"It would be pretty immoral to choose a random algorithm": Opening up algorithmic interpretability and transparency

*Webb, Helena
University of Oxford, Department of Computer Science
helena.webb@cs.ox.ac.uk

Patel, Menisha
University of Oxford, Department of Computer Science

Rovatsos, Michael
University of Edinburgh, School of Informatics

Davoust, Alan
Université du Québec en Outaouais, Department of Computer Science and Engineering

Ceppi, Sofia
PROWLER.io

Koene, Ansgar
University of Nottingham, Horizon Institute of Digital Economy Research

Dowthwaite, Liz
University of Nottingham, Horizon Institute of Digital Economy Research

Portillo, Virginia
University of Nottingham, Horizon Institute of Digital Economy Research

Jirotka, Marina
University of Oxford, Department of Computer Science

Cano, Monica
University of Nottingham, Horizon Institute of Digital Economy Research

1. Introduction

In recent years, a significant amount of public concern has emerged over the increasing pervasiveness of algorithms and the impact of automated decision-making in our lives (Floridi and Sanders, 2004; Koene et al., 2016; Binns, 2018). A number of high-profile cases have suggested that algorithms may inadvertently influence public opinion or produce outcomes that systematically disadvantage certain groups in society. Key examples include controversies over the roles

played by bots and algorithms in the 2016 US presidential election (Howard et al., 2018) and the placement of online advertisements for criminal background checks alongside searches for African-American sounding names (Sweeney, 2013).

What perpetuates these concerns and adds to their problematic nature is the lack of transparency surrounding the development of these algorithmic systems and their use (Pasquale, 2015). Algorithms developed and used by large corporations are widely used and yet proprietary, with their inner workings remaining hidden from direct scrutiny. In addition, due to the complexities of the problems they work on, many of the algorithms that now provide important services are inherently complex in their formulation. As a result, they are often only fully understandable to those who have specific technical knowledge and interest in them. This means that most of us are largely uninformed users, experiencing algorithms on a daily basis and yet unaware either of the issues, or of how to overcome them. Where there is a lack of transparency there is typically also a lack of accountability (Koene et al., 2017; Oswald, 2018). The use of algorithmic risk assessment scores to aid sentencing in US criminal courts has been accompanied by a number of controversies; one concerned the rejection of an appeal from a defendant to scrutinise the process through which his risk score and subsequent sentence had been produced (SCOTUSblog, 2017). It was ruled that knowing the outcome of the score was sufficient and that the defendant and his legal team did not have rights to access the proprietary risk assessment instrument itself.

The research reported in this paper is motivated by the desire to open up these algorithmic processes in order to make them more interpretable, transparent and subject to oversight. Some have argued for a "society in the loop" AI governance framework, where societal values would be embedded into algorithmic decision-making (Rahwan, 2018), comparable to the ways in which human judgment (from individuals) is used to train or control machine learning systems. Similarly, Responsible Research and Innovation approaches (Owen, Macnaghten and Stilgoe, 2012) advocate opening processes of innovation to include voices from across society. These perspectives highlight the need to elicit a collective judgment regarding particular algorithmic processes. Precisely how this can be achieved is challenging. How can we open up the 'black box' of algorithms to make them available for scrutiny by different groups with varying levels of technical literacy? On what basis should algorithms be judged? How does our judgment balance the interests of the different stakeholders affected by these processes and their outcomes?

This paper reports on empirical work to elicit the opinions of research participants regarding an algorithm to be used in a specific context. Presented with a limited resource allocation problem and several possible algorithms to solve it, participants were asked to choose their preferred and least preferred algorithms for the task. They were also given the opportunity to discuss these choices. Analysis of their choices and discussions shows that the participants made different preference selections but consistently invoked normative concerns when accounting for their choices. They also attended to their selections as strongly dependent on the context. This

discussion-based format formed a highly useful approach to begin opening up algorithmic interpretability and transparency.

2. Background: Exploring algorithmic transparency

It may be that in order to make algorithms more fair in their contemporary use, they should be made more transparent. So, how would this be achieved? Engendering transparency is no simple feat and many complexities exist. The notion of transparency itself has been explored extensively, with both the positive and more problematic sides in making 'the invisible more visible' revealed (for example, see Strathern, 2000). More specifically, in regard to transparency and algorithms, there exists a tension between the proprietary nature of algorithms on the one hand, and more scrutiny of algorithms to protect users on the other. Moreover, if we were to suggest all algorithms be transparent, then what does this mean in practice (Ananny and Crawford, 2016)? Users have different levels of technical literacy and access to information, so how can we *usefully* provide information to them that they can interpret in a beneficial way?

This multi-faceted problem is central to the project on which this study is based. UnBias[1] seeks to promote fairness in the design, development and use of algorithms. It explores issues surrounding the governance of algorithms; in particular, in understanding if algorithms and those who develop them could become more responsible for safeguarding users. This work largely involves interacting with stakeholder groups in order to investigate questions including:

- How can we develop ways of communicating to stakeholder audiences what algorithms do?
- How can we elicit perspectives from stakeholders that can inform the fairer design of algorithms?
- What kinds and forms of information support meaningful transparency by making algorithms available for interpretation and inspection by different stakeholder groups?

In order to explore these questions, it was necessary to devise ways to expose the complexities of algorithms to different groups of participants, including non-experts. Collecting quantitative and qualitative data from user groups can increase understanding of what a meaningful transparency might involve and how this might benefit contemporary debates over algorithm prevalence. The project research questions were operationalised into a unique study design based on a limited resource allocation problem. Participants were asked to comment on a specific set of algorithms within the context presented in the limited resource scenario. Data were collected

[1] https://unbias.wp.horizon.ac.uk/

through a series of discussion-based experiments utilising a research questionnaire. The study design is described next.

3. Study design and methods

3.1 Limited resource allocation case study scenario

In order to begin exploring algorithmic transparency, a case study was developed that required research participants to select and then discuss their preferred algorithms within a specific context. The scenario was that of a limited resource allocation problem, based on a real-world use case. It was presented to the participants as follows:

> *Students at the University of X are to be allocated coursework topics for their current course. There are 34 students and 34 topics. Each topic can only be allocated once and each student can only receive one topic to work on.*
>
> *Students have been given the opportunity to express their preferences by rating each topic according to how happy they would be if there were to be allocated it. They have rated each topic from a scale of 1 to 7 where 1 = very unhappy, 2 = unhappy, 3= slightly unhappy, 4 = indifferent, 5 = slightly happy, 6 = happy, 7 = very happy.*

The study team devised five algorithms that could be used to allocate the coursework topics in this scenario, and which differed in how they optimised for different objective functions based on the preferences given by the students. These preferences were given as numerical ratings, and interpreted as the *utility* that a student would receive from being allocated a specific topic, in a utilitarian, economics-inspired sense. The different algorithms either: i) maximised the sum of students' individual utilities (total utility), ii) maximised the lowest utility of any of the students for the allocation (focusing on limiting the "damage" to the student who was least well off given an overall allocation), or iii) minimised the sum of differences between the different students' utilities (aiming to reduce the total "distance" among all students' individual outcomes). Additional algorithms were obtained by combining several of these criteria, i.e. optimising for one while guaranteeing a certain level of another.

As this was a genuine scenario it was possible to run each algorithm on student preference rating data that had already been gathered. This generated a series of graphs and tables showing the outcomes of each algorithm in terms of utility and distance. These were then placed into a two-part questionnaire, which is provided in the appendix to this paper.

Part 1 of the questionnaire set out the case study scenario, as described above, and then presented tables and graphs showing the different utility values obtained by the students for each algorithm, as well as the mean of students' individual utilities, the total utility and the total distance between utilities. It then had a question section that asked respondents to select their most and least preferred algorithm for use in this context, and explain their selection.

Part 2 of the questionnaire provided the same graphs and tables but also provided a short explanation of each algorithm in terms of the optimisation criteria applied internally by the algorithms. A further question section asked respondents to select their most and least preferred algorithm for use in this context once again, and then explain their selection. The rationale for using a two-part questionnaire was to observe whether the type of information available about the algorithms made a difference to individual choices.

This case study questionnaire was then used in a series of discussion-based experiments, as is described next.

3.2 Discussion-based experiments and data analysis

Four groups of participants were recruited to take part in discussion-based experiment sessions using the limited resource allocation scenario questionnaire. The groups were comprised as follows:

> Group 1 - 9 participants, all undergraduate students studying computer science at a UK university.
> Group 2 - 7 participants, all post-graduates or post-doctoral level researchers in computer science based at a UK university.
> Group 3 - 10 participants, all with postgraduate-level experience in social science or law at a UK university.
> Group 4 - 13 working professionals from fields including academia, education, law, and industry[2].

Overall, 39 participants took part in these studies, the aim of which was to record which algorithms participants selected as most preferred and least preferred, and observe how they accounted for their choices. A further aim was to identify any systematic differences in preference selection between the groups.

On each occasion, the experiment was conducted in the following way: after brief introductions, the research team members facilitating the experiment outlined the case study scenario to the participants. Participants were then presented with Part 1 of the questionnaire and some time was taken to check their understanding of the graphs and tables. Participants completed the questionnaire individually by indicating their most preferred and least preferred algorithms and writing a short text to explain their choices. Participants were able to select more than one algorithm as preferred/least preferred, if necessary. Once all participants had completed the questionnaire, the research team facilitated a 10-to-20-minute group discussion. Participants were

[2] These professionals were part of a stakeholder group in the wider XXXX study and therefore had a pre-existing interest in current debates around algorithms.

asked first to report their questionnaire responses and then to explain the rationale for their selections. They were encouraged to debate with each other, in particular to explore the reasons behind differences of selection. They were also asked to comment on what further details might better help them in their decision-making. After this, participants were given Part 2 of the questionnaire and asked to complete it alone once again. After this, another group discussion was held with participants again asked to report, explain and debate their selections. They were also invited to comment on whether or not the extra information about the algorithms had led them to change their selections, and why.

The questionnaire responses were analysed quantitatively to identify patterns of selection within and across participant groups. The discussion sessions were audio recorded and transcribed. The transcripts were then analysed thematically (Richie and Lewis, 2003) to identify recurring patterns across the different groups with particular attention paid to the topics raised by participants in their discussions and the different kinds of categories and understandings they invoked (Silverman 2001, Coulthard, 1977, ten Have, 2004) in order to support their selections.

4. Findings

4.1 Quantitative findings

The quantitative analysis of the questionnaires presented a range of useful findings, which are discussed in brief here. There is a diversity in selections of least and most preferred algorithms, as shown in Figures 1 and 2.

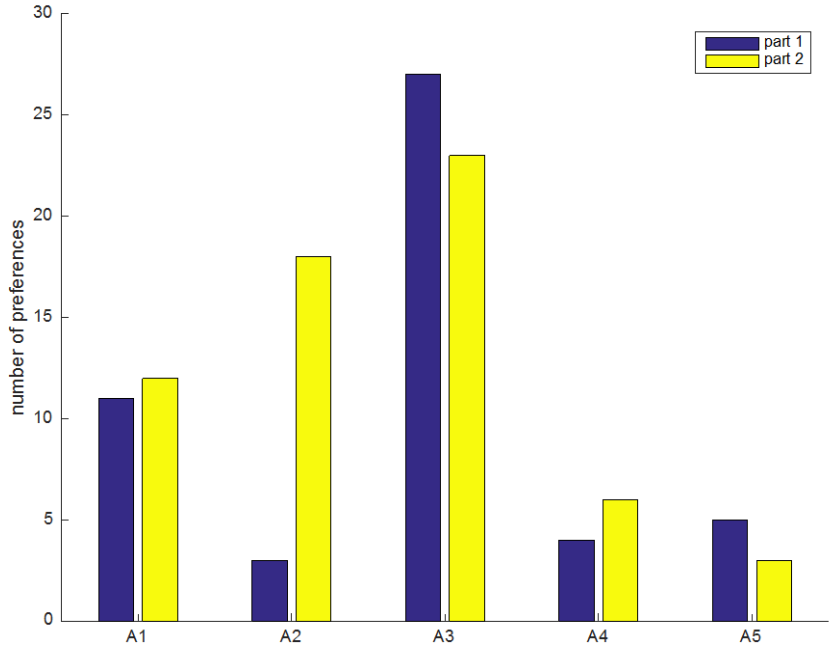


Figure 1: Most preferred algorithms in Parts 1 and 2 of the questionnaire.

