UNBIAS

EMANCIPATING USERS AGAINST ALGORITHMIC
BIASES FOR A TRUSTED DIGITAL ECONOMY

UnBias project key findings
Presented at our showcase event
1st October 2018 at the Digital Catapult, London

Research Councils UK Digital Economy Transforming Business and Society

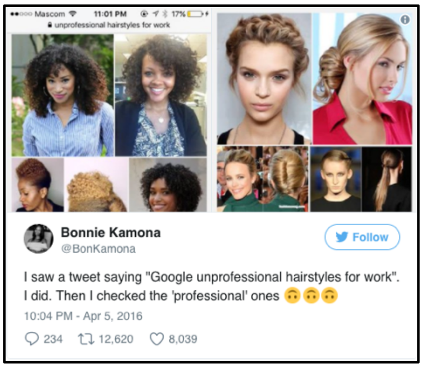
The University of Nottingham
UNITED KINGDOM · CHINA · MALAYSIA

UNIVERSITY OF OXFORD

THE UNIVERSITY OF EDINBURGH

What is the UnBias project about?

The UnBias project seeks to promote fairness online. We live in an age of ubiquitous online data collection, analysis and processing. News feeds, search engine results and product recommendations increasingly use personalisation algorithms to determine the information we see when browsing online. Whilst this can help us to cut through the mountains of available information and find those bits that are most relevant to us, how can we be sure that they are operating in our best interests? Are algorithms ever 'neutral' and how can we judge the trustworthiness and fairness of systems that heavily rely on algorithms?

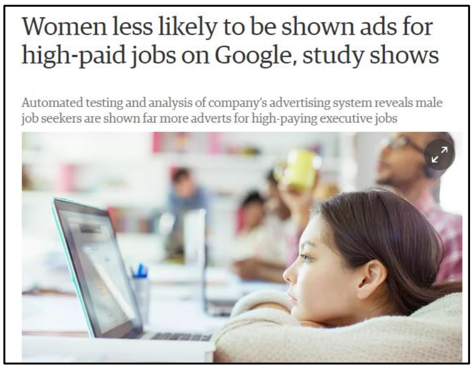


Bonnie Kamona
@BonKamona

I saw a tweet saying "Google unprofessional hairstyles for work". I did. Then I checked the 'professional' ones 🙃🙃🙃

10:04 PM - Apr 5, 2016

234 12,620 8,039



Women less likely to be shown ads for high-paid jobs on Google, study shows

Automated testing and analysis of company's advertising system reveals male job seekers are shown far more adverts for high-paying executive jobs

Our project investigates the user experience of algorithm driven internet services and the processes of algorithm design. We are particularly interested in circumstances in which algorithmic processes might (intentionally or unintentionally) produce biased or unfair outcomes – for instance in the form of helping fake content to spread on social media, producing search results that reinforce prejudiced attitudes, or the excessive personalisation of content and collection of personal data.

> **Mission:** To develop co-designed recommendations and materials for design, regulation and education to promote 'fairness' in algorithmic systems.

We focus in particular on the perspectives of different stakeholders and carry out activities that: 1) support user understanding about online environments, 2) raise awareness among online providers about the concerns and rights of internet users, and 3) generate debate about the 'fair' operation of algorithms in modern life.

UnBias is a collaboration between:
- HORIZON Digital Economy Research institute at the University of Nottingham
- Human Centred Computing at the University of Oxford
- Centre for Intelligent Systems and their Applications (CISA) at the University of Edinburgh.

We also work with creative studio Proboscis.

