Algorithmic Fairness in Online Information Mediating Systems

Ansgar Koene, Elvira Perez
University of Nottingham
{ansgar.koene,elvira.perez}@nottingham.ac.uk

Sofia Ceppi, Michael Rovatsos
University of Edinburgh
{sceppi,mrovatso}@inf.ed.ac.uk

Helena Webb, Menisha Patel, Marina Jirotka
University of Oxford
{helena.webb,menisha.patel,marina.jirotka}@cs.ox.ac.uk

Giles Lane
Proboscis
giles@proboscis.org.uk

ABSTRACT

This paper explores the challenges around fair information access when the limits of human attention require algorithmic assistance for 'finding the diamond in the coal mountain'. While often demanded by users, the seemingly intuitive concept of fairness has proven to be very difficult to operationalise for implementation in algorithms. Here we present two pilot studies aimed at getting a better understanding of the conceptualisation of algorithmic fairness by users. The first was a multi-stakeholder focus-group discussion, the second a user experiment/questionnaire. Based on our data we arrive at a picture of fairness that is highly dependent on context and informedness of users, and possibly inherently misleading due to the implied projecting of human intentions onto an algorithmic process.

ACM Reference format:
Ansgar Koene, Elvira Perez, Sofia Ceppi, Michael Rovatsos, Helena Webb, Menisha Patel, Marina Jirotka, and Giles Lane. 2017. Algorithmic Fairness in Online Information Mediating Systems. In *Proceedings of ACM Web Science Conference, Troy, NY, USA, June 2017 (WebSci'17),* 2 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Driven by competition for greater efficiency, online services are increasingly optimizing the user experience (as measured by attention capture) using Big Data algorithms that process massive-scale multi-dimensional data sets. The complexity of these algorithms is prompting concerns among service users, professional organizations and regulators [1, 3] regarding the breakdown of people's ability to (intuitively) grasp the general principles – let alone the more detailed elements – of these algorithms. In recognition of the potential for (inadvertent) abuse by means of algorithmic manipulation, computer scientists (e.g. [2]) are exploring how to design fair algorithms or verify the fairness of existing algorithms. From a computational perspective, there are many competing ways to operationalise the algorithm fairness. One might, for instance, prioritizes individual fairness over group fairness , "equality of outcomes" over "equality of treatment", etc. As confirmed by our results, this ambiguity reflects the more general problem that the concept of fairness, which initially appears to be very intuitive based on people's experience of human-human interactions, quickly reveals itself to be highly context-dependent when applied as a system design guideline. Even when referring to the same algorithmic system, two people discussing its fairness may be referring to different criteria depending on perspectives such as the ones mentioned above. In order to gain a better understanding of the multi-faceted experiences of algorithmic system fairness/unfairness by a range of user groups, the UnBias project (http://unbias.wp.horizon.ac.uk/) is running a series of studies comprising of "Youth Juries" deliberations with young digital natives, user observation studies and stakeholder engagement workshops. In this paper we present results from our first stakeholder engagement workshop and a questionnaire/discussion with computer science students.

WebSci'17, Troy, NY, USA
 978-x-xxxx-xxxx-x/YY/MM…$15.00
DOI: 10.1145/nnnnnnn.nnnnnnn

2 MULTI-STAKEHOLDER ENGAGEMENT

The 1st UnBias stakeholder workshop on "bias considerations of information mediating algorithms" took place on February 3rd 2017. The 26 participants represented four Small-Medium-Enterprises, three innovation incubators, two consultants, ten academics, five not-for-profit civil-society organizations, and two teachers. Prior to the workshop, participants were sent a questionnaire eliciting their experiences as a user and/or developer of algorithm-driven information systems. Included in the questions was a 'working definition' of fairness and questions regarding the usefulness of this definition and how they would propose to change it.

The 'working definition' defined fairness as "a context-dependent evaluation of the algorithm processes and/or outcomes against socio-cultural values. Typical examples might include evaluating: the disparity between best and worst outcomes; the sum-total of outcomes; worst-case scenarios; everyone is treated/processed equally without prejudice or advantage due to task-irrelevant factors".

Most of the stakeholders rated this definition as "good" or a "reasonable starting point". When asked how to change and improve the definition, the following points were highlighted:

Criteria relating to social norms and values: *(i)* Sometimes disparate outcome are acceptable if based on individual lifestyle choices over which people have control; *(ii)* Ethical precautions are more important than higher accuracy. *(iii)* There needs to be a balancing of individual values and socio-cultural values. Problem: How to weigh relevant social-cultural value?

Criteria relating to system reliability *(i)* Results must be balanced with due regard for trustworthiness. *(ii)* There needs to be independent system evaluation and system monitoring over time.

