension of fairness, and negative sentiment is associated with greater comprehension of demographic parity.[43] However, in order for algorithms to become more fair, developers must be educated and aware of the possible biases and discrimination that might occur as a result of the algorithms they develop. They must be able to identify the source of bias and take the proper steps to overcome it. In their review of 50 years of work on (un)fairness, Hutchinson and Mitchell[34] propose 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 encourage ML researchers to dig deeper into topics, define context- and use-dependent criteria, and question whether all subgroup dimensions can be served by a single model or if a different approach might be required.

Transparency and explanations in algorithmic decision-making

Transparency of algorithmic systems refers to methods for interpreting a complex model in a way that the user would understand, while it can act as a mechanism that helps accountability.[44] The opacity[29] of the algorithmic systems called for different transparency approaches to be investigated by the community, as a method that might enhance the perceptions of fairness.[45] With ML models being exploited for predicting sensitive individual information, classifying individuals into categories, and providing decisions that were previously taken by humans,[46] there is a need for interpreting those models in a way that the user would understand. Due to the complexity of the mathematical models that are employed in algorithmic systems, several methods have been proposed, such as excluding the use of sensitive attributes in some algorithms[44] and providing the user with visual[32] or verbal[31] explanations.

In light of the recently drafted General Data Protection Regulation, users of systems that involve automated decision-making have a right to obtain "meaningful explanations of the logic involved." Hence, methods and approaches that will allow users to understand the output of these opaque and automated systems in context are needed.[47] Meanwhile, users should be able to understand the possible consequences of applying the decision in the real world.[46] Therefore, explanations should be informative and easy to be interpreted by the users. Edwards and Veale[48] 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 tackle the contextual challenges early on.[34]

Explaining a system's decision, though, is not simple. Different levels of explanation are required depending on context and the audience, particularly with black box models.[48] While 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 users understand how their input could change the output of the system by resembling everyday human conversation. Studies with users, however, are ambiguous as to the type and level of explanation they prefer. Binns et al.[22] conducted a study using four different explanation styles (input influence, sensitivity, case-based, demographics). Their findings suggested that different explanations styles provide different justice perception. More specifically, they found that 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 with sensitivity-based explanations—explaining how much the value of a variable used in the model affects the output. However, they did not observe these when people were exposed to the same explanation style in different scenarios. Rader and Gray[1] found that explanations, in any form, help to raise awareness of how the system works and recognize 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,[31] emphasizing how fairness is subjective to each individual person.

On various occasions, the difficulty of dealing with and explaining potentially harmful outcomes has been highlighted. Take as an example the Google photos incident, where a Black American and his friend were mistakenly labeled by the system as "gorillas." Google worked 2 years to "solve" the problem and the final solution was just a work-around of removing the label from their lexicon. This shows the difficulties that companies, such as Google, and by extension 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 FATE for developers who are already in the business.

RESULTS

Quantitative results

Statistical 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).

Similar to previous work,[22,35] correlations were found between all constructs (see Table 1). Although we were expecting that all

Table 2. Pearson correlations for the six constructs of justice

		Agreement	Understanding	Appropriateness	Fair	Deserved	Trust
Agreement	Pearson correlation	1	.000	.000	.000	.000	.000
	sig. (two-tailed)		1	.401**	.365**	.319	.222**
Understanding	Pearson correlation	.413**		.000	.000	.001	.027
	sig. (two-tailed)	0.000	0.401**	1	0.678**	0.691**	0.535
Appropriateness	Pearson correlation	0.682**	0.000		0.000	0.000	0.000
	sig. (two-tailed)	0.000	0.365**	0.678**	1**	0.792**	0.703**
Fair	Pearson correlation	0.765**	0.000	0.000		0.000	0.000
	sig. (two-tailed)	0.000	0.319**	0.691**	0.792**	1**	0.685**
Deserved	Pearson correlation	0.795**	0.001	0.000	0.000		0.000
	sig. (two-tailed)	0.000	0.222*	0.535**	0.703**	0.685*	1**
Trust	Pearson correlation	0.574**	0.027	0.000	0.000	0.000	
	sig. (two-tailed)	0.000	0.000	0.000	0.000	0.000	0.000

constructs will correlate, according to the literature on perceived fairness and justice,[35] we were surprised to see that understanding of the process followed correlates with appropriateness of the factors, and understanding of the process correlates with deserved outcome (see Table 2).

Perception of fairness constructs

To examine a number of hypotheses regarding participants' perception of the fairness constructs in scenarios 1 and 2, we ran a series of Wilcoxon signed ranked tests. Surprisingly we found significant statistical differences in the participants' opinions regarding whether people who agreed with the decision also believe that the person in the scenario deserved the outcome (scenario 1: $z = 2.70$, $p = 0.007$; scenario 2: $z = 4.043$, $p < 0.001$). In both scenarios there was a considerable number of participants who selected the more positive options on the agreement scale, while also selecting the negative options on the deserved scale, indicating that they agreed with the decision but the person in the scenario did not deserve the outcome.[49] The results show also significant differences between the responses of the participants in scenario 1 on whether people who found the factors used in the decision-making process appropriate will also think that the decision-making process is fair ($z = -3.193$, $p < 0.001$), with participants in their majority agreeing that the factors used in the decision-making process were appropriate; however, they do not believe that the decision-making process was fair. In scenario 2 there was no statistically significant difference between the two scales.[49] For both scenario 1 and scenario 2 we did not get any significant differences between the people who indicated that the decision-making process was not fair and trust to the system's decision. It was also interesting to examine whether the different decisions in scenario 3 (cases A, B, and C) affected participants' perception of the constructs of agreement, understanding, appropriateness, fair process, deserved outcome, and trust. A within-subject analysis using ANOVA repeated measures followed by a Bonferroni post-hoc test was run. 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 the decision in case B (proportional outcome) perceived as the most just, while the decision on case C as the least. Comparing the participants' 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.[49]

Table 3. Differences between undergraduate and postgraduate students

	U	Z	P	M_U	M_P
Scenario 1: Understanding	1331	2.07	0.038	3	3
Scenario 3, case B: Sufficient information	764	−2.48	0.013	0.5	1
Scenario 3, case C: Agreement	813.5	−2.043	0.041	1	2
Scenario 3, case C: Appropriateness	795.5	−2.185	0.029	1	2
Scenario 3, case C: Fair	813	−2.06	0.039	1	2

Differences between undergraduate and postgraduate participants

Differences were also found between undergraduate and postgraduate participants. A series of Mann-Whitney U tests (see Table 3) were run to determine if there were differences between the two groups on their understanding of the process by which the decision was made; whether they believe the information provided was sufficient; whether they agreed with the decision; whether the factors used for making the decision were appropriate; and whether the decision was fair. Scenario 1 was the only scenario where statistically significant differences in understanding were found, indicating that postgraduates understood the process that the system is following in making a decision better compared with undergraduates. In scenario 3, case B, we found that undergraduates found that the information provided was less sufficient in this case compared with postgraduates, with statistical significance. In regard to the agreement with the decision, in case A and case B we did not find any statistically significant differences between the two groups. For case C,

Table 4. Themes emerged in scenario 1

Theme	Description	No.
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 5. Themes emerged in scenario 2

Theme	Description	No.
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

undergraduates appeared to agree less with the decision of the system compared with the postgraduates. In examining whether there is a difference in the perception of appropriateness of the factors considered for the system's decision in case C, undergraduates considered the factors used in the system for making the decision less appropriate compared with 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, with undergraduates considering the decision-making process less fair compared with the postgraduates. More detailed results can be found at Kasinidou et al.[49].

Qualitative results

Was the information provided in the scenarios sufficient?

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 analyze participants' free text responses we used thematic analysis, as described in the methodology section below.

Scenario 1: 59 participants elaborated on their responses to Q1, where 6 thematic areas emerged from their responses (Table 4). 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). A large portion of the responses argued that the algorithm should take into account more information about the accident (p64, p91) while others wondered "if there were any other elements which influenced the decision" (p79). Interestingly, some participants even mentioned the need to consider other factors even when they indicated they found the information sufficient. A few participants, despite the fact that they agreed that the information provided was sufficient, noted that "there could be other factors that are potentially more important" (p71). Out of the 23 responses discussing missing factors: 4 also discussed the similar cases, referring to a need for more information about the similarities with the other cases (p61); 2 discussed the process/weight of factors with respect to the "the inner working of the algorithmic decision-making system" (p69) and 1 also discussed the human/company policy.

Seventeen 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). Even when they agreed that the information provided was sufficient, three participants still noted the "need to specify that the decision was based on only 1 similar case" (p20). 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). The main idea, shared by the vast majority of the responses, can be summarized by participant 80: "Would be interesting to see how much each factor contributed to the decision." Some specifically asked, "I would like to see why Claire is a better driver than Sarah. Is it because of the higher percentage of miles and night drives?" (p37) or 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 be able to talk to the driver and better understand the driver's attitude. Seven 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).

Scenario 2: 52 participants elaborated on their answer, from which 5 thematic areas emerged (Table 5). The *process* of the decision-making, appearing in almost half (23) of the responses, making it the most often discussed theme. 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.

Table 6. 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

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]. Eleven responses fell under the catch-all *other* category as they did not mention any of the other themes.

Scenario 3: the three cases were analyzed 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 6). 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 was "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 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).

Defining fairness

In their attempt to provide a definition of fairness (Q4), participants referred to different concepts. Overall, 12 thematic areas emerged from the analysis of the participants' responses (see Table 7). It comes with no surprise that Objective Factors/Conditions is the most frequently used theme (42 of 99) in the definitions. Characteristically, participant 38 defined fairness as: "A computed decision that only factors features of merit and does not merely find a solution that is optimal but balances it with one which is egalitarian."

While some participants argued that the algorithm should "take into account all or most elements" (p9), others specified that [the algorithm] "should only use relevant factors" (p85) with respect to the goal or the system. Out of the 42 responses discussing Objective factors (Table 8): 13 responses also discussed the Outcome/Decision referring to the objectivity of the system's decision,[27] 11 also discussed the context[25,41] with respect to the factors that should be considered according to the system or the environment; and 9 also discussed Biases/Discrimination as a criterion for fairness[20] (see also below). Two themes each received 24 responses, tying for the second

Table 7. Themes emerged from defining fairness question: Name, description, and frequency

Category	Description	No.
Objective factors	the objectivity and/or appropriateness of factors	42
Decision/outcome	the quality of the decision or outcome	24
Biases/discrimination	producing outcomes with/without social biases or discrimination	24
Context	(not) taking into account the different situations/scenarios of deployment	21
Emotional/moral/ethical/norms	(not) considering ethics, emotions, morality, and/or social norms	18
Demographic characteristics	(not) using sensitive attributes (e.g., gender, race, age)	15
Training data	the impact of the dataset/information used to train the algorithm	13
Methods/rules	appropriate feature weights and/or procedures	13
Explainability/transparency	the algorithm/output is explainable, transparent, and/or understandable	12
Human intervention	the (positive or negative) impact of humans on the system/outputs	7
Disadvantaged groups	(not) considering impact on minorities/ disadvantaged groups	4
Other	[falls outside of the established themes]	9

most common theme: Biases/Discrimination and Outcome/Decision.

Participants often defined algorithmic fairness in opposition to what is unfair (i.e., Biases and/or Discrimination); "An algorithm that is robust to biases" (p33), "It should exclude discrimination factors" (p14). Another participant referred directly to social biases: "In particular fair algorithms do not reproduce, magnify or introduce social biases" (p62). One participant called out the role of the developers of the algorithm: "programmed by biased person/group of people based on opinionated criteria defined by them" (p36). One participant took the definition of fairness and directly applied it to algorithms: "Algorithms being fair? Idk. 'Impartial and just treatment or behavior without favoritism or discrimination' … by an algorithm" (p49).

Often, participants directly referred to the quality of the Outcome/Decision of the algorithm: that "fair" algorithms should "make appropriate suggestions" (p77), pick the "most suitable result" (p8), or offer a "fair solution" (p89). Some argued specifically that there should be "consistent" (p51) or "reproducible" (p70) outputs. While it seemed implied by many, one participant defined algorithmic fairness in relation to the "[l]ife impacting decisions" (p36) made by algorithms. Just over half (13) of the 24 responses discussing the *outcome/decision* also mentioned the *objective factors*. The responses seemed to imply that the "correct" factors would lead to the "best" outcomes, sometimes saying it explicitly: "Should only use relevant factors. Should have the highest chance of making the best decision," (p85).

Table 8. Co-occurrences in defining fairness question

	DG HI TD	DC	D/O	M/R	B/D	E/T	C	E/M/E/N	Other
Objective factors	1 4 3	7	13	3	9	4	11	3	0
Disadvantaged groups (DG)	0 0	1	0	0	0	0	2	1	0
Human intervention (HI)	2	0	1	0	3	0	1	0	1
Training data (TD)		2	2	1	3	2	1	2	0
Demographic characteristics (DC)			2	0	9	4	6	2	0
Decision/outcome (D/O)				4	5	3	6	5	0
Methods/rules (M/R)					0	1	2	6	0
Biases/discrimination (B/D)						4	4	1	1
Explainability/transparency (E/T)							4	3	0
Context (C)								8	0
Emotional/moral/ethical/norms (E/M/E/N)									0

Acronyms on top correspond to the categories in Table 7.

The third most common theme was Context with a total of 21 responses. The main idea, shared by the vast majority of the responses, can be summarized by participant 73: "Algorithmic fairness is not an absolute idea, but it is relative. [The] same algorithm may not be fair in every social, cultural, and political context." Others discussed the issue in more specific terms, such as factors ("for example, an algorithm that measures healthiness can take sex into account, but an algorithm to measure credit score shouldn't", p40) or "the community the algorithm will be used [in]" (p37). Some responses went down to the person-level, arguing that "[the algorithm] should treat individuals based on their personal merits/context/situation" (p60).

The most common co-occurrence with the Context theme was objective factors, with 11 responses discussing both themes. Participants usually made references to factors that were "relevant and appropriate for the task" (p39), sometimes specifying that this consideration was "based on the scenario [the algorithm is] used for" (p77). Some participants elaborated on this connection, also bringing in the role of Demographic characteristics: "Fairness depends on the domain and specific problem solved. An algorithm is certainly not fair if it bases its decisions on properties unrelated to the actual situation (e.g., gender, race, sexual orientation, religion, in many applications)" (p78).

From the remaining, Emotional/Moral/Ethical/Norms is a notable theme with a total of 18 responses. Often co-occurring with Context (8), this theme includes responses that argue for defining or judging algorithmic fairness in relation to certain pre-established concepts or rules. A common sentiment was that a fair algorithm would make "the most […] ethical choice" (p31), "align itself with […] social norms and values" (p73), and/or "reflect the moral rules of a certain society" (p61). A few participants mentioned finding solutions that are "egalitarian" (p38) or "that could be defended in court" (p67).

A number of responses explicitly mentioned Demographic Characteristics. Some discussed specific identity markers (most common being gender and race), while others were more vague, referring to algorithms using "facts that are not protected characteristics" (p42) or someone "belonging to [a] specific social group" (p60). One participant referred to these characteristics as "Factors that are not in people's control—age, gender, race, sexuality, etc" (p79). Perhaps unsurprisingly, 9 of the 15 discussing Demographic Characteristics also mentioned Biases/Discrimination; and, of course, the common thread was that the bias or discrimination would be "based on, e.g., demographic factors" (p22), or that the algorithm would "reproduce, magnify or introduce social biases" (p62).

Participants discussed the information used in the algorithm, whether as *training data* or as inputs. Responses ranged from discussing simply "numerous data" (p6) to having "good information with a clear basis" (p50). The former group sometimes referred to "hav[ing] enough information to judge" (p16), or to "keeping in mind all the available data" (p90) or similarly, that the algorithm has been "pre-trained on a correct dataset with a large number of cases" (p92). One participant gave a thoughtful response referring to proxies in the dataset reflecting structural biases toward certain social groups[50]: "Primarily it is important to me that nobody is discriminated based on characteristics that are not subject to human action (like gender, race, nationality, ethnicity). However, the problem is that, due to past discrimination, action-related characteristics also allow conclusions about non-action-related characteristics." (p69).

Participants sometimes referred to the Methods/Rules of the algorithm, arguing for giving factors/variables "the appropriate importance" (p46). 6 of the 13 responses discussing Methods/Rules of the model also discussed Emotional/Moral/Ethical/norms. The responses ranged from discussing "ethical" decision-making (p91) based on "the rules that are fixed and accepted by the public" (p19) to arguing for selections "without emotional and moral judgment" (p3).

Some responses defined algorithmic fairness in relation to the *explainability/transparency* of the decision and/or process. The definitions varied quite often, discussing the need for "[a]lgorithmic decisions [to be] made in an explainable, transparent way" (p39). While some participants wanted this transparency to "make the process clearly understandable and contestable to all the stakeholders" (p73), other participants wanted the transparency to "mak[e] it as clear as possible to the user how decisions are made" (p63), including the role/weight of factors especially in the case of contested factors such as gender/race (p68). For other participants, the decisions themselves needed to be "[j]ustified with an explanation that is understandable and rational" (p58), some elaborating that this "allow [s the decision] to be judged in light of fairness" (p22). Only one participant discussed concrete methodology to achieve algorithmic fairness through "the use of a general argumentative framework that combines the knowledge base (a set of rules and input data) and preference reasoning" (p91).

The remaining themes received few responses. The *human intervention* theme appeared only in seven responses; some responses cited unwanted human intervention as a reason to employ a fair algorithm (p2), while other responses referred to human intervention in the creation of algorithms as a source of unfairness (p36), as well as other ideas. The least common theme, *disadvantaged groups*, received only four responses. Nine responses fell under the catch-all *other* category, which includes thoughtful responses that do not mention the other themes (such as "in my opinion we don't have to concentrate whether an algorithm is fair but whether an algorithm is unfair", sic, p34).

Consideration of fairness

When asked whether they would consider fairness in their system (Q5), most of the participants (66.7%) responded affirmatively (4–5), 19.2% indicated that they would not consider fairness (1–2), and 14.1% seemed undecided (3). In line with our instructions—to answer the following question only if the participant selected 3–5—only 81 out of the 99 participants answered the question about choosing a part of the system to focus on (Q6), in order to promote fairness. As a reminder, the participants were also asked to explain their choice using free text (Q7).

The majority of the respondents (37 out of 81; 46%) chose to focus on the Training data. Some participants justified their answer on that the Training data are "the most important factor in the algorithm fairness" (p2), "essential in order to create and understand scenarios algorithms" (p55), and "more useful and reliable, as all we do in next parts all depend on the data we use" (p21). Furthermore, participants discussed that "[a]ny

Table 9. Themes emerged from sources of unfairness question: Name, description, and frequency

Category	Description	No.
Sensitive attributes	the use of irrelevant and/or demographic factors, such as gender, race, age	60
System/model	the procedures followed by the model; features weights	29
Dataset	the dataset/information used for training the model or as input	19
Human influence	the impact of humans on the system (such as social biases)	8
Other	[falls outside of the established themes]	15

algorithm merely perpetuates the bias of its creators and the data it is fed" (p76) and that "[b]iased data will lead to problems" (p62). "Machine learning algorithms learn based on the training data" (p40), if the training data are biased then the algorithm will be biased (p40, p42, p68). Others discussed that they would collect more training data (p22, p93) and they would remove sensitive attributes (p39, p40, p69) to ensure an unbiased dataset. One particular participant who chose to focus on Data also mentioned "modeling the algorithm also Needs to be done in a fair way, and there Needs to be an Explanation on why one applicant was rejected (Output)" (p68).

Twenty-six participants (32%) mentioned Modeling Algorithm. Some participants explained that they chose the Modeling Algorithm since they can control input and data through the algorithm (p28, p41) by "set[ting] which factors are more important than others" (p46) and generally can "decide how to interpret and act upon [the data]" (p51). Others explained that "[r]emoving the data might not remove the unwanted bias" (p58); but "biases can be removed (or at least reduced) from input and training data via pre-processing or careful selection of rules" (p78). Some participants discussed that Modeling Algorithm is the most important part of the process to ensure unbiased algorithms (p37, p57), "[it] gives the greatest flexibility to tune toward fairness" (p60) since data cannot be changed (p54, p92).

Thirteen participants (16%) selected the Input that the user provides to the system as the part of the process on which they would focus. The main idea behind their choice is "by restricting the input categories to only those which should be considered (i.e., not considering the "protected categories"—gender, age, ethnic background), we ensure at least some basic fairness" (p67). Some others discussed that they would focus on both the Input data and the Modeling Algorithm (p91) by "disregard[ing] discriminating factors, such as age and race" (p56). On the other hand, some participants discussed the need for more information to make the algorithm fair (p13, p15, p82).

Only 5 out of the 81 respondents (6%) said that they would focus on the Output, suggesting that the "[r]esults are the most important part of the data model" (p3) and that they "would focus on explaining the output to the user" (p63), "so that a human can identify potential biases/short sighted learning behavior" (p80).

Table 10. Co-occurences in cause of unfairness question

	DS	SM	HI	Other
Sensitive attributes	9	14	3	3
Dataset (DS)		10	4	0
System/model (SM)			5	0
Human influence (HI)				0

Causes of unfairness

When asked to ponder potential causes of "unfairness" in the hiring decision-making system described (Q9), the majority of participants discussed one or more of four themes (see Table 9). The most often discussed theme was Sensitive Attributes, appearing in 60 responses. These responses made references to specific characteristics of job applicants—such as gender, ethnic background, and age—and often framed the unfairness in terms of discrimination based on these factors: e.g., "If decisions were based on gender, age, ethnic background" (p58). A large portion of the responses explicitly discussed characteristics—with gender, ethnic background, and age appearing much more often than other characteristics in the provided scenario—but at times, participants simply made references to "protected" (p52) or "unrelevant" (p2) factors. A few participants discussed factors that were relevant, such as skills and competency, and stated any factors falling outside of determining those would "cause unfairness in output" (p27).

While some participants believed unfairness came from choosing to consider these factors at all ("Gender and Ethnic background shouldn't be considered by the algorithm," p40), others referenced the impact of the factors on the process or the final decision ("high weight on ethnic background, age, or gender," p57). The latter group makes up the 14 co-occurrences of Sensitive Attributes and System/Model as seen in Table 10. A smaller group of responses (9) included a reference to the dataset, often bringing up the biases embedded in historical data50; for example, participant 98 pointed out: "When using historical data to build a model, the model could discriminate based on gender or ethnic background. This is because there was and still is to some extent, unfair representation of people of all genders and racial backgrounds on most sectors, especially the IT sector."

Twenty-nine responses discussed the System or Model itself as a cause of unfairness. While a lot of responses discussed the lack of/inappropriate weight on factors as described earlier, a few responses blamed "[t]he design of the system" (p14) or simply proposed [most likely a lack of] "reliability" (p83) as a possible cause for unfairness.

Out of the 19 responses mentioning some part of the Dataset, those that did not discuss the Sensitive Attributes or System/Model usually discussed having too little data (p12) or specifically having "[n]ot enough training data" (p24). Others made references

Table 11. Themes emerged from transparency strategies question: Name, description, and frequency

Category	Description	No.
Explanation of the algorithm	explaining the process followed by the system; how the factors were used/ weighed	53
Explanation of the output	explaining the output to the user; why a specific decision was made	25
Training data	using the training or output datasets to offer transparency	4
Auditing the algorithm	analyzing outputs or model	3
Documentation	a system report/document available to the public	3
Do not know	participant does not know what to do to make the system more transparent	14
Nothing	participant would do nothing to make the system more transparent	9
Other	[falls outside of the established themes]	7

to "The data collected, or the training data used for training the model" (p90) without specifying what may be causing unfairness within this data. One participant interestingly made a reference to potentially malicious Human intervention in the Dataset, as well as the information security of the System/Model: "Some users have provided false information, or the system has vulnerabilities that have not been detected before" (p15).

Seven responses simultaneously discussed Sensitive attributes, Dataset, and System/Model, usually framing unfairness through inappropriate weighing of sensitive attributes found in the Input data (e.g., p43). Two responses discussed all four themes, mentioning that the weight of these factors (p38)—or how the outputs of the models are perceived/used (p69)—is the responsibility of people creating the algorithm.

Transparency strategies

In the responses discussing ways to make the system more transparent (Q8), 8 themes emerged (see Table 11). Participants often discussed *explanations of the algorithm/model* (53 responses) and *explanations of the output* (25 responses), with some responses (13, see Table 12) in the overlap of these themes. Many responses made a vague reference to "add[ing] and explanation of the process" (sic, p26) or "explain[ing] the main steps" (p84) in order to make the system more transparent. However, quite a few participants specified their hypothetical strategy, such as making explicit the factors and their weights within the model ("displaying the different factors of the decision and how much weight they had", p63), applying explainability measures (e.g., "counterfactuals," p54; "PCA"/principal-component analysis, p38; "decision trees," p90; "radar graph for each user," p43), or allowing the decision-making process to be contestable/malleable (p73). These explanation strategies, especially when elaborated upon, often overlapped with explanations of the Output.

Table 12. Co-occurrences in transparency question (only other/ nothing co-occur once, from those not shown)

	EA	TD	Auditing the algorithm
Explanation output	13	3	1
Explanation algorithm (EA)		2	2
Training data (TD)			0

The responses that suggested *explaining the output*—which did not discuss explaining the algorithm/model as described above—often discussed showing previous inputs and outputs similar to the case in question (p95) or disclosing the full set of outputs such that each output can be judged in relation to others ranked similarly (e.g., p75, p17).

Four participants suggested using the *training data* to make the system more transparent. They discussed the "validation of training data" (p65) and explaining the output based on its similarity to cases in the training data (p75, p76, p92). Only three responses discussed *auditing the algorithm/model*, either suggesting a "dedicated algorithm for data tracking" (p15), an analysis of the outputs (p21), or explanatory audit strategies ("The process leading to the output should be displayed; the means of doing this would depend on the algorithm (ex. a decision tree, the list of relevant propositions, etc.)," p51). In a similar manner, three responses suggested *documentation* for added transparency, establishing "[m]ore effective communication" (p16) or "system report [that] can be posted online" (p4).

A few participants thought about the strategy for explaining the system/model more fundamentally, suggesting choosing a more explainable algorithm to begin with: "with the current state of the art I would avoid the more black box ML approaches like neural networks and favor models with a clear structure, like decision trees and Bayesian networks" (p60). However, the participants did not always seem confident that it would be an efficient system or that there were other paths to transparency ("I would try choosing an algorithm that is easier to interpret while still having adequate performance. Not sure what I would do if that didn't work," p40). Some participants, while inclined to make the system more transparent, seemed disillusioned with the "unexplainable" nature of many ML systems: "I would, but I know that this is a difficult thing to do. Not even computer scientists are able to fully understand their ML systems and the resulting decision-making process. So how can other people?" (p69).

Similarly, seven participants said they *didn't know* how to make the system more transparent, while nine participants explicitly said they *wouldn't do anything*. Out of the nine saying no, three elaborated on their answer: one stated the user "just need[s] to know the outcome" (p18), one was worried that "if [the system] was transparent, there must be someone try to cheat" (sic, p5), and another refused the system's use in this context completely ("No because people don't trust algorithm and they shouldn't. The manager should rely on his experience not the algorithm," p66).

Accountability

The last part of the study examined the concept of accountability and how the participants perceive it. Most of the participants (75.7%) agreed with (4–5 on the Likert scale) the statement that "their team" would be held accountable, compared with 38.4% who agreed that "the system" would be held accountable and 6.1% who agreed that "neither the system or my team" would be held accountable (Q9).

In the free text explanations of their choices (Q10), participants remarked that "the team creates the system so that the team should be held accountable" (p1) and that "since the system was built by me and my team, as part of the team I should take fully responsibility on this occasion" (p7). In a similar tone participant 46 noted "The system just makes it much quicker and simplified but the code behind it is written by the team developing the algorithm, the system provides the output by using the steps that is programmed to follow, no blame on the system."

Some participants justified their view that the team would be held accountable with the fact that the system is not autonomous, and instead a human chooses the attributes/factors that the system uses to make decisions. Some participants wrote thoughtful (relatively) lengthy answers that were very clear on this point: "You can't blame a system. It works how it's designed. Not sure if it's my team's fault though, it depends on whether the higher-ups enforced what attributes to use. Why rank gender and ethnicity if not because you plan to discriminate?" (p49); "The system would have made the decisions. However, the factors were only considered as the system was programmed to consider them." (p70); "It makes no sense to hold the system accountable for the decisions. At the end of the day, humans programmed the system, whether it was developed as a black box using historical data or not, the humans should be held accountable by not properly tuning the parameters of it to be less discriminatory." (p98).

An important aspect indicated by participants is the "black box" nature of decision-making systems and how this makes it difficult to decide who is to be held accountable in the case of unfairness. Characteristically, one participant wrote: "Depends on how closely people can look at the system. If it were just speculation based on the end result without looking at the system directly it may be hard to judge (hence less likely to be held accountable)." (p48). A participant also commented on the suitability of ML how these are employed for solving complex problems "Machine learning algorithms are a mostly bad solution for very complex problems." (p66). Finally, participant 98 (above) also touches the opacity of these algorithmic systems but still believes the team is responsible for any unfair outputs.

Other participants felt that both the team and the system should take the responsibility mainly since there must be a company behind the system: "My team developed the system—hence it is our responsibility. Considering AI, the system is clearly accountable as well, but possibly due to design, data or other developer issues." (p24). Another participant pointed out that "Ultimately it is the system, which is at fault; however institutionalizing blame leads to deflection (i.e., try suing a corporation versus suing a person), and my team would therefore have to be held most accountable for what has happened." (p39). On the other hand, some participants sharing this opinion felt that the humans should be accountable for the unfairness of the system. "It is our responsibility to create fair systems." (p71). "Both the system and the team is at fault, since the "Fairness" was coded by the team." (p89).

DISCUSSION

In this study, we seek to better understand future developers' perception of topics related to algorithmic fairness, transparency, and accountability. To do that we provided participants with a questionnaire where their knowledge and previous training, with respect to topics related to algorithmic fairness, were recorded. We investigated the relationship between six constructs related to fairness and justice[22,35] 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.[25] In addition, participants were asked to define algorithmic fairness in their own understanding and given a scenario of an algorithmic system, to report what could be possible causes of unfairness, whether they would consider methods for algorithmic transparency, and their responsibility toward accountability, in case the system behaves unfairly to some parts of the population.

Level of education, knowledge, and training on fairness, accountability, and transparency

The results emphasized that soon-to-graduate and postgraduate students with degrees in ML, Computer Science, Data Science, etc., lack the necessary knowledge in topics related to fairness in algorithmic systems. Our finding that students coming from different parts of the world lack knowledge on these topics is one of the most important findings in this work and adds on previous research,[24] reflecting the need for incorporating seminars, modules, and training courses in computing-related degrees. Educating current and future developers is an important step in the process of developing more fair AI systems.

This may be partly due to the lack of formal training in topics related to FATE, which the majority of our participants never received. Previous work[26] reported evidence of statistically significant changes in perception and attitudes of students toward algorithmic fairness and transparency just after an hour of lecture and discussion. However, it must be noted also that the proportion of participants who claim to have very little knowledge is lower than the proportion who have never received formal training; some participants may have found information on FATE elsewhere or overestimated their knowledge in the area out of personal interest.

As was expected we found differences related to the participants' level of academic education. Postgraduate students appeared to understand the decision-making process and state that they were not sure whether the information provided for this scenario was sufficient, compared with undergraduates. This shows the experience that postgraduate students have over undergraduates especially as was reflected in the free text replies.[49] Postgraduate students appear to have understood the system well enough to be able to challenge the factors, values/weights, and the model overall. 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. Pierson[26] reported evidence of statistically significant changes in perception and attitudes of students toward 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 FATE in algorithmic systems. They also need to be aware that the systems they are developing have an impact (positive or negative) to the society.

Perception of fairness constructs

Aligned with previous work[25,40,41] we found that factors are context and output dependent, something that is also obvious in our qualitative analysis. Results showed that, in the case that inappropriate (scenario 2) factors were involved in the decision-making process—specifically age and gender—participants were reluctant to believe the process was fair. This finding confirmed our expectations and are in line with other work,[39,51] 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 another report,[40] where the "ratio" decision was found to be more fair than the "equal" decision, supporting thus Liu et al.'s calibrated fairness[52] instead of the "treating similar people in a similar way" approach.[18]

Regarding the (lack of) information/explanations provided in the scenarios, participants indicated the need for more concrete examples (e.g., how gender, race, and/or loan repayment rate of the individual is taken into consideration in the decision-making process) so they can judge the process, the decision, and their fairness better. Therefore, for developers to enable full judgment 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.[15]

Algorithmic fairness

Overall, we can see that algorithmic fairness is subjective for each individual. Previous work[15] emphasizes that algorithmic fairness is not a construct that is easily defined. Many students felt that algorithmic fairness had to do with the factors within the model; how they were selected or related to the objective of the system. Objectivity, objective factors, and "egalitarianism" as it is referred by a number of the respondents, aims at leveling the opportunities for people in the society rather than providing equal share to everything. This is a concept that has been discussed in previously[15,16,53] with respect to algorithmic fairness.

One sentiment kept emerging, even if participants used different combinations of themes to state it: algorithmic fairness is often defined in opposition to discrimination based on demographics, which can be seen in the (use of) outputs by algorithmic systems. This can be observed quantitatively from how often the Biases/Discrimination, Outcome/Decision, and Demographic characteristics themes (co-)occurred. Participants in this study are more likely to be familiar with reports on media about algorithmic systems behaving unfairly to certain populations, such as the COMPAS recidivism risk calculation algorithm,[53] Google Photos' "gorilla" tag mistakenly applied on photos of Black people,[15] and the gender- and race-biased gender inferences in facial analysis algorithms (reported in Gender Shades[54] and covered by the documentary *Coded Bias*),[22] which increase awareness of algorithmic discrimination based on demographics. In fact, participant 34, whose response could not be clustered under the themes we determined, was explicit in this choice of framing: "in my opinion we don't have to concentrate whether an algorithm is fair but whether an algorithm is unfair."

Context appeared to be usually associated with fairness, implying that algorithmic fairness is not a metric that is detached from the real situation it is deployed in. There has been a discussion by scholars on the importance and challenges of representing context when designing the algorithmic model[25,41] that will drive a system. Our participants, intentionally or unintentionally, have also identified this relationship.

We were not surprised to see that some respondents emphasized the role of Human intervention in the process, emphasizing the role of the developer and the possibility of their individual biases transferred into the algorithmic model through rules and/or the training data. Practitioners who were interviewed in previous work[24] also emphasized the need for introducing diversity in development teams toward mitigating unfairness in ML systems, as well as the need for approaches that will steer more diverse training datasets, in order to minimize biases passed into the system and amplified.[11,55] Consistent with previous work,[24,34] respondents selected the Training data and the Modeling algorithm as the most important components they would focus on for making the system more fair. They emphasized that potential biases and discrimination will be learnt by the system if such exist in the training data, and the potential impact they or the team developing a system can have on eliminating the biases.

Interestingly, much fewer participants made references to Emotional/Moral/Ethical/Norms than we expected, with only one response relating to law. Moral, Ethical, and Emotional considerations in the respondents' definitions were rightfully co-occurring with context. Work in philosophy[56,57] is exploring the relationship between the concepts of this theme and algorithmic decision-making however, context as well as cultural norms are

dimensions that cannot be disconnected from the discussion. Even within these responses, the attitude toward how norms should be treated was not consistent: some participants felt that the system's Method/Rules should respect and align with established norms, while others felt that fairness can be achieved when we have separated the algorithmic process from those norms.

Causes of unfairness

When describing the potential causes of "unfairness" based on the scenario of the hiring decision-making system, a notable portion made references to the role of attributes relating to demographics and other "sensitive" characteristics in the system's decision. This is consistent with their responses on defining fairness where the majority made references to objective factors and those that have been focused on for analyses in the most prominent research in the field of fairness, accountability, and transparency [of algorithms].[54,58,59] Maybe the respondents, being students, were more familiar with those through the academic and scientific literature. Consistent with their definitions on fairness, participants discussed the (input and/or training) datasets as potentially a main cause of unfairness in the system described, implying or explicitly discussing having more or "better" data as a solution to unfairness. This may reflect the belief that ML models, as based on mathematical methods, are neutral and that biases only enter the system through datasets which reflect social phenomena.

Transparency

Explaining the algorithm/model was the most preferred strategy for Transparency, although some responses opted (also) for explaining the output/decision. Students showed awareness of specific methods for explanations, often offering transparency with respect to the feature selection and weights. As seen from their definitions of fairness with respect to—and theories on causes of unfairness in—algorithmic systems, features/factors are considered a critical part of making the process and its results "fair." The participants offering explanations about the *output only* often suggested disclosing cases that were similar to the user's case, whether it be from the training dataset or the full set of outputs. This may indicate awareness of the complexity of ML systems and a desire to make the explanations more understandable to the non-expert users.