Figure 1 shows participants' most preferred algorithms in both Parts 1 and 2 of the questionnaire. Algorithm 3 was the most popular choice in each part - selected 27 times in Part 1 and 24 times in Part 2. However, in both cases almost half the preferences were split amongst algorithms other than algorithm 3. A similar diversity of opinion was found in the selection of least preferred algorithms – as shown in Figure 2. Potential interpretations emerge of the value perspectives that participants drew on to produce their respective answers. Participants who chose A3 as preferred may have adhered to a value framework that focuses on maximising overall satisfaction, whereas participants who chose A1 may have emphasised the importance of minimising disparity between satisfactions of the students.

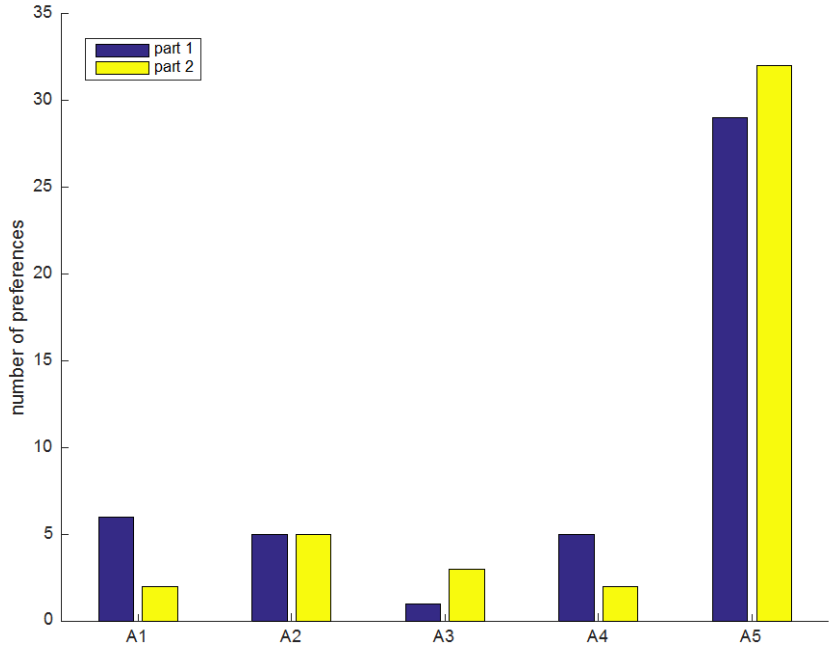


Figure 2: Least preferred algorithms, in Parts 1 and 2 of the questionnaire

The quantitative analysis also showed that some participants did change their preferences between Parts 1 and 2 of the questionnaire. For reasons of simplicity in this paper we represent these changes globally in Figures 1 and 2 through the different coloured bars. However, our analysis also included a more detailed breakdown of changes by individual participant. Changes in selection most often related to a change of most preferred algorithm; in the Part 2 responses we particularly observed an increase in preferences for algorithms that offered a trade-off between multiple criteria over those that optimised a single criterion. Overall around 30% of the algorithms selected as the most and least preferred in the Part 1 of the questionnaire were not selected as such in Part 2 by the respective participants. Eight participants completely reversed their preferences over the two parts of the exercise – with their least preferred algorithm in Part 1 becoming their most preferred in Part 2.

Due to the format of the experiment it is not clear what led to the change of opinion: it could have been due to the further information - and therefore a greater level of transparency provided about each algorithm in Part 2 – or it could alternatively have been a result of participants being persuaded by arguments put forward in the discussion session following the completion of Part 1 of the questionnaire and/or participants' own reflections as the experiment continued. Some relationships between participants' responses and their professional or educational background were observed. For instance, undergraduate participants were more likely to change their responses from Part 1 to Part 2 than any of the other participants. However, these relationships were not very strong and require further testing with a larger sample size before firm conclusions can be made. Overall, although the quantitative analysis did not yield statistically significant results, it does point towards some interesting observations regarding the ways that participants drew on the information available to them to produce their preference selections. These observations can be further unpacked through the qualitative analysis, as discussed next.

4.2 Qualitative findings

Qualitative analysis of the discussion transcripts produced a range of relevant findings revealing recurring topics raised by the participants and patterns in the ways that they discussed their preference selections. Three key findings were particularly illuminating. When discussing algorithm preference, participants: 1) consistently raised normative issues around fairness as relevant; 2) consistently related their preference to the (real or imagined) context in which the algorithms were to be applied; and 3) displayed varying levels of familiarity with technical features of algorithms. These findings are discussed below and illustrated with examples from the data transcripts - quoted in italics.

4.2.1 Fairness

Moral references, in particular references to fairness, were ubiquitous when participants discussed their preferred and least preferred algorithms. When asked to explain their selections, participants routinely began by using terms that mirrored the wording of the questionnaire – preference and student happiness. When they then went on to justify these selections it was highly noticeable that they did so using vocabulary and invoking categories that treated the selections as somehow normative. That is, rather than simply positioning the reason for their selected responses as a matter of personal preference, they instead oriented to them as a matter of right and wrong behaviour. In the three instances below each participant uses a term that explicitly relates the task to normative matters.

> Example 1: *It would be pretty immoral to choose a random algorithm, right?*
>
> Example 2: *For me I feel like whether I'm a student or a teacher, I think A1 still would be the fairest because five majority say they are slightly happy with it…*

Example 3: *I just knew that A5 was unjust because it had one – at least one unhappy person*

In Example 1 the participant states that choosing randomly would be normatively inappropriate, 'immoral', treating the process of selecting an algorithm – and by extension the consequence of that selection – as a matter of right or wrong behaviour. In Examples 2 and 3 the participants use the language of fairness – 'fairest' and 'unjust' – to provide a rationale for their most and least preferred algorithms. They indicate that their selection is based on their perception that a specific algorithm or its outcomes is more or less acceptable than the others. Participants across all groups consistently drew on moral terms and categories to explain and justify their selections, and references to fairness were by far the most common way in which they did so. As seen in the report of the survey findings, this did not necessarily lead to agreement amongst participants: whilst participants typically oriented to the relevance of fairness in their decision-making, they sometimes applied understandings of fairness in different ways, which led to differences of opinion.

Example 4: *With A1 at least everyone is slightly happy*

Example 5: [selecting, with equivocation, A2 as preferred] *They're my main priorities, like to maximise the total utility, but I don't want to have the people be very unhappy*

Example 6: *A2…actually I think it is the most balanced one*

As participants continued to rationalise their selections, they revealed the understandings of fairness that underpinned their choices. The most common understanding across the dataset, illustrated in Examples 4 and 5, begins with the assumption that each student feeling happy with their coursework allocation would be a positive and desirable outcome. As it is not possible for each student to achieve maximum happiness, fairness was located in the optimum distribution of happiness levels across the students. Fairness in this sense was often expressed by participants in terms of finding the best 'balance' of distribution – as seen in Example 6. The exact nature of that balance was a matter for debate; for some it equated to everyone feeling happy to some degree at least (Example 4) whereas for others it allowed for a higher level of variation as long as marked unhappiness was avoided (Example 5).

Example 7: *A1 …does get kind of fair result, which doesn't mean it's the best result, but it's just kind of equal.*

Example 8: [on selecting A5 as the least preferred] *yeah on the one hand it's kind of really objective, I think fairness isn't really about like the objectiveness, as such; it's accounting also for preferences…I mean like equality isn't about treating everyone the same, it's about taking into account their specific circumstances*.

Example 9: *But the student may not deserve a better topic*.

Despite the importance placed on balance, it was frequently referred to as insufficient on its own. Algorithms 1 and, in particular, 5 were often justified as least preferred on the basis that whilst they achieved an even or 'equal' (Example 7) distribution by minimising distance, this wasn't enough to achieve a good result. Taking this further some participants commented that this kind of balance was not necessarily fair, marking out a difference between sameness and fairness or equality, as seen in Example 8. Meanwhile some participants rejected understandings of fairness as based on balance in preference for alternatives – such as fairness as based on individual merit (Example 9). These distinctions and alternative understandings demonstrate the ways in which participants in the discussions constructed fairness as relevant to the task at hand but also complex and not universal. They also displayed a clear recognition that there are tensions between fairness objectives that cannot be easily reconciled.

4.2.2 Context

In the discussion sessions, participants' reasoning about algorithm preference was highly bound up with matters of context. Across the four groups, participants routinely articulated their choices in relation to the context of the task or various hypothetical contextual situations. Even when technical features of the algorithms were discussed, different contextual circumstances were also invoked. The repeated references to context suggest that selections about algorithm preference were not made in reference to abstract features of the algorithm alone but rather in relation to the application of the algorithm within the particular scenario of the case study. It was also notable that when asked during the task what additional information would be useful to support their decision-making, participants often asked for this kind of contextual detail rather than any more technical information about the algorithms themselves. At times they also constructed imagined contexts, based on their assumptions about the case study or personal experiences of student coursework tasks.

Example 10: *Like it doesn't have to be like an actual outcome, but I think it's always good to, you know, give an example.*

Example 11: *So we don't know anything about the students, right. Are they generally unhappy?*

Example 12: *Well, yeah, if the student is suitable for the project. That for me actually would be more important.*

These examples illustrate the difficulty of understanding algorithms and their implications in a solely technical or abstract sense, rather than judging them universally. In Example 10, the participant is emphatic that contextual information - real or hypothetical - is beneficial to facilitate a better understanding. The participant proposes that an example of outcomes would be bene-

ficial to the completion of the task. In Example 11, the participant asks a question about the wider context of the scenario which, even if rhetorical, suggests that this kind of further information would aid decision-making. Similarly, in Example 12, the participant attempts to reason about the specific context of the task, and in doing so articulates what criteria would be important for him, if he were to allocate projects. His individual perspective is that the more appropriate criterion is the suitability of a student to a project - detail that was not available in the questionnaire - rather than students' preferences. The ubiquity of references to context across all four groups suggests that participants did not reflect on the algorithms in abstract terms when selecting their preferences, but rather grounded their reasoning in the details of the specific case study given. It is also noticeable that, even when participants shared an understanding of the (real or imagined) context at hand, this did not necessarily mean they were in agreement about algorithm preference. These factors suggest that it would be very difficult to determine any kind of globally preferred algorithm that could span across contexts. Moreover, the frequency of participant requests for more detail about context and their occasional construction of an imagined context suggests that the algorithms themselves were understood within the context of their application rather than in abstract terms. In keeping with the quantitative findings, this indicates the importance of information to participant preference selection, albeit in a rather different way.

Across the groups expressions of algorithm preference tended to change frequently according to the nature of the context that was being considered. Participants drew on contextually-informed reasoning to explicitly justify their own stances.

> Example 13: *So the university might actually be keener on having one of the algorithms where there's a higher level of people happy... and that could, you know, feed into the overall [university] feedback as well.*
>
> Example 14: *And the reason I would go with A3 is because it doesn't seem like it's a bad thing. You're going to get a project you don't like that much, it's not that big of a deal.*
>
> Example 15: *…like if one means I'm going to die if I try to do this then you wouldn't want that. Or there's no way I'm going to ever pass the class because you gave me something I hate. Or if it just means I really don't like this, it's not the same.*

As highlighted by Examples 13, 14 and 15, there was an orientation by participants towards considering the consequentiality of algorithmic outcomes in the specific context of the case study scenario. This consideration of consequences was treated by them as hugely important in their own reasoning and decision-making about preferred algorithms. In Example 13, the participant identifies why from a university-based perspective that prioritises positive student feedback, allocating projects to maximise student happiness would be preferable. In Examples 14 and 15, the participants focus on the consequences for a student of receiving a less desirable project. In Example 15, the participant juxtaposes different extremities of outcome for the students, con-

trasting the seriousness of hating a project versus simply not liking it, and stating that this would make a difference to preference selection. This was another common feature in the discussions: levels of concern for the design and use of an algorithm were expressed as related to how problematic and serious the consequences of that use were for those affected by it. Once again, this finding indicates the difficulty in attempting to identify a cross-context preferred algorithm.

> Example 16: *It's objectively fairer because you've put certain parameters into a computer and it has then spat out a choice, whereas if you've got a professor, a lecturer or whoever doing that, then they have their - they're bringing in their own knowledge about you, your knowledge about your classmates.*
>
> Example 17: *I would not trust the algorithm more than the professor and his knowledge about me could actually enhance my project because he could then explain to me, "I gave you this one because I know you have potential. Just take the four weeks' time and you'll do great. I know you can do a better job with this one than the one you preferred ..."*
> Example 18: *An algorithm is just as unbiased as the programmer who created it.*

The notion of context was also invoked in relation to the process of decision-making itself. In particular, participants displayed consideration over whether knowledge of context would enhance or problematise the appropriate allocation of projects. These discussions once again involved references - either explicit or implicit - to fairness. In both Examples 16 and 17, automated decision-making is contrasted with human decision-making. Each example presents conflicting viewpoints on whether intricate familiarity with the local context makes the process of allocating projects more or less fair. The participant expressing a view in Example 16 sees algorithms as `objectively fairer' alluding to a professor potentially making biased decisions given his or her contextual knowledge of students. In Example 17, the wider contextual knowledge is seen as a tool to `enhance' the experience of the student, alluding to the ability of a professor to make fluid and appropriate decisions based on less stringent criteria than those the algorithm is constrained by. The reasoning of the professor is thus seen to go beyond the preferences of students to what may be better for their work. There was no consensus among participants and groups on whether the existence of contextual knowledge when making decisions was good or bad. Discussion was complex, nuanced and at times contradictory, with wider contextual knowledge seen as necessary, and then later viewed negatively as introducing bias to situations. Some participants even began to scrutinise the assumption that algorithms are neutral by considering production of an algorithm as a whole. In Example 18, the participant refers to the process through which an algorithm is produced as relevant to fairness, suggesting that since it will likely hold the values of those who designed it, it will be just as biased as its designer. This argument undermines that presented in instances such as Example 16, which attributes a neutrality to algorithms.

4.2.3 Technical features of algorithms

Much discussion in the sessions focused on the characteristics of the 5 different algorithms. Participants frequently referred to specific features of an algorithm or its results in support of their preferences. However, a given feature might be referred to both positively and negatively by different participants. All groups asked questions to clarify their understanding of the algorithms and, as noted above, were eager to learn more about the context in which the algorithms would be applied. Participants from technical backgrounds were noticeably more fluent and familiar using technical terminology whereas those from non-technical backgrounds required assistance to understand the meaning of key terms such as `utility' and to interpret the graphs shown on the questionnaires. Examples 19 to 21 illustrate the kinds of difficulties of understanding described by student participants from non-technical backgrounds.

> Example 19: [participant in non-technical student group when asked to comment immediately after reading part 1 of the questionnaire] *This makes no sense to me whatsoever*
>
> Example 20: [participant in non-technical student group, in response to being asked how far she would be able to make a decision based on just a technical description of the algorithms] *I won't get it. It's just for me a bit like gibberish... I agree with everyone else how this is just a bunch of words and... what helps me most [in this task] is that... we actually had an example.*
>
> Example 21: [participant in non-technical student group, commenting on how well he is able to understand the technical description of the algorithms] …*for me the descriptions are there because we've done the work already this afternoon, where you've basically had to explain how it works and what the terms mean and what you mean by distance and utility and things like that, reading through this now makes still not perfect but a reasonable amount of sense.*

Although this is not a surprising finding, the varying levels of familiarity with technical features of algorithms displayed by participants is significant. It demonstrates that within and across communities there will be different levels of understanding and that particular effort might be necessary to address the lack of understanding of some members. Non-technical participants were explicitly told that they were not expected to understand the questionnaires on first reading and extended periods of time were given to inviting and answering participant questions. Technical participants were similarly invited to ask questions but these were generally less forthcoming and participants were perhaps less likely to willing, given the subject matter, to admit lack of understanding in front of their peers. Further inspection of the data could perhaps usefully reveal instances where participants from various backgrounds inadvertently revealed a lack of understanding and help to identify common misunderstandings that could be addressed. More generally the data can also provide insights into the kinds of information that might better help individuals, including those from particular demographics, to understand details of algorithmic features and processes. These observations have implications for the calls for transpar-

ency and interpretability made in contemporary debates over the role of algorithms in modern life - as discussed below.

5. Discussion

This paper has reported on an empirical study designed to begin unpacking issues around algorithmic interpretability and transparency. A questionnaire and discussion-based approach provided participants with an opportunity to examine algorithms and their consequences in a specific context. An experiment was devised in which groups of participants were asked to select their most and least preferred algorithms from a predefined selection of five options. The task was contextualised in a scenario that required the allocation of coursework topics to undergraduate students. Four groups of participants took part in the study, undertaking questionnaire and discussion tasks in a two-stage process.

Quantitative analysis of the questionnaire responses showed that even though presented with the same case study scenario, participants selected different algorithms as their most and least preferred. Some participants did change their responses between Parts 1 and 2 of the questionnaire, and in the Part 2 responses there was an increase in preferences for algorithms that offered a trade-off between multiple criteria over those that optimised a single criterion. The quantitative findings point to some interesting interpretations of the kinds of value frameworks participants drew on when making their selections, and these were further unpacked in the qualitative analysis.

Qualitative thematic analysis of the discussion sessions revealed that when asked to explain their preferences, participants across the different groups raised the same core issues. They consistently invoked normative understandings of right and wrong to justify their selections, specifically using the language of fairness to argue that the preferred algorithm should be the fairest one. Opinions about which algorithm was fairest and what constituted fairness did differ however, and participants frequently attended to the difficulty or even impossibility of a single algorithm producing a fair result in all cases. Closely connected to references to fairness were references to context. Participants expressed the need for further knowledge of the context in which the algorithm would be applied or even created imagined contexts in order to aid their decision-making. Context was drawn on to support different preferences and to raise questions over the relative consequences of the application of the algorithm and over the process of decision-making itself. Finally, references were also made to the extent to which an algorithm and its consequences could or could not be easily understood and what further information would be needed to aid this understanding.

These findings make clear that even when provided with the same information, participants make different preference selections and rationalise them differently. The issues raised by participants as important to their selections resonate closely with values that have come to the fore in current debates over algorithm prevalence. There appears to be a community-level associa-

tion of a preferred algorithm as being a fair algorithm. Competing models of fairness are drawn on in expressions of preference although there may be some general favouring of models that balance out or trade off different relevant criteria such as maximising utility and minimising distance. In addition, although it is not possible to reach global agreement on fairness, it does appear possible that some groups sharing certain characteristics might be able to reach consensus. Due to the small sample of the research reported in this paper, we do not make any firm claims regarding the relationships between participants' selections and their demographic characteristics. However, in subsequent work we have drawn on a slightly amended version of the limited resource allocation scenario and questionnaire and used them with a larger and broader sample. Our initial results indicate some interesting demographic differences, with preference selections clustering around alternate algorithms according to the nationalities of the participants. These differing preference selections appear to connect to the different kinds of ethos underpinning the education system in the countries where the participants come from. This suggests a strong relationship between cultural context and perception of fairness in a specific scenario. The work we report here does indicate that agreement can be reached over which algorithms are definitely not fair. Furthermore, given the overall priority given by users to fairness, if a particular fairness model could be identified as applicable in a given scenario, then it might be possible for consensus to be reached around which algorithm is preferred. However, the importance participants placed on context suggests that it may be very difficult to safely claim that an algorithm is fair if it's applied in many different contexts. It also implies that regulatory oversight needs to use a context specific approach with requirements specified on the basis of the application domain. This finding therefore raises significant implications for algorithm design and governance.

These findings demonstrate the complexities around algorithmic transparency. When given information about a set of algorithms and the outcomes they would produce, participants in this study were able to express opinions over their appropriate application in a given scenario; they were also able to draw on the features of the algorithm as a means to articulate in detail the rationale for their own preferences. Participants engaged enthusiastically with the task and seized the opportunity to debate core issues around algorithm design, fairness, transparency and interpretability. Following the study sessions, various participants made requests for further copies of the questionnaires in order to share them with others. This demonstrates that the task provides a meaningful way to enable individuals from various backgrounds to think through relevant issues. A further conclusion from the study is that this task can also be given to the developers of algorithms as a professional development exercise to help them to identify different user perspectives.

These findings also highlight that care needs to be taken to provide information that users find relevant and that is presented to them in ways that they can understand and draw on effectively. Algorithmic transparency and interpretability are crucial but complex values. It may be that users from different educational and professional etc. backgrounds need to be given information in different ways. It might also be necessary that users of various kinds are given oppor-

tunities to express and overcome both explicitly and implicitly expressed instances of lack of understanding. Further work can be done in this area to identify what forms of information-giving best support transparency and interpretability, for instance in terms of volume of information provided, the use of technical terms and the alternate use of text, visualisation and graphics etc. In our study, all the participants had a relatively high level of education and existing awareness of algorithms; it is likely that other demographics in the general population would require very different forms of explanation to help them understand the matters at hand and express their preferences.

This study provides empirical findings that can contribute to contemporary discussions over the importance of algorithmic interpretability and transparency. The findings highlight some challenges and questions that are important for further work in this area. One challenge is to develop further mechanisms to open up algorithms for inspection and observation. An important question is: since, as indicated by these findings, transparency and interpretability are to be valued, how can they be embedded into the design and development of algorithms? How can we enable and encourage companies to open up their processes so that proprietary algorithms can become meaningfully transparent and how can this be done in a way that mitigates the potential economic consequences of this? Ultimately, who should hold the responsibility to be transparent and oversee transparency processes? These are crucial issues that need to be addressed by further work in this area.

6.References

Ananny, M. and Crawford, K. (2016), "Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability", *New Media and Society*, Vol. 20 No. 3, pp. 973-989. DOI = 10.1177/1461444816676645

Binns, R. (2018), "Fairness in Machine Learning: Lessons from Political Philosophy", in Friedler, S. A. and Wilson, C. (Eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research),* Vol. 81, pp. 1-11, PMLR, New York.

Coulthard, M. (1977), *An Introduction to Discourse Analysis,* Longman: London.

Floridi, L. and Sanders, J.W. (2004), "On the morality of artificial agents", *Minds and Machines* Vol. 14 No. 3, pp. 349-379.

ten Have, P. (2004) *Understanding Qualitative Research and Ethnomethodology*. SAGE Publications: London.

Howard, P. N., Woolley, S., and Calo, R. (2018), "Algorithms, bots, and political communication in the US 2016 election: The challenge of automated political communication for elec-

tion law and administration", *Journal of Information Technology & Politics*, Vol. 15 No. 2, pp. 81-93. DOI = 10.1080/19331681.2018.1448735

Koene, A., Perez, E., Carter, C. J., Statache, R., Adolphs, S., O'Malley, C., Rodden, T. and McAuley, D. (2016), "Privacy concerns arising from internet service personalisation filters", *ACM SIGCAS Computers and Society*, Vol. 45 No. 3, pp. 161-171.

Koene, A., Perez, E., Webb, H., Patel, M., Jirotka, M., Ceppi, S., Rovatsos, M. and Lane, G. (2017), "Algorithmic fairness in online information mediating systems", *Proceedings of the 2017 ACM on Web Science Conference*, WebSci 2017, Troy, NY, USA, June 25-28, 2017, pp. 391-392. DOI = 10.1145/3091478.3098864.

Oswald, M. (2018), "Algorithm-assisted decision-making in the public sector: framing the issues using administrative law rules governing discretionary power," *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences,* Vol. 376 No. 2128. DOI = 10.1098/rsta.2017.0359.

Owen, R., Macnaghten, P. and Stilgoe, J. (2012), "Responsible research and innovation: Drom science in society to science for society, with society", *Science and Public Policy* Vol. 39 No. 6, pp.751-760.

Pasquale, F. (2015), *The Black Box society: The secret algorithms that control money and information.* Harvard University Press, Cambridge M.A.

Rahwan, I. (2018), "Society-in-the-loop: programming the algorithmic social contract", *Ethics and Information Technology*, Vol. 20 No. 1, pp. 5-14.

Ritchie, J., Lewis, J. (2003), *Qualitative research practice: A guide for social science students and researchers*, SAGE Publications" London.

SCOTUSblog (2017), "Loomis vs Wisconsin". *Supreme Court of the United States Blo,g* Available at: http://www.scotusblog.com/case-files/cases/loomis-v-wisconsin/ (accessed 12/03/2018).

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

Strathern, M. (2000), "The tyranny of transparency", *British Educational Research Journal*, Vol. 26 No. 3, pp. 309-321.

Sweeney, L. (2013), "Discrimination in online ad delivery", *ACM Queue*, Vol. 11, No. 3. DOI= 10.1145/2460276.2460278.

APPENDIX – LIMITED RESOURCE ALLOCATION EXPERIMENT QUESTIONNAIRE PARTS 1 AND 2

UnBias project Questionnaire Part 1 **ID _____**

Consider the problem of allocating coursework topics to students where each student must be assigned exactly one topic, and each topic can only be assigned to one student.

Students express their preferences by assigning every topic a score on a scale from 1 to 7 representing how happy they would be if the topic were assigned to them (1 = very unhappy, 2 = unhappy, 3= slightly unhappy, 4 = indifferent, 5 = slightly happy, 6 = happy, 7 = very happy).

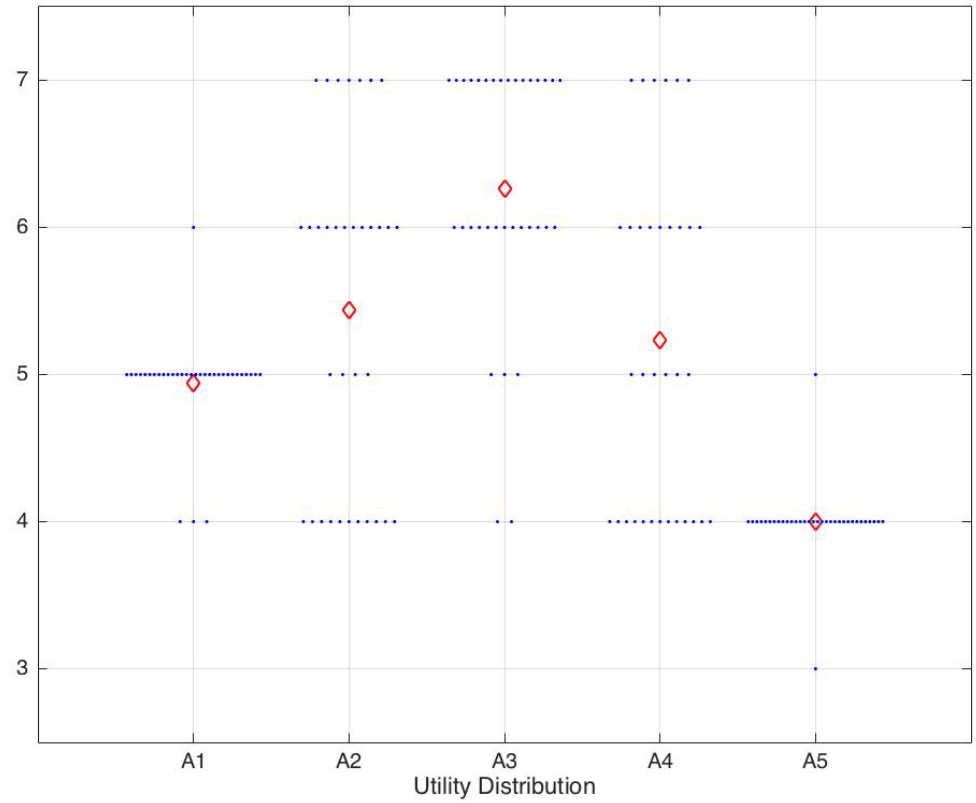
The graphs below show the distribution (blue dots) and the mean (red diamond) of students' utilities and distance between the utilities of all students computed by different algorithms.

Student's utility = the happiness level achieved based on the score the student gave to the project the algorithm assigns to him/her.

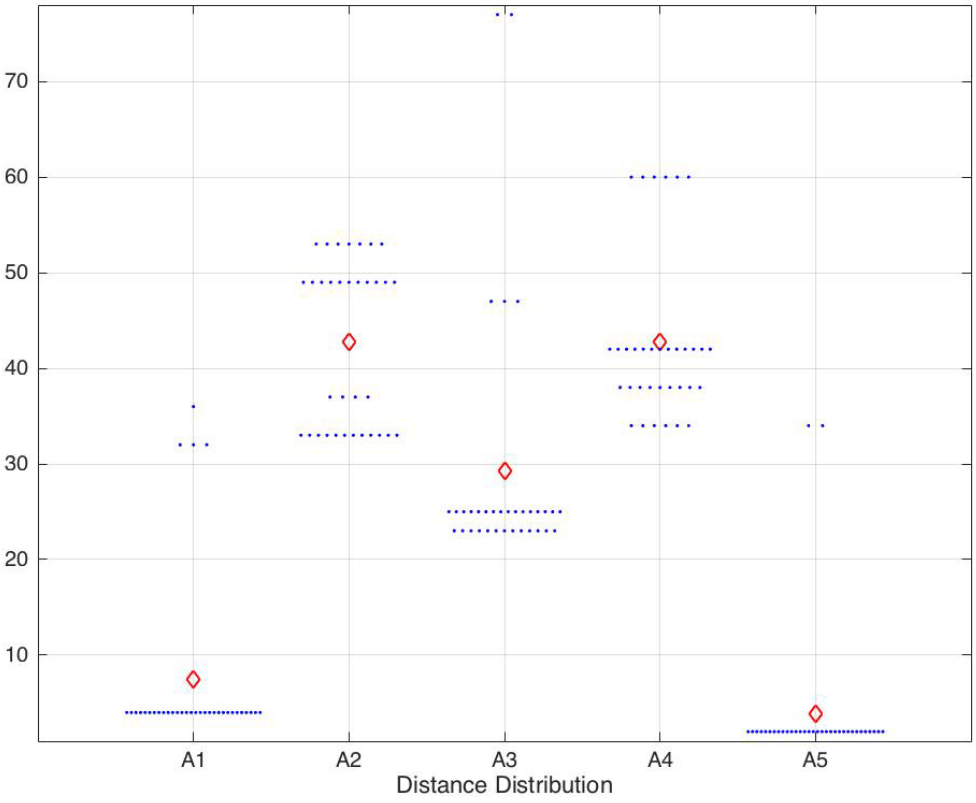
Student's distance = the total difference between the student's utility and those of all other students, given the projects assigned to everybody by the algorithm

For each algorithm, the table below shows the sum of all student's utilities (total utility) and the sum of students' distances for all students (total distance).

Algorithm	A1	A2	A3	A4	A5
Total Utility	168	185	213	178	136
Total Distance	252	1454	994	1452	132



Utility Distribution



Distance Distribution

Given the allocations computed by each algorithm, which of them would you prefer most, and which would you prefer least? You can list more than one algorithm in each line.

Most Preferred Algorithm(s): _____

Least Preferred Algorithm(s): _____

Please give reasons for your assessment:

_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____

UnBias project: Questionnaire Part 2 **ID _____**

Consider the problem of allocating coursework topics to students where each student must be assigned exactly one topic, and each topic can only be assigned to one student.

Students express their preferences by assigning every topic a score on a scale from 1 to 7 representing how happy they would be if the topic were assigned to them (1 = very unhappy, 2 = unhappy, 3 = slightly unhappy, 4 = indifferent, 5 = slightly happy, 6 = happy, 7 = very happy).