Criteria relating to (non-)interference with user control/ agency: *(i)* Subjective experience of fairness depends on the user's objectives at the time of use and therefore requires an ability to

tune the data and algorithm. *(ii)* Individuals should be able to limit the data collection about them and its use. Inferred personal data are still personal data. Any meaning that is assigned to the data must be clearly justified towards the user. *(iii)* It must be possible to demonstrate and explain the reasoning and behaviour of the algorithm in a way that can be understood by the data subject. *(iv)* If the algorithm is not indispensable to the task, it should be possible to opt-out of the algorithm but still use the other components of the service. *(v)* Users must have freedom to explore algorithm effects, even if this would increase the ability to figame the systemfi. *(vi)* There need to be clear means of appeal/redress for impact of the algorithmic system that the user cannot control.

In a dissenting opinion one participant rated the original working definition of fairness as "way off", stating that:

I struggle with the underlying antrhopomorphisation of algorithms when one speaks of fialgorithm fairnessfi. In my view, the concept of fairness is as you rightly noted deeply enshrined in the specific socio-cultural codes of the respective group of actors. Since algorithms as such constitute only the tools that actors in the social context use to achieve (some of) their objectives , one should also judge fairness probably more along the behaviour of these actors, their objectives, and methods. This implies that both the process of the algorithm and its outcomes need to be taken into account.

The 'working definition' of fairness and the suggestions provided by the stakeholders focused on scenarios like search engines and news recommendation systems, in which algorithms can optimize the outcome for each user independently. In the next section/study we considered cases where such independently optimized outcomes are not possible.

3 COLLECTIVE DEFINITION OF FAIRNESS

Many multi-user scenarios have a combinatorial nature, i.e., decisions applied to one user also affect other users such that decision optimization must consider the requirements and preferences of all users at the same time. Examples of such scenarios are sharing economy applications, where users seek peers to form teams, and situations in which resources are limited and/or variable in type, e.g. hotel rooms. Consider for example ride-sharing. Assume there are four users, A, B, C, and D, each car can transport at most two users, and users have preferences for peers. In particular, assume A prefers to travel with B over D over C, B prefers to travel with C, C with B, and D is indifferent. If an algorithm tries to solve this problem by considering each user individually, it would suggest to A to travel with B, violating the wish of B to travel with C. By considering the collective of users at once, the algorithm could propose a feasible solution in which A travels with D and B with C.

Generally, such scenarios are characterized by collectives of users with conflicting preferences and constraints (due to the combinatorial nature of the problem) that limits feasible solutions. Note that users may have conflicting ideas also about the fairness concept to use. To explore this fairness selection problem we conducted a pilot-experiment aimed to (i) explore problems related to identifying a collectively approved definition of fairness, and thus the most appropriate algorithm to use in a constrained collective scenario, and (ii) observe how transparency of the values embedded in an algorithm affects users' judgment.

We considered the problem of allocating coursework topics to students when each student must have exactly one topic and each topic can be assigned to at most one student. Each student expressed his/her preferences over each topic by assigning them a score from 1 to 7 representing the desirability of that topic to the student.

To identify possible student-topic assignments we used five algorithms differing in the adopted fairness concept as defined by the trade-off between maximizing the aggregated level of preference-matching summed over all students vs minimizing the preference-match differences between students.

We gave a two part questionnaire to students with graphs showing the preference score each student gave to their assigned topic and the difference between the level of preference-matching reached for every student. In the first part of the questionnaire, students had chose the most and least preferred algorithms only on the basis of above mentioned graphs, while in the second part, they were also given the description of the algorithms. The two parts of the questionnaire were sequentially presented in order to observe how transparency of the algorithms' values affected the outcomes.

The results are summarized by three observations. First, the outcome confirms the importance of transparency; indeed there was a drastic change in algorithm selections between the first and the second part of the questionnaire. Second, since students' preferences of least and most preferred algorithms were spread across several options, we conclude that users judge algorithms using different criteria even when experiencing exactly the same scenario. Finally, this pilot experiment proves that for constrained collective scenarios it is impossible to guarantee that, independently and without the support of a mediating agent, users can agree on an algorithm to use. The different fairness criteria users adopt can conflict with each other making it impossible for an algorithm to guarantee them simultaneously.

4 CONCLUSION

Combining the results from both our studies we conclude that there is no unique, globally approved, definition of fairness. Several crucial characteristics of fair algorithms can be highlighted that are, however, not always simultaneously achievable.

Our findings show the need to extend the study on (i) fairness concepts, in order to coherently integrate the stakeholders' suggestions on defining fairness and highlight their conflicts, and (ii) the design of techniques that can be adopted to (partially) coordinate the heterogeneous fairness preferences of groups of users.

5 ACKNOWLEDGEMENT

This work forms part of the UnBias project supported by EPSRC grant EP/N02785X/1.

REFERENCES

[1] IEEE 2016. *Ethically Aligned Design. A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems. Version 1.* IEEE. http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html.

[2] Suresh Venkatasubramanian Sorelle A. Friedler, Carlos Scheidegger. On the (im)possibility of fairness. In *arXiv:1609.07236.*

[3] U.S. Executive Office of the President 2016. *Big Data: A Report on Algorithmic Systems, Opportunities, and Civil Rights.* U.S. Executive Office of the President. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf.