A Responsive Engagement Approach to Promote the Development of 'Fairer' Algorithms

A Responsive Engagement Approach to Promote the Development of 'Fairer' Algorithms

Helena Webb[1], Alan Davoust[2], Michael Rovatsos[3], Menisha Patel[1], Ansgar Koene[4], Marina Jirotka[1]
[1]University of Oxford, Oxford, United Kingdom
[2]Université du Québec en Outaouais, Gatineau, Canada
[3]University of Edinburgh, United Kingdom
[4]University of Nottingham, United Kingdom
helena.webb@cs.ox.ac.uk
alan.davoust@uqo.ca
mrovatso@inf.ed.ac.uk
menisha.patel@cs.ox.ac.uk
ansgar.koene@nottingham.ac.uk
marina.jirotka@cs.ox.ac.uk

Abstract: There is much contemporary concern about 'unfairness' in algorithmic systems. Public controversies have arisen over lack of transparency and accountability in the development and application of algorithmic systems, as well as their potential to produce outcomes that are systematically unfavourable to certain groups. As a result, a variety of fairness criteria and metrics have been proposed to guide the development of algorithmic systems. However, it is unclear whether and how wider society can be involved in deciding which of these different fairness criteria should be favoured. Our work addresses this question by drawing on Responsible Innovation (RI) and 'society in the loop' (STIL) approaches. These suggest that the development of 'fairer' algorithmic systems may be facilitated through responsive engagement with societal stakeholders. We conducted an exploratory study to determine whether it is possible to present a complex set of algorithms to lay stakeholders in a way that enables them to make informed decisions about them. We presented participants with two limited resource allocation scenarios and a set of algorithms; we then asked them to select which of the algorithms they most and least preferred for the allocation. We collected quantitative data recording participant selections and qualitative data capturing how participants explained and justified their selections. We found that participants were able to meaningfully interrogate the algorithms presented to them and displayed grounded understanding of the consequences of different selections. Whilst there was no overall consensus in either scenario, participants displayed patterns in their reasoning. They consistently treated their decisions as contingent on specific understandings of fairness and context, and different interpretations of these matters accounted for different preference selections. These insights and the approach itself can be incorporated into co design processes for contemporary algorithmic systems.

Keywords: algorithms, fairness, society in the loop, Responsible Innovation, explanation, transparency, engagement

1. Introduction

There is much contemporary concern about 'unfairness' in algorithmic systems. Public controversies have arisen over the potential for automated systems to produce outcomes that are systematically unfavourable to certain demographic groups. High profile instances include the roll-out of facial recognition software systems that tend to accurately identify fair skinned men more often than they identify darker skinned women (Buolamwini and Gebru, 2018), and the discovery that searches made on platforms such as Google can produce results that reflect, and arguably reinforce, stereotyped and prejudicial attitudes towards gender and race (Guarino, 2016). An important aspect of the problem is that these systems tend to be developed and applied through complex, unshared processes which are concentrated into the hands of a few, whilst their outcomes have far-reaching consequences. A further complicating factor is that the high degree of technical complexity underpinning algorithmic systems means that they can be difficult for lay people to understand. Opening up the processes of algorithm development and application to non-specialists requires the identification of methods that can enable lay people to meaningfully interrogate an algorithm and develop informed judgments about it.

We explore the potential to move towards a *responsive engagement* approach to promote the development of 'fairer' algorithmic systems. This involves as a first step engaging with participants and soliciting their informed opinion regarding the application of particular algorithms in a given context. This provides a means to collect

nuanced judgments from societal stakeholders which can then be made available to developers and incorporated into processes of design and development. In this paper we: discuss relevant literature relating to 'society in the loop' and Responsible Innovation approaches; outline the methods we adopted to develop and test our responsive engagement approach; set out our findings; and discuss their implications.