Very few responses discussed audits of the model as a measure of transparency. Audits are often used in the research community to examine the system's treatment of data from different groups of people.[54,58,59] Audits were an important part of the discussion with practitioners in previous work,[24] where respondents were requesting for methods and procedures for performing system-level audits in a more systematic way as an approach toward fairness. This shows that experience acquired by practitioners allowed them to appreciate other methods of transparency, compared with the students in our sample, who lack knowledge, training, and experience in the topic. This may point to a heightened awareness of unfairness with a lack of knowledge regarding solutions.

Finally, the majority of the participating students agreed that their team should be held *accountable*. This is promising as it may be an indication that the future generation of developers understands their responsibility of delivering "fair-behaving algorithms" to their users—and the possible consequences in case the system they develop behaves unfairly to some parts of the population. This is also evidence toward a direction of holding humans more accountable for algorithmic unfairness, rather than blaming AI and algorithms alone.[60] Furthermore, they believed that, whatever the system produces, it does so because the team chose to program it that way or the training data carry certain societal stereotypes. This line of thinking often mused on the retaliation measures available in the case of unfairness, such as legal action. However, interestingly, participants often remarked that when the system is opaque, it may be difficult to determine who is accountable for the unfairness in the case of unfairness. Therefore, there may be a link between the level of transparency and the accountability of a system, which warrants further research.

Limitations

Limitations of our study should be noted when interpreting this work. First, the participants of this study were students in computing-related degrees; hence, their responses should not be generalized to the general population. Second, since we did not have equal representation of the participant's program and country of study, we did not run any comparative statistical analysis based on these parameters. At this stage, our focus was to explore their perceptions and understanding rather than quantitatively compare their responses. Future work should try to gather a larger number of participants in a similar study and understand possibly also the cultural differences that might exist between countries, race, gender, etc. Finally, while interpreting our analyses, it must be noted that the frequencies only indicate *whether a certain theme was mentioned or not*; we did not analyze how often the theme was discussed as a positive or negative influence on algorithmic fairness. The most common threads in the responses are not the only kind; what one participant considers a crucial part of fairness may be the same thing another participant considers a *barrier* to fairness. Future work needs to examine algorithmic fairness from this lens.

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. Most importantly, we need to understand how developers perceive FATE in the systems they develop, which will potentially decide on behalf of a human, and on some occasions for matters with real social impact.

This paper provides some insights on how future developers perceive algorithmic FATE 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. Despite the calls for increased education on "ethical" development of technology, the majority of our participants have not taken any training or courses on the topic, and therefore do not consider themselves knowledgeable about algorithmic fairness. We find there is a need for systematic education on algorithmic fairness.

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 favor 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.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Styliani Kleanthous (styliani.kleanthous@ouc.ac.cy).

Materials availability

This work did not generate any new material other than the results presented in this manuscript.

Data and code availability

This work did not generate any data and/or code that could be made available to others.

Methodology

To understand how future developers, perceive FATE in algorithmic decision-making, we conducted an online survey that ran between September 2019 and May 2020. Participants were asked to fill in an online questionnaire that had two parts embedded. The first part focused on examining the participants' perception in terms of different constructs related to "justice" as described in Colquitt and Rodell.[35] The second part focusing on the participants' understanding of topics related to FATE in the context of a specific scenario.

Participants were presented with three scenarios where algorithms made decisions that influenced humans. Scenarios were adopted from previous work and the context of the scenarios describe cases that most adults have experience with, or they are aware of.[34,40] Two of the three scenarios were used to trigger the participants' judgment 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 behavior. This scenario was adapted from Binns et al.[22]

Scenario 2: passengers on overbooked airline flights being automatically selected for rerouting: 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 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, and was a 55-year-old male, instead of Lisa, who was single, traveling alone, and a 35-year-old, female.

Scenario 3: applying for a personal financial loan. This scenario was adapted from Saxena et al.[40]

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 $50,000.

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.

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.

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 Colquitt and Rodell[35] 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 behavior.
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," or "unsure." Subsequently, participants self-reported (yes, no, or other write-in) whether they had 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 five-point Likert scale (from 1 [not at all] to 5 [very knowledgeable]), providing also free text explanations on their training and knowledge.

In addition, we collected qualitative responses on the participants' understanding of topics related to fairness in the context of a different scenario. Firstly, participants were asked to respond to the following question in free text: (Q4) "How would you define algorithmic fairness?" Then participants were presented with the following scenario:

You work as a part of a team developing a system to filter and rank CVs for the hiring manager at a company, to help them shortlist the best candidates. The system will rank applicants based on the following attributes: Gender, Age, Ethnic Background, Work Experience, Education, Skills, Knowledge, Competencies, Personality Traits.

Participants were asked, given the above scenario, (Q5) whether they would "consider dimensions of fairness in [their] system" (5-point Likert scale: from 1 [strongly disagree] to 5 [strongly agree]) and if they answered with a 3 or higher, to indicate (Q6) "on which part of the process [they] would focus" from the following options: (1) Input (the input that the user provides to the system), (2) Output, (3) Modeling algorithm, or (4) Data (used for training the algorithm/for learning). The next question asked the participant to (Q7) "explain why [they] chose to focus on that part of the development process" using free text.

The survey continued with two more free text questions, (Q8) "Would you do anything to make the system more transparent so the user (manager) will understand how the system took a decision? How?" and (Q9) "In your opinion what would be a possible cause of unfairness in the above system?" In the subsequent section, we focused on questions of accountability, preceding the questions with "If this system behaves unfairly to some parts of the population": (with the subtitle "For example the system might discriminate over people of certain ethnicity or discriminate between male and female"). Then, we presented the participants with three statements and asked them to indicate (Q10) whether they agree with each on a Likert scale (1 [strongly disagree] to 5 [strongly agree]) that: my team would be held accountable; the system would be held accountable; neither the system nor my team would be held accountable. The final question of the survey asked the participants to (Q11) "Please explain [their] answers" to the above statements with free text.

Thematic analysis

Qualitative research includes a range of analytical methods applicable in various contexts such as content analysis,[61,62] participatory action research,[63] and systematic analysis.[64] Thematic analysis has become a widely used tool for analyzing qualitative data[65] and report patterns and/or classifying data into thematic categories that are essential to a better description of a phenomenon.[66] Participants' free text responses in Q1, Q4, Q7, and Q8 were coded and thematically analyzed[67]: three researchers independently read the data thoroughly, attentively, and analytically in order to identify significant elements

in the responses. Then, they used the information identified as relevant in the reading phase to generate initial codes by grouping elements of data according to similarities or perceived patterns. Then, the categories identified by the three researchers were compared, the disagreements discussed, and sometimes a dimension's definition amended to come to a final consensus. Once the final categories were agreed, researchers compared the responses they identified for each category and discussed any disagreements. We allowed multiple categories per answer and calculated the co-occurrence of themes in responses in an attempt to capture the interplay of different themes in participants' perceptions. Descriptive statistics were used to understand the questions with Likert scale or otherwise multiple-choice answers.

Participants

We recruited respondents using snowball sampling. We emailed the survey to colleagues at 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 has received ethical clearance by the Cyprus National Bioethics Committee.

A total of 60.6% of our respondents were male, with 47.5% in the age group of 18–24 years, 35.4% between 25 and 32 years, 10.1% between 33 and 40 years, and 7.1% above 40 years. 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%), with 7% in the US, 4% in Israel, and 3% in China, Brazil, and Australia.

Understanding the participants' responses required to first understand their previous experience with (Q2) and knowledge in topics related to algorithmic fairness (Q3); 22.2% of our participants had taken some kind of training on the above topics, while the majority (77.8%) had not. Of those who had taken some kind of training, most had either worked in a related project ("I worked in the project XY[4] for one year and a half," participant 61, p61); took a short training ("My previous employer was trying hard to build fair products, and thus most of the employees went through a short training," p62); or read about these topics out of interest ("I only read some articles on explainable AI and the importance thereof," p22). Others had a more formal, course-related experience ("I have been studying the ethics in computing for last 5 years," p73; "Took a course in Algorithmic Transparency," p20). We did not expect many participants with formal training or academic courses, as the area is relatively new.

Interestingly, when asked how knowledgeable they were in these topics, fewer participants (59.6%) selected the lower end of the scale (1–2, not knowledgeable) than the proportion of participants that did not get any training. In a similar vein, while approximately 22.2% of our participants have been trained on these topics, only 14.1% selected the upper end of the scale (4–5, very knowledgeable). Despite the fact that the participants are students at undergraduate or postgraduate level, and some are PhD students, only one reported being very knowledgeable (5). Even though participants indicated they were trained they might feel that they are not knowledgeable enough in these topics or they might be underestimating of their knowledge and understanding on the topics. Nevertheless, there is a limitation in the training and education that developers are getting on algorithmic FATE and this was emphasized also in previous work.[24]

ACKNOWLEDGMENTS

This project is partially funded by the Cyprus Research and Innovation Foundation under grant EXCELLENCE/0918/0086 (DESCANT) and by the European Union's Horizon 2020 Research and Innovation Programme under agreement nos. 739578 (RISE) and 810105 (CyCAT).

AUTHOR CONTRIBUTIONS

J.O. and S.K. defined the initial conceptualization of the study and designed the methodology. S.K. was responsible for supervising the process from beginning to end and contributing with M.K. to the data collection. P.B., M.K., and S.K. took part in the data analysis, applying qualitative and quantitative formal methods. All authors contributed to the interpretation of the results and to the article write up.

DECLARATION OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

We worked to ensure gender balance in the recruitment of human subjects. We worked to ensure ethnic or other types of diversity in the recruitment of human subjects. We worked to ensure that the study questionnaires were prepared in an inclusive way.

Received: June 22, 2021
Revised: September 13, 2021
Accepted: October 5, 2021
Published: November 3, 2021

REFERENCES

1. Rader, E., and Gray, R. (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), pp. 173–182. https://doi.org/10.1145/2702123.2702174.
2. Thorson, K., Cotter, K., Medeiros, M., and Pak, C. (2019). Algorithmic inference, political interest, and exposure to news and politics on Facebook. Inf. Commun. Soc. 1–18. https://doi.org/10.1080/1369118X.2019.1642934.
3. McCradden, M., Mazwi, M., Joshi, S., and Anderson, J.A. (2020). When your only tool is a hammer: ethical limitations of algorithmic fairness solutions in healthcare machine learning. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society ((AIES '20). Association for Computing Machinery), p. 109. https://doi.org/10.1145/3375627.3375824.
4. Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). ProPublica. Machine bias: there's software used across the country to predict future criminals. And it's biased against blacks. ProPublic. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. (Accessed 25 October 2021).
5. Mujtaba, D.F., and Mahapatra, N.R. (2019). Ethical considerations in AI-based recruitment. In 2019 IEEE International Symposium on Technology and Society (ISTAS) (IEEE), pp. 1–7.
6. Bosch, N., D'Mello, S.K., Baker, R.S., Ocumpaugh, J., Shute, V., Ventura, M., Wang, L., and Zhao, W. (2016). Detecting student emotions in computer-enabled classrooms. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence ((IJCAI'16). AAAI Press), pp. 4125–4129.
7. Cowgill, B. (2018). Bias and Productivity in Humans and Algorithms: Theory and Evidence from Resume Screening, *29* (Columbia Business School, Columbia University).
8. Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2018). Human decisions and machine predictions. Q. J. Econ. *133*, 237–293.
9. Bozdag, E. (2013). Bias in algorithmic filtering and personalization. Ethics Inf. Technology *15*, 209–227.
10. Chen, I.Y., Johansson, F.D., and Sontag, D. (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.), pp. 3543–3554.

11. Baker, P., and Potts, A. (2013). 'Why do white people have thin lips?' Google and the perpetuation of stereotypes via auto-complete search forms. Crit. Discourse Stud. *10.2*, 187–204.
12. Otterbacher, J., Bates, J., and Clough, P. (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), pp. 6620–6631. https://doi.org/10.1145/3025453.3025727.
13. Antoine, A. (2016). The 'three black teenagers' search shows it is society, not Google, that is racist. Guardian *10*. https://www.theguardian.com/technology/2016/jun/09/three-black-teenagers-anger-as-google-image-search-shows-police-mugshots. (Accessed 25 October 2021).
14. Kyriakou, K., Kleanthous, S., Otterbacher, J., and Papadopoulos, G.A. (2020). Emotion-based stereotypes in image analysis services. In Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization, pp. 252–259.
15. Binns, R. (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), pp. 514–524. https://doi.org/10.1145/3351095.3372864.
16. Fazelpour, S., and Lipton, Z.C. (2020). Algorithmic fairness from a non-ideal perspective. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society ((AIES '20). Association for Computing Machinery), pp. 57–63. https://doi.org/10.1145/3375627.3375828.
17. Chouldechova, A. (2017). Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. Big data *5*, 153–163.
18. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (New York, NY, USA: Association for Computing Machinery), pp. 214–226. https://doi.org/10.1145/2090236.2090255.
19. Lahoti, P., Gummadi, K.P., and Weikum, G. (2019). ifair: learning individually fair data representations for algorithmic decision making. In 2019 IEEE 35th International Conference on Data Engineering (ICDE) (IEEE), pp. 1334–1345.
20. Lahoti, P., Gummadi, K.P., and Weikum, G. (2019). Operationalizing individual fairness with pairwise fair representations. Proc. VLDB Endowment *13*, 506–518.
21. Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. (2013). Learning fair representations. In Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28 (ICML'13), III–325–III–3 JMLR.org.
22. Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., and Shadbolt, N. (2018). ''It's reducing a human being to a percentage'': per- ceptions 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), pp. 1–14. https://doi.org/10.1145/3173574.3173951.
23. Grgić-Hlača, N., Redmiles, E.M., Gummadi, K.P., and Weller, A. (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 (CHE), pp. 903–912. https://doi.org/10.1145/3178876.3186138.
24. Holstein, K., Vaughan, J.W., Daumé, H., Dudik, M., and Wallach, H. (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), pp. 1–16. https://doi.org/10.1145/3290605.3300830.
25. Lee, M.K. (2018). Understanding perception of algorithmic decisions: fairness, trust, and emotion in response to algorithmic management. Big Data Soc. *5*. 2053951718756684. https://doi.org/10.1177/2053951718756684 arXiv:10.1177/2053951718756684.
26. Pierson, E. (2017). Demographics and discussion influence views on algorithmic fairness. arXiv, 1712.09124 [cs.CY].
27. Wang, R., Harper, F.M., and Zhu, H. (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), pp. 1–14. https://doi.org/10.1145/3313831.3376813.
28. Woodruff, A., Fox, S.E., Rousso-Schindler, S., and Warshaw, J. (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), pp. 1–14. https://doi.org/10.1145/3173574.3174230.
29. Eslami, M., Vaccaro, K., Lee, M.K., Elazari Bar On, A., Gilbert, E., and Karahalios, K. (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), pp. 1–14. https://doi.org/10.1145/3290605.3300724.
30. Rader, E., Cotter, K., and Cho, J. (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), pp. 1–14. https://doi.org/10.1145/3173574.3173677.
31. Thi Ngoc Trang Tran, Atas, M., Felfernig, A., Le, V.M., Samer, R., and Stettinger, M. (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), pp. 13–21. https://doi.org/10.1145/3320435.3320437.
32. Tsai, C.-H., and Brusilovsky, P. (2019). Evaluating visual explanations for similarity-based recommendations: user perception and performance. In Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization (Larnaca, Cyprus) (UMAP '19) (Association for Computing Machinery), pp. 22–30. https://doi.org/10.1145/3320435.3320465.
33. Marcinkowski, F., Kieslich, K., Starke, C., and Lünich, M. (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), pp. 122–130. https://doi.org/10.1145/3351095.3372867.
34. Veale, M., Van Kleek, M., and Binns, R. (2018). Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making (CHI '18) (Association for Computing Machinery), pp. 1–14. https://doi.org/10.1145/3173574.3174014.
35. Colquitt, J.A., and Rodell, J.B. (2015). Measuring justice and fairness. In The Oxford handbook of justice in the workplace, R.S. Cropanzano and M.L. Ambrose, eds. (Oxford University Press), pp. 187–202. https://doi.org/10.1093/oxfordhb/9780199981410.013.8.
36. Hutchinson, B., and Mitchell, M. (2019). 50 years of test (un)fairness: lessons for machine learning. In Proceedings of the Conference on Fairness, Accountability, and Transparency ((FAT* '19). Association for Computing Machinery), pp. 49–58. https://doi.org/10.1145/3287560.3287600.
37. Leben, D. (2020). Normative principles for evaluating fairness in machine learning. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society ((AIES '20). Association for Computing Machinery), pp. 86–92. https://doi.org/10.1145/3375627.3375808.

38. Lee, M.K., and Baykal, S. (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 ((CSCW '17). Association for Computing Machinery), pp. 1035–1048. https://doi.org/10.1145/2998181.2998230.

39. Harrison, G., Hanson, J., Jacinto, C., Ramirez, J., and Blase, U.R. (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), pp. 392–402. https://doi.org/10.1145/3351095.3372831.

40. Ani Saxena, N., Huang, K., DeFilippis, E., Radanovic, G., Parkes, D.C., and Liu, Y. (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 ((AIES '19). Association for Computing Machinery), pp. 99–106. https://doi.org/10.1145/3306618.3314248.

41. Green, B., and Hu, L. (2018). The Myth in the Methodology: Towards a Recontextualization of Fairness in Machine Learning. Machine Learning (The Debates workshop at the 35th International Conference on Machine Learning (ICML)).

42. Brown, A., Chouldechova, A., Putnam-Hornstein, E., Tobin, A., and Vaithianathan, R. (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), pp. 1–12. https://doi.org/10.1145/3290605.3300271.

43. Saha, D., Schumann, C., McElfresh, D.C., Dickerson, J.P., Mazurek, M.L., and Carl Tschantz, M. (2020). Human comprehension of fairness in machine learning. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society ((AIES '20). Association for Computing Machinery), p. 152. https://doi.org/10.1145/3375627.3375819.

44. Lepri, B., Oliver, N., Letouzé, E., Pentland, A., and Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes. Philos. Technology *31*, 611–627.

45. Diakopoulos, N. (2016). Accountability in algorithmic decision making. Commun. ACM *59*, 56–62.

46. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM Comput. Surv. *51*, 42. Article 93. https://doi.org/10.1145/3236009.

47. Dourish, P. (1997). Accounting for system behaviour: representation, reflection and resourceful action. Comput. Des. context, 145–170.

48. Edwards, L., and Veale, M. (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*, 18.

49. Kasinidou, M., Kleanthous, S., Barlas, P., and Otterbacher, J. (2021). I agree with the decision, but they didn't deserve this: future developers' perception of fairness in algorithmic decisions. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (Virtual Event, Canada) (FAccT '21) (Association for Computing Machinery), pp. 690–700. https://doi.org/10.1145/3442188.3445931.

50. O'Neil, C. (2016). Weapons of Math Destruction: How Big Data Increases In-Equality and Threatens Democracy (Crown Publishing Group).

51. Dietvorst, B.J., Simmons, J.P., and Massey, C. (2015). Algorithm aversion: people erroneously avoid algorithms after seeing them err. J. Exp. Psychol. Gen. *144*, 114.

52. Liu, Y., Radanovic, G., Dimitrakakis, C., Mandal, D., and Parkes, D.C. (2017). Calibrated fairness in bandits. arXiv, preprint arXiv:1707.01875.

53. Binns, R. (2017). Fairness in machine learning: lessons from political philosophy. Proc. Machine Learn. Res. *81*, 1–11.

54. Buolamwini, J., and Gebru, T. (2018). Gender Shades: intersectional accuracy disparities in commercial gender classification. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency Proceedings of Machine Learning Research, *Vol. 81*, S.A. Friedler and C. Wilson, eds. *(PMLR), pp. 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html.*

55. Kyriakou, K., Barlas, P., Kleanthous, S., and Otterbacher, J. (2019). Fairness in proprietary image tagging algorithms: a cross-platform audit on people images. In Proceedings of the International AAAI Conference on Web and Social Media, *13*, pp. 313–322.

56. Glymour, B., and Herington, J. (2019). Measuring the biases that matter: the ethical and casual foundations for measures of fairness in algorithms. In Proceedings of the Conference on Fairness, Accountability, and Transparency (New York, NY, USA: (FAT* '19). Association for Computing Machinery), pp. 269–278. https://doi.org/10.1145/3287560.3287573.

57. Wong, P.-H. (2019). Democratizing algorithmic fairness. Philos. Technology *33*, 1–20. https://doi.org/10.1007/s13347-019-00355-w.

58. Kay, M., Matuszek, C., and Munson, S.A. (2015). Unequal representation and gender stereotypes in image search results for occupations. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (Seoul, Republic of Korea) (CHI '15) (New York, NY, USA: Association for Computing Machinery), pp. 3819–3828. https://doi.org/10.1145/2702123.2702520.

59. Sweeney, L. (2013). Discrimination in online ad delivery. Queue *11*, 10:10–10:29. https://doi.org/10.1145/2460276.2460278.

60. Silberg, J., and James, M. (2019). Tackling Bias in Artificial Intelligence (And in Humans) (McKinsey Global Institute).

61. Berelson, B. (1971). Content Analysis in Communication Research (Hafner), pp. 16–25.

62. Hsieh, H.-F., and Shannon, S.E. (2005). Three approaches to qualitative content analysis. Qual. Health Res. *15*, 1277–1288.

63. Houh, E.M., and Kalsem, K. (2015). Theorizing legal participatory action research: critical race/feminism and participatory action research. Qual. Inq. *21*, 262–276.

64. Robert Thompson, A. (2012). Qualitative Research Methods in Mental Health and Psychotherapy: A Guide for Students and Practitioners (John Wiley & Sons).

65. Braun, V., and Clarke, V. (2006). Using thematic analysis in psychology. Qual. Res. Psychol. *3*, 77–101.

66. Fereday, J., and Muir-Cochrane, E. (2006). Demonstrating rigor using thematic analysis: a hybrid approach of inductive and deductive coding and theme development. Int. J. Qual. Methods *5*, 80–92.

67. Thomas, D.R. (2006). A general inductive approach for analyzing qualitative evaluation data. Am. J. Eval. *27*, 237–246.