Algorithm Fairness in Online Information Systems

Ansgar Koene[1], Sofia Ceppi[2], Helena Webb[3], Menisha Patel[3], Elvira Perez[1], Giles Lane[4], Marina Jirotka[3], Michael Rovatsos[2]

(1) University of Nottingham, (2) University of Edinburgh, (3) University of Oxford, (4) Proboscis

Introduction to UnBias

UnBias investigates the user experience of algorithm driven services and the processes of algorithm design to provide policy recommendations, ethical guidelines and a stakeholder co-produced 'fairness toolkit'. We focus on a wide range of stakeholders and carry out activities that

1) support user understanding of algorithm mediated services
2) raise awareness among 'smart' systems providers about the concerns and rights of users
3) generate debate about 'fair' operation of algorithms in modern life.

Multi-Stakeholder Engagement

Engaging with academics, civil-society, industry, regulators:

- What constitutes a fair algorithm?
- What kinds of (legal and ethical) responsibilities do internet companies have to ensure fair and unbiased algorithmic decisions?
- What might enhance user awareness/trust in online algorithms?
- How might concepts of fairness be built into algorithmic design?

The discussion was structured around four case studies:

1) gaming the system – anti-Semitic autocomplete and search results;
2) news recommendation and fake news;
3) personalisation algorithms;
4) algorithmic transparency.

How to bump Holocaust deniers off Google's top spot? Pay Google

Google 'is unhappy' with Holocaust denial beating the truth in its search results – but it probably makes more money that way

Prisoners being transported to Auschwitz, 1944. Photograph: Universal History Archive/UIG via Getty Images


Facebook fires trending team, and algorithm without humans goes crazy

Module pushes out false story about Fox's Megyn Kelly, offensive Ann Coulter headline and a story link about a man masturbating with a McDonald's sandwich

I worked on Facebook's Trending team – the most toxic work experience of my life

The trending module was meant to have 'learned' from the human editors' curation decisions. Photograph: Bloomberg/Bloomberg via Getty Images

Defining Algorithm Fairness

Candidate definition: *"A context-dependent evaluation of the algorithm and/or outcomes against socio-cultural values"*.
E.g.: equal treatment without prejudice due to *task-irrelevant* factors.

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) Needs for independent system evaluation and monitoring over time.

Criteria relating to (non-)interference with user control:

(i) Subjective fairness experience depends on user objectives at time of use, therefore requires an ability to tune the data and algorithm.
(ii) Users should be able to limit data collection about them and its use. Inferred personal data is still personal data. Meaning assigned to the data must be justified towards the user.
(iii) Reasoning/behaviour of algorithm demonstrated/explained in a way that can be understood by the data subject.
(iv) If not vital to the task, there should be opt-out of the algorithm
(v) Users must have freedom to explore algorithm effects, even if this would increase the ability to "game the system"
(vi) Need for clear means of appeal/redress for impact of the algorithmic system that the user cannot control.

Dissenting opinion

"I struggle with the underlying antrhopomorphisation of algorithms when one speaks of "algorithm fairness". 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."

Collective Fairness

Many multi-user/limited 'resource' scenarios are characterised by 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.
E.g. problem of allocating course-work topics to students when each student must have exactly one coursework topic and each topic can be assigned to at most one student.

- Each student expressed preferences over each topic by assigning them a score from 1 to 7 representing how desirable that topic is.
- 5 Algorithms used to determine 'optimal' course-work assignment using different fairness criteria:
  - ❖ A1: minimise total distance while guaranteeing at least 70% of maximum possible utility
  - ❖ A2: maximise the minimum individual student utility while guaranteeing at least 70% of maximum possible total utility
  - ❖ A3: maximise total utility
  - ❖ A4: maximise the minimum individual student utility
  - ❖ A5: minimise total distance

Where: utility = score the student gave to the assigned topic
distance = sum of absolute difference between student's utility and utility for each other student

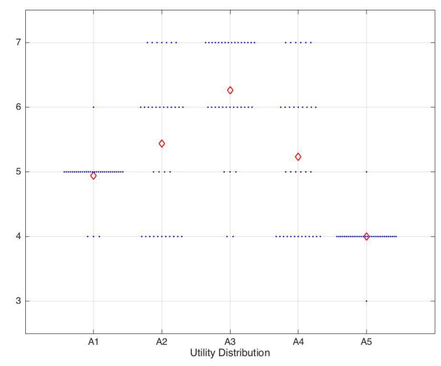

Figure 1: Student's utility distribution (blue dots) and mean (red diamond)

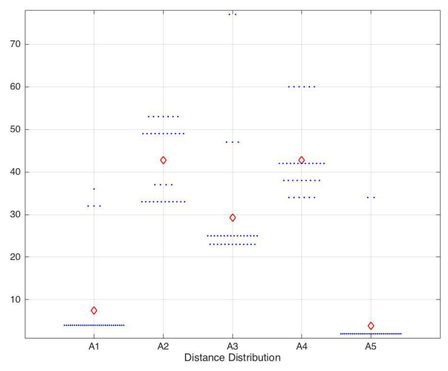

Figure 2: Student's distance distribution (blue dots) and mean (red diamond)

We asked the student to rank their preference for each algorithm:

1. on the basis of knowing only the topic assignment outcomes
2. knowing the fairness criteria used by each algorithm

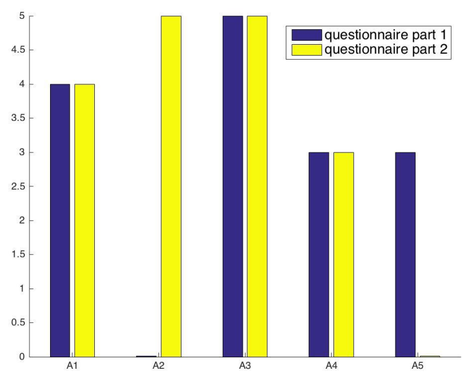

Figure 3: Algorithms most preferred by the students when they are not aware of the values embedded in the algorithms (blue) and when a description of the algorithms is provided (yellow) to them. Note that students were allowed to select more than one algorithm.

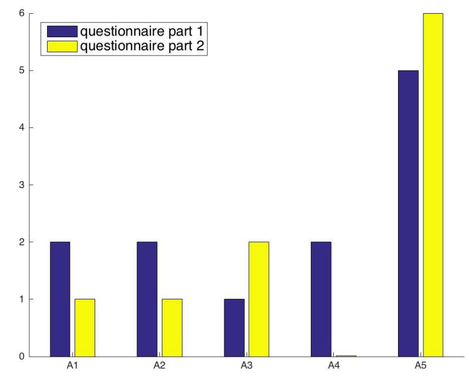

Figure 4: Algorithms least preferred by the students when they are not aware of the values embedded in the algorithm (blue) and when a description of the algorithms is provided (yellow) to them. Note that students were allowed to select more than one algorithm.

Individual students preferred different algorithms.
Preferences differed with/without knowledge of optimisation criteria.

Research Councils UK Digital Economy – Transforming Business and Society
EPSRC Engineering and Physical Sciences Research Council
horizon DIGITAL ECONOMY RESEARCH
http://unbias.wp.horizon.ac.uk/
The University of Edinburgh
University of Oxford
The University of Nottingham UNITED KINGDOM · CHINA · MALAYSIA