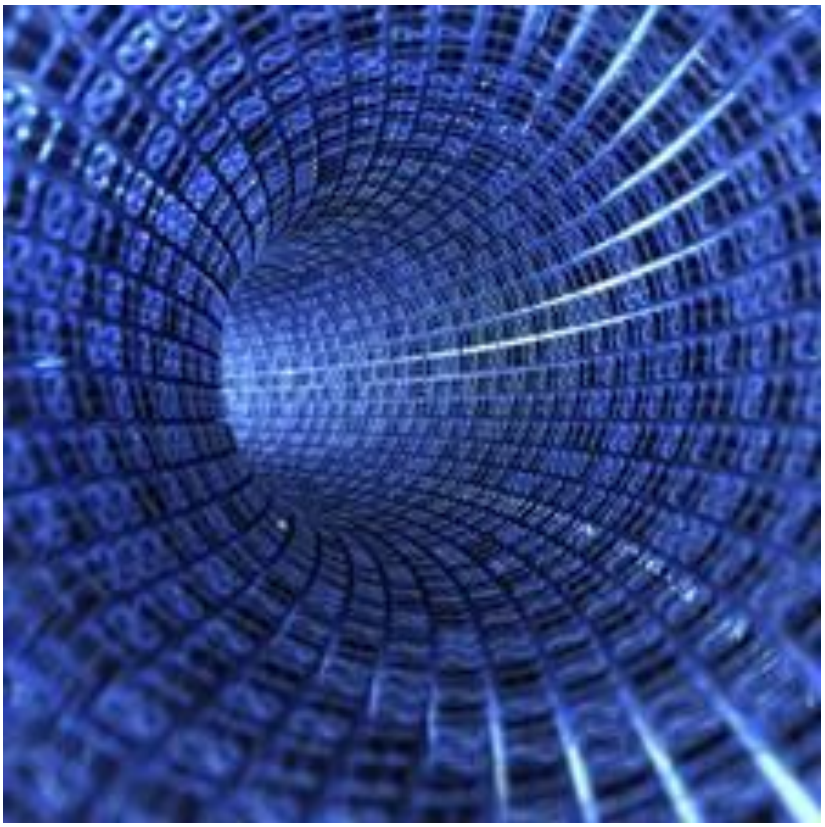
Summary report of outcomes: 2nd UnBias Stakeholder workshop



Report authors:
Ansgar Koene, University of Nottingham
Helena Webb, University of Oxford
Menisha Patel, University of Oxford

Workshop date: 19 June 2017

Report publication: 27 November 2017

UNBIAS
EMANCIPATING USERS AGAINST ALGORITHMIC BIASES FOR A TRUSTED DIGITAL ECONOMY

Research Councils UK Digital Economy Transforming Business and Society

The University of Nottingham
UNITED KINGDOM · CHINA · MALAYSIA

UNIVERSITY OF OXFORD

Table of Contents

Executive summary

On June 19th 2017 the second UnBias project stakeholder engagement workshop took place at the Wellcome Collection, London, UK. The workshop brought together participants from academia, education, NGOs and enterprises in order discuss issues relevant to the project. Specifically, participants conducted a *limited resource allocation* task and explored ideas for an *empathy toolkit* to help online providers and other stakeholders to help them understand the concerns and rights of internet users.

The conduct of the **limited resource allocation** task required participants to select their preferred and least preferred algorithms in a specific scenario. The results revealed that when given the same scenario, participants express different preferences over algorithm selection. This highlights the difficulty faced by any system in selecting an algorithm that is globally approved by users. There do appear to be some similarities of preference amongst participants with similar (professional or educational) backgrounds but it is unclear what difference the provision of further information about algorithms (i.e. a degree of transparency) makes on participant preference.

When asked to explain their algorithm selections in discussion, participants consistently used the language of fairness. They invoked normative understandings of right and wrong to argue that the preferred algorithm should be the fairest one. They also related their preferences to the (real or imagined) context in which the algorithms were to be applied and frequently expressed a desire to have further information in order to make 'better' decisions.

In the **empathy toolkit discussion** participants were asked to debate questions over the nature of empathy and the potential format for a planned UnBias project empathy tool. This tool will be designed to help online providers and other stakeholders understand the concerns and rights of internet users. In the discussion participants commented that empathy can bring benefits when it leads to positive actions. In the context of technology, empathy can lead to actions that are supportive of users – such as improved design and attention to the long term consequences of innovation. However they also warned that empathy does not always lead to supportive action. Increasing the empathy of platforms towards their users might not lead to benefits; instead the need for platforms to generate profit (in particular through advertising revenue) could lead to empathy being used to manipulate rather than help users. There are various forms that our planned tool could take, but it would need to ensure that empathy, or other positive values, could be fostered without providing a pathway for the greater manipulation of users and user data.

The **next steps** for the UnBias stakeholder engagement work-package are to run further workshops to present and discuss our emerging project findings and gain feedback on the resources we are preparing for different audiences.

The report of the first workshop can be accessed here
http://unbias.wp.horizon.ac.uk/2017/04/28/report-of-1st-wp4-workshop-report/

Part 1: Limited Resource Allocation task

In the first half of the workshop participants undertook a limited resource allocation task. They were presented with a (real) scenario in which university students were to be allocated coursework topics. Each topic could only be allocated once and all students had previously graded how happy they would be – from a scale of 1 to 7 – if each topic were to be allocated to them. Participants were asked to complete a questionnaire which presented them with 5 algorithms to achieve the distribution of coursework topics. In this, participants were asked to select their preferred and least preferred algorithm for this task. The questionnaire was in two parts: the first part presented them only with the results of the algorithm – indicating how many students received a coursework topic as 1, 2 etc.- and the second part included a brief description of each algorithm. Participants completed the questionnaires alone and then discussed them as a group. The questionnaires are provided in Annex 1 and Annex 2 of this report, respectively.

This task was devised by the UnBias project team members at the University of Edinburgh and we have carried it out on a number of other occasions with students from computer science and social science/law backgrounds. We have collated the results from the various times the task has been undertaken and analysed them quantitatively and qualitatively. Here we report on or cumulative findings.

1) Quantitative analysis

The quantitative analysis of the questionnaires did not yield statistically significant results, but nevertheless we can make the following observations:

- Participants expressed different preferences over their least and most preferred algorithms. Algorithm 3 was the most popular choice in each part of the task but in both cases almost half the preferences were split amongst algorithms other than algorithm 3 (see Annex 3, yellow bars). This result clearly supports the hypothesis that, when given the same scenario, participants may express different preferences over algorithm selection. This highlights the difficulty faced by any system in selecting an algorithm that is globally approved by users. In part 2 we observe an increase in preferences for algorithms that offer a trade-off between multiple criteria (e.g. utility and distance) over those that optimise a single criterion
- The questionnaire was split into two parts in order to assess whether the provision of extra information had an impact on preference selection. Change of preference across the two parts did occur: around 30% of the algorithms selected as the most and least preferred in the first part of the questionnaire were not marked as such in the second part (see Annex 3, blue bars). Moreover, we observe a complete change of selection by eight participants - out of 39 in total - especially regarding their least preferred algorithms. However, it is unclear whether the change in preference selection was prompted by the further information provided in part 2 of the questionnaire or by other factors such as the discussion with other participants that took place at the end of part 1 or personal reflection by participants between the two parts of the activity.
- There were some relationships between participants' responses and their (professional or educational) background. The most homogenous preference selections came from the computer science student group whereas the stakeholder group was more diverse in its selections.

2) Qualitative analysis

We analysed transcripts of the discussions that took place amongst each group taking part in the task. We observe that:

- The questionnaire asked participants to select algorithms in terms of preference, but when asked to explain their selections in discussion they consistently used the language of fairness. They invoked normative understandings of right and wrong to argue that the preferred algorithm should be the fairest one. For instance, from the stakeholder group: "*A3 seems the strongest one but the algorithm that produces these results seems to kind of result in unfairness because it's …very simplistic."* Opinions about which algorithm was fairest differed, as did ideas about what actually constituted fairness. For instance, another quote from the stakeholder group: "*What's more important. Is it, first, equality? Second lifting people that are at the bottom?"* Participants frequently attended to the difficulty or even impossibility of a single algorithm producing a fair result in all cases.
- Participants consistently related their preferences to the (real or imagined) context in which the algorithms were to be applied. They drew on the given context to support their preferences and also expressed the need for further information - about the demographics of the students, the nature of the course etc. - in order to aid their decision making. For instance, some felt they might make a different selection dependent on whether the students risked falling their entire course if they did not do well and others wanted to check if the algorithms might be systematically discriminating against, for example, female students.
- Across the different groups participants displayed varying levels of familiarity with technical features of algorithms.

The aim of this task was to assess the potential for a single preferred algorithm to be agreed upon by users. Our results demonstrate that this is unlikely as it is apparent that even when provided with the same information, participants make different preference selections and rationalise them differently. It may be possible however, via a larger study, to cluster together participants with particular characteristics to reach agreed preference on a sub-community level. Meanwhile, the importance placed by participants on context suggests that it would not be possible to determine a single preferred algorithm that could be applied across different scenarios. It does however indicate the perceived relevance of context to algorithm design.

Our findings resonate with the values that have come to the fore in current debates over algorithm prevalence -- in particular the values of fairness, transparency and explainability. We identify a community-level association of a preferred algorithm as being a fair algorithm. Competing models of fairness are drawn on in expressions of preference, once again undermining the potential to identify a single preferred, fair algorithm. However, 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 and 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.

Part 2: Empathy toolkit discussion

The second half of the workshop focused on the notion of empathy. A key aim of the UnBias project is to develop an empathy tool for online providers and other stakeholders to help them understand the concerns and rights of internet users. We asked our workshop participants to discuss the following questions in relation to the development of this tool:

- What's the difference between fairness and empathy?
- How can you encourage people to feel empathy for others?
- Who in particular could benefit from the UnBias empathy tool and when might they most benefit from using it? What are the potential advantages and disadvantages of this kind of tool?
- What might the UnBias empathy tool look like? (format, content, etc.)

Participants discussed first in small groups and then in a plenary session. These discussions were highly illuminating and raised a number of important points for the UnBias project to consider.

1) The nature of empathy

Empathy is a difficult concept to define. Our participants described it as a psychological process that is experienced and displayed by individuals in different ways. In the context of online platforms, empathy was described as including respect for users' attention and time, and respect between user and platform.

Participants noted differences between empathy and the more social concept of fairness. Although closely connected, there can be tensions between the two. When we understand someone's perspective very well, we might be drawn to accommodate them at the exclusion of others. So feeling empathy might lead us to unfairly prioritise certain groups or individuals.

2) The value of empathy for technology and design

Empathy can bring benefits when it leads to positive actions. In the context of technology, empathy can lead to actions that are supportive of users – such as improved design and attention to the long term consequences of innovation. One participant stated:

… often when we're doing things, we're thinking short to medium term. It's very difficult …to design very long-term outcomes and often there's not a lot of incentive to do that. So empathy could be … a very important way of perhaps triggering the need for longer term perspectives

It was also noted that empathy would form a valuable inclusion in the teaching of digital ethics and that it might be fostered naturally and positively in scenarios where trust in platforms already exists

3) The downsides/risks of empathy

One of the most important insights arising from the discussion was that empathy is not always positive and does not always lead to supportive action. There was much scepticism that increasing the empathy of platforms towards their users would lead to benefits. Instead, it was felt that the need for platforms to generate profit (in particular through advertising revenue) would lead to empathy being used to manipulate rather than help users. As one participant said:

… when [the platform's] business model is to extract every single thing they possibly can from us, for their own financial gain and benefit, how likely is it, or even reasonable that we can encourage them to empathise with their users, who essentially are just a piggy-bank to raid.

Participants also pointed to recent controversies over Facebook's contagion study and their targeted advertising towards teenagers that their personalisation mechanisms have identified as emotionally vulnerable as instances in which empathy had been acted on in a negative way. In addition, comparisons were drawn with certain perspectives from robot ethics which state that robot technologies should not be developed to look too similar to humans. This logic states that robots that appear 'falsely' human might elicit strong emotional attachments from users and this might lead them to assume that the robots have certain (human) capabilities that they do not actually have.

4) Ideas for empathy tools

With the warnings regarding the potential harms of empathy in mind, participant suggestions for our empathy tool focused on ways to ensure that empathy, or other positive values, could be fostered without providing a pathway for the greater manipulation of users and user data. Ideas included:

- A technical tool that allows you to see – through visualisations or aggregated, anonymised data – the online experiences of others to help users understand different perspectives.
- A game that allows users to try out different forms of code and see their consequences, in order to understand how processes affect different groups of people.
- A 'red flag' system for designers to alert them to the possibility that they might be (unintentionally) discriminating against certain groups
- A set of corporate empathy guidelines– similar to the current use of corporate responsibility guidelines.
- The multidisciplinary training of designers, in particular with a focus on helping them to understand the diverse perspectives of different population demographics

Our participants emphasised that an empathy tool would be about humanising data, so that designers think about human users rather than numbers. They also emphasised that to be effective, empathy would need to be accompanied by transparency from platforms. From another perspective, we know that trust and reputation is of importance to platform owners, developers and policy makers. Therefore, our tool could help these groups to realise that empathy can enhance user trust and platform reputation – this way they will more likely see the relevance of the tool. Participants also warned against empathy becoming a mere 'tick box' exercise that involves no genuine reflection. Finally, a number of participants concluded that it is not the platforms or algorithms themselves that need to develop empathy, but rather the developers and designers behind them.

[W]e don't necessarily want empathic algorithms, but we want to increase the empathy of the people designing the algorithms to think about the complexity of the people whose behaviours are going to be acted upon by them, rather than saying, oh you know, this is a series of nice little sort of boxes, we can typify people, stereotype people into…So perhaps it's measures of success might be about how much can we move away from personas, scenarios, stereotypes and more to appreciate just the extraordinary complexity of people on all kinds of spectrums.

Annex 1: Questionnaire part 1

UnBias: 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 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).

Table temporarily removed pending forthcoming academic publication

Images temporarily removed pending forthcoming academic publications

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:

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

Annex 2: Questionnaire part 2

UnBias: 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 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).

Table temporarily removed pending forthcoming academic publication

Images temporarily removed pending forthcoming academic publication

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:

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

Annex 3: Participant choices of most preferred Algorithm

Image temporarily removed pending forthcoming academic publication