2. Background and related work

2.1 Society in the loop

Fairness has become an important concern in the engineering of algorithmic systems, i.e. systems that deliver automated decisions, usually based on machine learning or optimisation algorithms. One great difficulty however, is the lack of a clear yardstick to measure or validate fairness. Instead, we have a variety of possible fairness criteria, some applicable to binary decision problems (Hardt et al, 2016; Kusner et al, 2017) and others applicable to other classes of problems, e.g. fair division or allocation problems (Moulin, 2004). Many of these criteria have been shown to be incompatible or in trade-offs with one another (Kleinberg, Mullainathan and Raghavan, 2016; Binns, 2018), and some of them are possibly misguided (Friedler, Scheidegger and Venkatassubramanian, 2016).

A promising direction of research seeks ways of directly incorporating human judgment into algorithmic decisions, following the "society in the loop" (SITL) principle (Rawhan, 2018). Within the SITL perspective, an important approach has been to crowdsource human judgments on fairness or other ethical questions and systematically aggregate these judgments into a decision algorithm capable of handling new cases. The main instance of this approach is the MIT "Moral Machines" project, that crowdsourced human judgments on a large number of 'trolley problems' relating to self-driving cars (Bonnefon, Shariff and Rawhan, 2016; Noothigattu et al, 2018). Within other communities, such as Human Computer Interaction (HCI), the focus has been on understanding people's perception of fairness in different contexts (Gal et al, 2016; Binns et al, 2018; Veale, Van Kleek and Binns, 2018; Woodruff et al, 2018). An important conclusion from the recent Woodruff (2018) study cited above is the need to engage with users in order to understand their fairness concerns and develop trust.

2.2 Responsible Innovation

The SITL approach overlaps with Responsible Innovation (RI), also known as Responsible Research and Innovation (RRI), as both fields emphasise the value of an inclusive process for the development of ethical technologies. RI emerged from concerns surrounding the societal and ethical consequences of novel technologies (Owen, Macnaghten and Stilgoe, 2018) and holds concepts of responsibility and fairness at its core. Central to RI is to enable an inclusive, reflexive and accountable research and innovation process. This is for the most part achieved through the development of processes and mechanisms which ensure the involvement of relevant stakeholders throughout the entirety of the research and innovation life cycle (Von Schomberg, 2013). Bringing together stakeholders and developers to facilitate informed and mutually beneficial engagement is recognised as a challenging task but one that offers great benefits to achieving societally desirable outcomes. In relation to the development of algorithms, this would likely involve a contextualised consideration of an algorithm to determine the most relevant stakeholders. Following this, mechanisms such as stakeholder workshops and focus groups could be integrated into the cycle so that stakeholders could share their views with developers in a meaningful way. Importantly, those developing the algorithm would take these perspectives and concerns into account and find ways to embed them into their ongoing work.

2.3 Towards a responsive engagement approach to algorithm development

SITL and RI offer solutions to help facilitate the development of `fairer' algorithms. Both specify this should be an inclusive process that involves the participation of societal stakeholders to form a vital feedback mechanism for developers. There is little precedent for such an approach: while several studies have qualitatively investigated people's perception of fairness regarding different algorithmic processes (Friedler, Scheidegger and Venkatassubramanian, 2016; Lee and Baykal, 2017; Binns et al, 2018; Woodruff et al, 2018) to the best of our knowledge, only one practical case study (Lee, Kim and Lizandro, 2017) proposed to give the stakeholders a say in how an algorithmic system *should* work. This is an important difference. We argue, - concurring with Lee, Kim and Lizandro - that preference elicitation must go beyond asking what is conceptually fair. Focusing

on the conceptual leaves out key contextual factors, which might include elements such as incentives, cultural factors and the broader social consequences of an algorithm. These features can be highly contingent to preference selection so should form part of investigations into participant preferences. For this reason, our approach grounded the set of algorithms to be discussed within specific contexts.

3. Methods

The research described here formed part of a broader project called 'UnBias', which explored users' experiences of algorithmic systems (Webb et al, 2018). We conducted an exploratory study to gauge the capacity for soliciting the informed preferences of stakeholders as part of a responsive engagement approach to algorithm design. We ran focus groups and an online survey in which participants were presented with a set of five algorithms and asked to select their most and least preferred algorithms to resolve limited resource allocation problems. The problems were presented in two specific case study scenarios, both concerning the allocation of courses to university students. We collected quantitative and qualitative data for analysis. This elicited opinion regarding the optimisation criteria for non-opaque algorithms to be applied in a specific context.

Case study scenarios
We developed two case study scenarios, each based on a limited resource allocation problem. The first was introduced as follows:

> In the undergraduate Computer Science programme at the University of X, second year students take an elective (non Computer Science) course. This course will represent 5% of their academic credit for the year. There are 48 students and exactly 48 places available across 10 possible courses they could take. The number of places available to students on each course is as follows:
>
> Digital photography = 5 places
> Introduction to film studies = 4 places
> *[...]*

Eight other course titles followed, with a comparable numbers of places, totalling exactly 48 places. It was then explained that the University programme organisers wanted to assign these places to the students using an algorithm that would consider the students' preferences, but could not necessarily satisfy them all:

> The students have expressed their preferences for which course they would most like to take. They have rated each of the 10 courses on a scale from 1 to 7 representing how happy they would be if the course was assigned to them:
>
> 1 = very unhappy
> 2 = unhappy
> 3 = slightly unhappy
> 4 = indifferent
> 5 = slightly happy
> 6 = happy
> 7 = very happy

The ratings that the students gave the courses were intended to quantify the *utility* that they would receive from their allocation. Presented as such, this provided information to participants in an intuitive form.

Then five possible algorithms to assign the courses were presented:

1) Priority to high-performing students: the students are ranked according to their grades, and are given their choices in this order, with the last students being forced to pick among the remaining places.
2) Efficiency: The algorithm tries to satisfy everyone's preferences without considering grades, and selects the allocation that maximises the sum of utilities received by all students.
3) Maximin: The algorithm again tries to satisfy everyone's preferences regardless of grades, and selects the allocation that maximises the lowest utility value received by any individual student.
4) Efficiency, grade-weighted: The algorithm attempts to satisfy everyone's preferences, but weighs each student's utility with the student's grade average. The algorithm selects the allocation that maximises the weighted sum of utilities. This favours high-performing students, although not as strongly as algorithm 1.
5) Maximin, grade-weighted: in this algorithm, first the best possible maximin value is calculated (algorithm 3), without considering grades. Then utilities are weighed by student grades, and the allocation is

recalculated to maximise the weighed sum, but this time under the constraint that no student receives less than the previously calculated maximin value.

These algorithms were chosen in order to cover key dimensions and concepts of distributive fairness (Sen, 1974; Moulin, 2004) as well as to cover notions brought up by participants in an earlier pilot study of this approach (Webb et al, 2018). The algorithms were explained in more detail, and their effects were illustrated with histograms representing the probabilities, under a given algorithm, that the students would receive an allocation they rated 1, 2, ... or 7. These histograms were generated by simulating many resource allocations with preference data sampled from an earlier pilot study. For each student in the simulation, a grade was given randomly, sampled from the grade distribution of British GCSE exams.

The second scenario was formulated in an identical way, but the students considered were medical students and they were to be allocated medical specialisms such as oncology or gynaecology, which would largely determine their future professional activity. This 'raised the stakes' by requiring decision-making on an issue involving longer term and more serious consequences. The introduction of this change was motivated by the results of our pilot study, where we had already observed the importance of context, and we were interested to assess the influence of such context variables on participants' preferences.

3.1 Study format and participants

Participants were presented with the two scenarios and the set of five algorithms. For each scenario they were asked to select their most and least preferred algorithms for the allocation. It was possible to select more than one algorithm for each response. They were also asked to give reasons for their selections. This was done in two formats. In an online questionnaire participants ticked boxes to indicate their preferences and then typed free text responses to explain their reasoning. We received 72 responses to the questionnaire. Participants were a mix of postgraduate students and working professionals. We also held three focus group sessions. In these, participants filled out a paper questionnaire to select and explain their preferences for scenario 1 and then discussed their answers as a group. This was then repeated for scenario 2. The discussions were audio recorded. Two focus group sessions were held in Austria and involved 14 postgraduate students and 16 professionals respectively. One session was run in the UK with a group of 4 postgraduate students. In total our dataset consisted of 106 algorithm preference selections with written explanations, supplemented by 6 hours of audio recording. The selections were analysed quantitatively. The written explanations and verbal discussions were analysed qualitatively using an inductive, thematic analysis approach (Silverman, 2001).

4. Findings

The analysis produced a wealth of findings that help us to understand how our participants interrogated and made sense of the algorithms presented to them. The combination of quantitative preference selections and qualitative explanations was particularly illuminating. Our key findings highlight that:

1) Our approach presented participants with a set of algorithms in a way that enabled them to make an informed choice over algorithm preference;
2) There was no consensus over preferred and least preferred algorithms but participants did tend to shift their selections when the context of application changed;
3) Participants consistently drew on normative and contextual reasoning in their selections. Differences of opinion occurred when they invoked alternative understandings of fairness and context to support their selections;
4) There may be some systematic reasons underpinning differences of opinion.

We now detail these findings and illustrate them with examples from our data. Overall, the findings demonstrate that our approach offers a valid methodology that can inform the co-design of algorithms with non-expert users.

4.1 Engaging people

Our participants were able make sense of the information presented to them in the two case study scenarios and the set of five algorithms. This understanding was displayed in the comments that accompanied preference selections. For instance, participants consistently and accurately: used the same vocabulary as the

descriptions given; described the allocations shown in the graphs; referred to the features of the algorithms; deduced the likely consequences of the allocations; and matched the characteristics of the algorithms chosen as preferred/least preferred.

All participants can be considered lay users as none were professionals involved in algorithm design. However, given the complexity inherent to algorithmic systems and their terminology, we were particularly interested in how well participants from non-technical backgrounds understood and drew on the information that was presented to them. Therefore, we collected demographic details about participants' educational and professional backgrounds and cross-referenced them to their qualitative responses. In the vast majority of instances, the responses of non-technical participants mirrored our overall observations. We identified fewer than 5 comments that indicated a potential misunderstanding. The number of non-technical participants was approximately 20% of the total. This is relatively small but nevertheless we can conclude that our participants demonstrated through their responses that they were accurately interrogating the algorithms and making informed decisions. This suggests it is possible to present lay participants with a set of algorithms in a way that enables them to make an informed preference selection.

4.2 Diversity of opinion

We were interested to determine whether a diverse range of people would reach agreement on the preferred algorithm for a particular scenario. In this respect the answer is clearly negative: our participants made very different preference selections. Extensive discussion in the focus groups illuminated the reasoning behind these selections, but did not bring the participants towards any consensus. However, there were some interesting patterns, which can be described in relation to the two main fairness dimensions that the proposed algorithms touched on: firstly, the trade-off between maximising overall utility and ensuring that everyone obtains some minimal level of utility, and secondly regarding the priority to be given (or not) to high-performing students.

In both scenarios there was a clear tendency to choose algorithms that guaranteed a minimal level of utility to all students: algorithms 3 and 5 together represent 77% of the participants' choices, with over 60% of participants selecting only algorithms from this set across both scenarios. In contrast, algorithms that maximised efficiency (algorithms 2 and 4) accounted for only 17% of choices, with fewer than 10% of participants choosing only algorithms from this set. However, approximately one third of participants did not take a clear stand on this question, selecting algorithms from both sets as most preferred in one scenario or the other. We also note that overall there was no clear change in this dimension from scenario 1 to scenario 2.
On the other hand, participants' preferences regarding the other fairness aspect (i.e. whether high-performing students should be given priority, and to what extent) exhibited a different pattern: the participants were quite evenly split between wanting to include this aspect and not, but there was a clear shift from the first scenario to the second, towards more priority given to high-performing students.

This can be visualised by placing the five algorithms along a spectrum reflecting the priority given to high-performing students. At one end, algorithms 2 and 3 did not differentiate based on student grades, and at the opposite end, algorithm 1 gave very strong priority to high-performing students. Algorithms 4 and 5 struck a balance between these, and can be placed in the middle. We can add further intermediate points to place the preference profiles of respondents who selected several preferred algorithms with different priority levels, e.g. algorithms 3 (left) and 5 (middle): we can locate this profile at the second point from left. Using this idea, we can visualise the shift from scenario 1 to scenario 2 (Figure 1). We observe a very clear tendency towards more priority for strong students, but also a significant number of respondents who did not change their preferences and a small number of switches in the opposite direction.

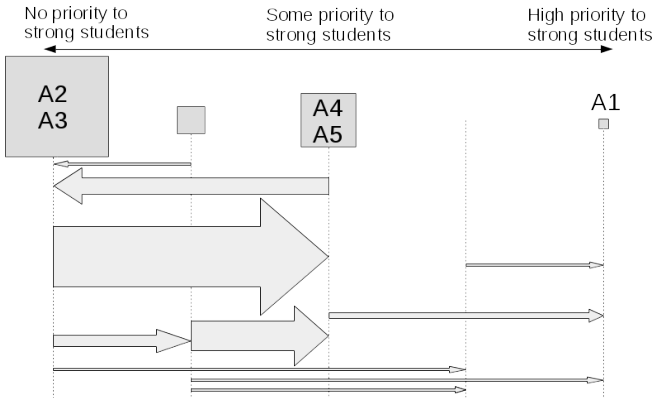

Figure 1: Visualisation of the shift in preferences from scenario 1 to scenario 2: the squares represent participants who did not change their preferences between scenarios (the size of the square is proportional to the number of respondents of each profile type), and arrows represent shifts between profile types. Again, the arrows are proportional to the number of respondents shifting their preferences the same way.

4.3 Understanding People's Preferences

The qualitative analysis enabled us to identify the reasoning that participants drew on when they made and justified their selections. Two key concerns dominated participant decision-making: fairness and context. Participants consistently treated their selections as contingent on these issues. However, their understandings of what constitutes fairness and what role context plays in the algorithm allocations differed. These considerations were deeply grounded in the specifics of the case studies - suggesting that generalising these preferences to other scenarios would not be accurate.

Fairness

Participants consistently invoked fairness as a consideration in their selections - even though the questionnaire did not directly prompt them to make selections on this basis. Participants indicated that they selected least preferred algorithms because they were unfair and most preferred as they were most fair/least unfair. Fairness arguments were put forward using various vocabulary and participants invoked different understandings of what constitutes fairness in their explanations. A major point of difference occurred over the inclusion of consideration of academic merit in the allocation. As the illustrative examples below show, this was positioned as both fair (Example 1) and unfair (Example 2), and, as noted above, some participants changed their views as the scenario changed (Example 3).

Example 1 referring to scenario 1: *I think [algorithm] 5 strikes the best balance. It will allow most of the best students to excel in something that they are interested in, and giving lower performing students a good chance of studying something they want...*
Example 2 referring to scenario 1: *I am not happy with any system that gives preference to the highest performing students. I would prefer a more egalitarian system that tried to minimise disadvantage.*
Example 3 referring to scenario 2: *This was an interesting change and has made me think the opposite to the previous example! For specialism I think the achievements of the students should be taken into account...*

Some respondents associated fairness with sameness, preferring algorithms that produced an even distribution of satisfaction across the student population. Sameness was particularly associated with fairness by participants who preferred algorithms that did not include consideration of academic merit (Example 4). For those participants that did prefer algorithms including academic merit, fairness was often constituted as a balance between the features of happiness and merit (Example 5).

Example 4 selecting algorithm 3 in scenario 1: *...it does not give preference to academic performance so it seems fairer as everyone (regardless of academic performance) gets something that they rate at least a four.*
Example 5 selecting algorithm 5 in scenario 1: *I think it is the fairer and more balanced algorithm. It takes into account student performance but also preferences*.

Participants invoked further models of fairness. For some participants, the happiness of individual students was treated as less relevant than the collective allocation across the cohort. Another interpretation emphasised social and/or welfare issues. This was particularly invoked when participants' least preferred algorithms included consideration of academic merit. Several participants made reference to socio-structural issues such as class and wealth. Here academic performance was positioned as (a potential) result of social inequalities, and algorithms that prioritise it as deepening those inequalities.

In many instances the fairness models articulated by participants are incompatible: preferences to either include or exclude academic merit as a consideration cannot be reconciled within one scenario, and neither can preferences for balance versus sameness. Understandings that emphasise social and welfare issues are also very likely to conflict with those that emphasise collective outcomes. One final model of fairness is illustrated in Example 6; some participants highlighted fairness as contingent not on the algorithm selected but the process in which it was applied.

Example 6: ... *It is not really important what algorithm you use because, from my perspective, it's more important that you make transparent that you pre-deliver equal and from that point on, I think students won't care much about what the exact algorithm is.*

Context

Participants consistently explained their preferences in relation to the contextual detail of the scenarios. In most cases, these references concerned the perceived impact of the allocation. Algorithms were selected on the grounds of their positive consequences and rejected on the grounds of their negative ones. The impacts often related to the ways that students might be affected by the allocation in terms of psychological well-being and academic/career development. At times they also related to impacts beyond the students in the allocation; for instance, through comments that people would lose trust in educational institutions if they were viewed to be treating students unfairly, and through reflections on the wider societal consequences that might arise from having 'unhappy doctors'.

Participants referred to scenario 2 as more 'serious' due to the greater potential for long lasting (negative) consequences. The increased potential for serious impact was drawn on as both a rationale for changing (Example 7) and not changing (Example 8) preference selection in the move from Scenario 1 to 2. In a few cases the entire premise of an algorithmic allocation was rejected (Example 9).

Example 7: *Since people will be allocated a choice that will probably determine the rest of their career, the argument for minimax over utility is even stronger than the first scenario.*
Example 8: *This differs from the first scenario in two key ways. Firstly, these students are in their 6th year of study and their choice of specialism will affect their future career path. Secondly, their study of medical practice has profound consequences for the health and safety of the population at large. In light of this Algorithm 5 is most preferable as it considers both the well-being of the students as well as their aptitude...*
Example 9: *That's nothing one can decide by algorithm!*

Sometimes participants also reflected on context in terms of the factors lying behind the algorithm. This often involved comments highlighting the social and psychological factors that might lie behind academic grades - and thus make an allocation based on them unfair. Whilst most participants drew on an unspoken assumption that the preferences expressed by students regarding the courses in each scenario were `genuine', a few pointed out that they might be made in order to game the algorithm and manipulate the allocation (Example 10).

Example 10: *In algorithms 2 to 5, I have a good chance to get what I want if I select value 7 for one specialism and value 1 for all the other ones, whatever my marks were...*

Context and fairness were intertwined. Participants invoked and drew on a wide range of contextual factors as relevant to their decisions over what was and was not considered fair. For instance, an understanding that social inequalities might shape academic grades led some participants to reject algorithms including considerations of academic merit as unfair. The emphasis placed by participants on context suggests it can be very difficult to make meaningful preference decisions when algorithms are presented as abstract and without context. Similarly, participants cannot make meaningful decisions about fairness without a context to ground

their reasoning within. Furthermore, given that participants ground the reasoning for their preference selections contextual detail - and tend to change their preferences when the scenario changes - there is no evidence to suggest preference selections can be generalised.

4.4 Systematic patterns in participant reasoning

Participant interpretations might not always relate only to personal preferences. During the two focus group sessions in Austria, participants largely expressed aversion to any algorithm that included consideration of academic merit. During the discussion periods, they explicitly related this to the 'no one left behind' principle of the Austrian education system that fosters inclusivity ahead of merit. Participants referred to their own selections as consistent with this cultural principle. In responses to the questionnaire similar relationships were drawn between understandings of fairness and the educational contexts of countries including Norway and France. In the UK-based focus session, one participant postulated that women might be more likely to reject algorithms including considerations of academic merit as they would be more likely to have direct experience of the ways that the demands of child rearing can negatively affect academic performance. The potential for cultural or demographic factors to influence preferences presents an interesting area for further study; it also reveals another complicating factor in reaching universal agreement in algorithm preference.

Additionally, participants consistently labelled algorithms as unfair more readily than fair. When explaining their selections, they also stated they found it easier to select their least preferred and often made choices based on what was ‘least worst’. They indicated that they were not always totally comfortable with their selections. This difficulty in making decisions about what is ‘fair’ has implications for the use of stakeholder engagement methods to inform the development of `fairer' algorithms. This is discussed next alongside other implications of our findings.

5. Discussion

This study was motivated by contemporary controversies surrounding `unfairness' in algorithmic systems and concerns that development processes lack transparency and accountability. Drawing on SITL and RI perspectives, we are interested in the capacity for a responsive stakeholder engagement method to foster the co-design and development of ‘fairer’ algorithms. We conducted an exploratory study to assess the value of this approach by collecting participant preferences over algorithms to be used in limited resource allocation scenarios. Our findings demonstrate that:

- It is possible to present a set of algorithms to lay stakeholders in a form that enables them to make an informed preference selection.
- It is unlikely that a diverse range of participants will reach consensus on the preferred algorithm for a particular scenario.
- Participants' preference selections are underpinned by reasoning relating to: (i) fairness - the best algorithm for a scenario is referred to as the one that is the most fair, or least unfair; and (ii) context - preference selection is contingent on the detail of the situation at hand. Participants' understanding of fairness and context, and the roles they played in the scenarios, differed and this led to the spread of preference selections made. It is worth noting that participants also grounded understandings of fairness in broad perceptions of context - invoking considerations of individual, institutional and societal factors.
- There may be systematic reasons that underpin preference selections - for instance cultural contexts might influence participant understandings of fairness

Our approach and findings have genuine novelty. Our methodology moves beyond conceptual fairness and instead elicits views that are grounded in context. As a result, our findings indicate the kinds of understandings that participants draw on when asked to select their preferred algorithms in a given scenario and the kinds of complexities this presents for developing ‘universally fair’ algorithms. These complexities are suggested in the existing literature and we here provide empirical evidence to show how they become visible in the concerns expressed by lay users. We intend to develop our approach further, in particular to conduct data collection with more stakeholders from non-technical backgrounds, and explore the relationships between expressed preferences and demographic factors. Nevertheless, we conclude that our approach offers a valuable step towards a responsive engagement method to promote the development of `fairer' algorithms through co-

design. We have shown that we can successfully engage with participants to elicit their informed opinion and that nuanced insights that can be gathered from analysis. As a next step these insights can be incorporated into processes of design. In particular, this can contribute to requirements gathering for algorithm design and the approach itself offers a suitable method for user-centred requirements elicitation (Sutcliffe, Drew and Jarvis, 2011). More generally findings derived from this approach contribute to the debates occurring within HCI, computer ethics and machine learning communities over the problem of `unfairness' in algorithmic systems and the processes through which `fairer' outcomes might be achieved. In this study we have focused on non-opaque algorithms. However, our approach and the insights derived from our findings could also apply to some systems that use opaque algorithms as we are, in essence, focusing on the optimisation criteria the algorithms would be trained to satisfy rather than attempting to explain how the algorithms work.

We finish with a series of recommendations. We suggest that in order to optimise this kind of responsive engagement methodology:

1) Quantitative and qualitative methods should be combined. Quantitative measures display the spread of opinion and capture demographics; qualitative data provide access to the rationale underpinning selections and differences in opinion.
2) Algorithms should be presented to participants within a context. This produces more meaningful results and enables participants to reveal the issues that are salient to them.
3) For similar reasons, participants should be asked to consider fairness (either explicitly or implicitly) within a particular scenario. We must go beyond asking the question of what is fair in an abstract sense.
4) Preference selections should be treated as relevant only to the particular scenario and not generalised to other scenarios: generalisations are unlikely to be accurate.
5) Participants should be encouraged to describe what they perceive to be unfair in addition to what the perceive to be fair. Further work should also identify practices to help participants to more easily articulate their understandings of fairness.

6. Acknowledgements

We would like to thank all the participants who gave their time to take part in our questionnaire and focus group studies. This research was conducted as part of the project 'UnBias: Emancipating Users against Algorithmic Biases for a Trusted Digital Economy'. This project was funded by the UK's Engineering and Physical Sciences Research Council (EPSRC). Project reference EP/N02785X/1.

7. References

Binns, R. (2018) Fairness in Machine Learning: Lessons from Political Philosophy. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency,* PMLR 149-159.

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J. and Shadbolt, N. (2018) 'It's reducing a human being to a percentage': Perceptions of Justice in Algorithmic Decisions, *Proceedings of the 2018 Conference on Human Factors in Computing Systems,* ACM, 377.

Bonnefon, J., Shariff, A. and Rahwan, I. (2016) The Social Dilemma of Autonomous Vehicles, *Science,* 352(6293), 1573-1576.

Boulamwini, J. and Gebru, T. (2018) Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency,* PMLR 81:77-91.

Friedler, S.A., Scheidegger, C., and Venkatasubramanian, S. (2016) On the (im) possibility of fairness, arXiv *pre-print arXiv:1609.07236.*

Gal, Y., Mash, M., Procaccia, A.D. and Zick, Y. (2016) Which is the Fairest (Rent Division) of Them All? *Proceedings of the 2016 ACM Conference on Economics and Computation,* ACM, NY, USA, 67-84.

Guarino, B. (2016) Google faulted for racial bias in image search results for black teenagers, *The Washington Post* (10 June 2016).

Hardt, M., Price, E., Srebro, N. et al (2016) Equality of Opportunity in Supervised Learning, *Advances in Neural Information Processing Systems,* 3315-3323.

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2016) Inherent trade-offs in the fair determination of risk scores, arXiv *pre-print arXiv:1609.05807.*

Kusner, M.J., Loftus, J., Russell, C. and Silva, R. (2017) Counterfactual Fairness, *Advances in Neural Information Processing Systems,* 4066-4076.

Lee, M.K. and Baykal, S. (2017) Algorithmic Mediation in Group Decisions: Fairness Perceptions of Algorithmically Mediated vs. Discussion-Based Social Division, *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work,* 1035-1048.

Lee, M.K., Kim, J.T. and Lizandro, L. (2017) A Human Centred Approach to Algorithmic Services, *Proceedings of the 2017 Conference on Human Factors in Computing Systems,* ACM, 3365-3376.

Moulin, H. (2004) *Fair Division and Collective Welfare,* MIT Press, Massachusetts.

Noothigattu, R., Gaikwad, S.S., Awad, E., Dsouza, S., Rahwan, I., Ravikumar, P. and Procaccia, A.D. (2018) A Voting Based System for Ethical Decision Making arXiv *pre-print arXiv:1709.06692.*

Owen, R., Macnaghten, P. and Stilgoe, J. (2012) Responsible Research and Innovation: From Science in society to science for society, *Science and Public Policy,* 39,6, 751-760.

Rahwan, I. (2018) Society-in-the-loop: programming the algorithmic social contract, *Ethics and Information Technology,* 20,1, 5-14.

Sen, A. (1974) Rawls versus Bentham: an axiomatic examination of the pure distribution problem, *Theory and Decision,* 4,3-4, 301-309.

Silverman, D. (2001) *Interpreting Qualitative Data: Methods for Interpreting Talk, Text and Interaction,* Sage, London.

Sutcliffe, A., Thew, S. and Jarvis, P. (2011) Experience with User-Centred Requirements Engineering, *Requirements Engineering,* 16, 4, 267-280.

Veale, M., Van Kleek, M. and Binns, R. (2018) Fairness and Accountability Design Needs for Algorithmic Support in High Stakes Public Sector Decision- Making, *Proceedings of the 2018 Conference on Human Factors in Computing Systems,* ACM, 440.

Vom Schomberg, R. (2013) A Vision of Responsible Research and Innovation, In R. Owen, M. Heintz and J. Bessant (Eds.) *Responsible Innovation*, John Wiley, London, 51-74.

Webb, H., Patel, M., Rovatsos, M., Davoust, A., Ceppi, S., Koene, A., Cano, M. (2018). It would be pretty immoral to choose a random algorithm Opening up algorithmic interpretability and transparency. *ETHICOMP 2018*. Available at: https://ora.ox.ac.uk/objects/pubs:864762.

Woodruff, A., Fox, S.E., Rousso-Schindler, S. and Warshaw, J. (2018) A Qualitative Exploration of Perceptions of Algorithmic Fairness, *Proceedings of the 2018 Conference on Human Factors in Computing Systems,* ACM, 656.