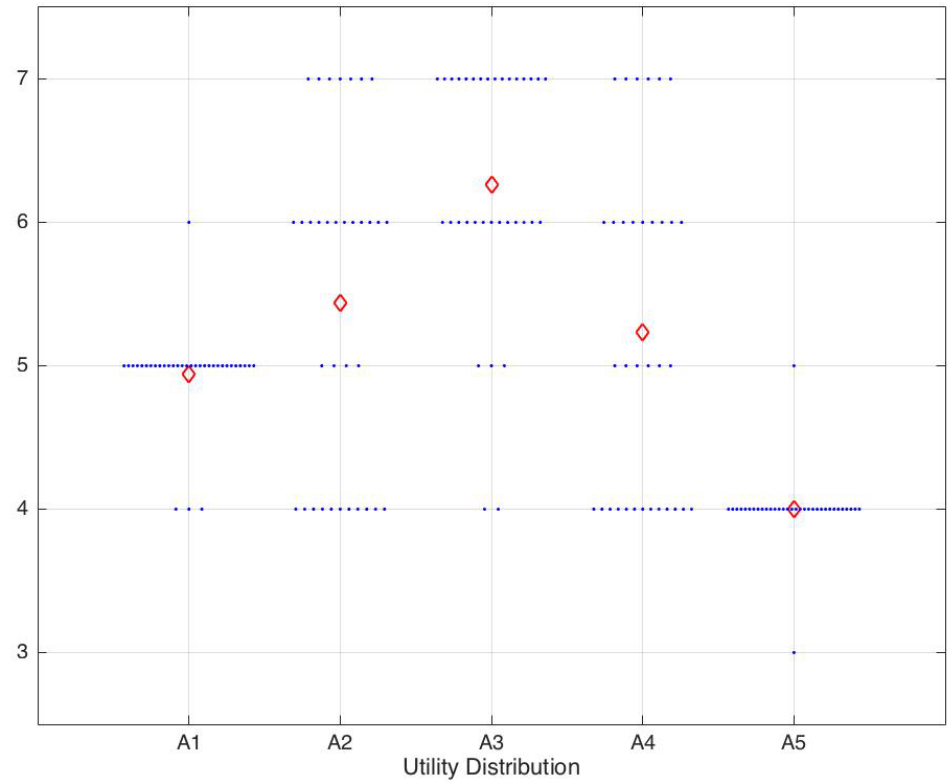
The graphs below show the distribution (blue dots) and the mean (red diamond) of students' utilities and distance between the utilities of all students computed by different algorithms.

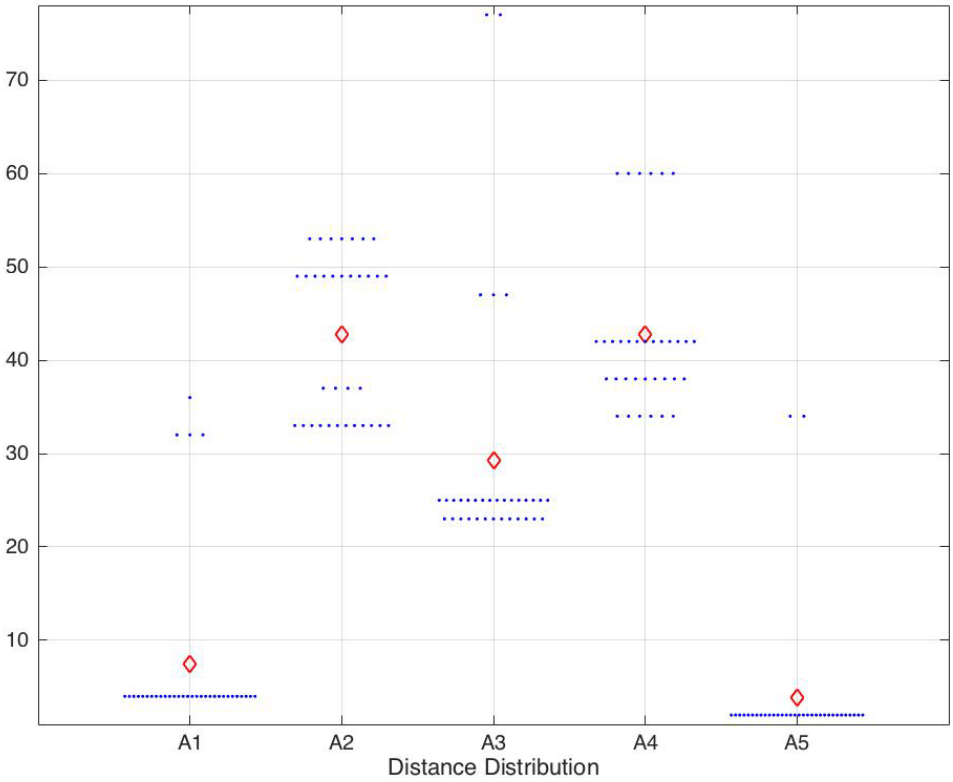
Student's utility = the happiness level achieved based on the score the student gave to the project the algorithm assigns to him/ her.

Student's distance = the total difference between the student's utility and those of all other students, given the projects assigned to everybody by the algorithm

For each algorithm, the table below shows the sum of all student's utilities (total utility) and the sum of students' distances for all students (total distance).

	A1	A2	A3	A4	A5
Total Utility	168	185	213	178	136
Total Distance	252	1454	994	1452	132


7
6
5
4
3
A1
A2
A3
A4
A5
Utility Distribution


70
60
50
40
30
20
10
A1
A2
A3
A4
A5
Distance Distribution

The following is an informal description of how each algorithm works:

Algorithm 1 (A1) minimises the total distance while guaranteeing at least 70% of the maximum possible total utility.

Algorithm 2 (A2) maximises the minimum individual student utility while guaranteeing at least 70% of the maximum possible total utility.

Algorithm 3 (A3) maximises total utility.

Algorithm 4 (A4) maximises the minimum individual student utility.

Algorithm 5 (A5): minimises total distance.

Given the explanation of how the algorithms work and the allocations computed by each algorithm, which of them would you prefer most, and which would you prefer least? You can list more than one algorithm in each line.

Most Preferred Algorithm(s): _____

Least Preferred Algorithm(s): _____

Please give reasons for your assessment:

_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____