The project is funded under the Trust, Identity, Privacy and Security (TIPS) programme of the UK's Engineering and Physical Sciences Research Council (EPSRC). Project reference EP/N02785X/1

The project consists of four interconnected Work Packages (WPs):

WP1 [led by Nottingham] uses 'Youth Juries' workshops with 13-17 year-old "digital natives" to co-produce citizen education materials on properties of information filtering/recommendation algorithms;

WP2 [led by Edinburgh] uses co-design workshops and Hackathons to explore challenges in the algorithmic system design process related to bias/fairness of algorithmic decisions, e.g. trade-offs in limited resource allocation tasks;

WP3 [led by Oxford] explores users' experiences of algorithm driven online platforms via observations and other fieldwork methods;

WP4 [co-led by Oxford and Nottingham] explores policy dimensions, including information and education governance frameworks for responding to the demands of the changing landscape in the usage of algorithmic decision and recommendation systems. These are developed through broad stakeholder focus groups including representatives of industry, regulators, third-sector organizations, educators, lay-people and young people (a.k.a. "digital natives").

Project team

Nottingham
Professor Derek McAuley
Dr Elvira Perez Vallejos
Dr Ansgar Koene
Dr Liz Dowthwaite
Dr Virginia Portillo
Dr Helen Creswick
Monica Cano

Oxford
Professor Marina Jirotka
Dr Helena Webb
Dr Menisha Patel

Edinburgh
Professor Michael Rovatsos
Dr Alan Davoust
Dr Sofia Ceppi

Proboscis
Giles Lane

Key findings

As the UnBias project has developed we have enjoyed a great deal of collaboration and cooperation across the workpackages. We have found various linkages between our areas of work and our results and outputs complement each other. Across the project, we have identified 3 key findings

1. The existence of concerns over the contemporary prevalence algorithm driven online platforms
2. A desire for change to improve the user experience of these platforms
3. Opportunities for change – we have identified capacity to address unfairness in algorithmic systems through various means, in particular via education, engagement and policy change.

The existence of concerns over the contemporary prevalence algorithm driven online platforms

In our youth jury sessions conducted with young people, many participants recognised the benefits of using personalisation and filtering processes, believing them to be convenient by making their internet searches quicker, and useful in showing information that was new and interesting to them. However, many also expressed concerns that these processes can be inaccurate and often miss the nuances in their searches. They also expressed worries about filter bubbles that limit the amount of other information that they might see when online.

> Youth jury participants
>
> *"I watch quite a lot on Netflix at night, and they recommend you films that have like the same actor in it. And it's like you're not watching the film because of the actor. You're watching the film because of the film."*
>
> *"I don't really know because I feel like sometimes there might be important stuff in the world going on that might be filtered out for me for some reason when actually I might be quite interested in it, so yeah."*
>
> *"If an algorithm is working in a way that if you Googled 'Trump' and then it only gave you negative or positive articles, then that's wrong I think."*
>
> *"It seems like they're controlling you and they already know what you're going to do, so you follow what they show you."*

Some youth jurors felt that the level of personalisation that they were comfortable with depended on the context, for example music and entertainment recommendations were seen as useful, however the personalisation of newsfeeds was met with concern. Some also expressed resignation and powerlessness that they could do little to change or influence the ways that the algorithms operate.

Similar kinds of concern were raised by the stakeholders on our panel. These professionals from industry, law, research, education and NGOs highlighted the ways in which the mechanisms through which individual users access content online can be problematic. For instance, tailored advertising can have very negative effects. They told us of the limitations of personalisation. It can be useful for advertising purposes and of course helps to keep free access to platforms. But the assumptions being made about users can be patronising and inaccurate. Furthermore, people can be fearful of invasions of privacy and their vulnerabilities being targeted.

Stakeholder participants

"So you, your Facebook and your friend's Facebook might be getting very different ads tailored to your usage, and was very very influential if we view the report from Cambridge Analytica in terms of putting our personalised adds through Facebook that swayed people's opinions and that was seen as content, so that was Fake News in that sense."

"It's very condescending isn't it. Some silicon-valley fundamentalist decided they know better what a person, that and that age, and that and that place should like … who are they?"

"I think that the main problem is that consumers expect the internet to tell them the truth. "I can find anything I want online. And I double check the news, because I can find out if they are telling me they are not the truth, so false news, fake news, whatever". So they think they are able to search for information properly, but maybe they aren't and they do not even know that they could be biased in their search behaviour".

Our stakeholders told us that lack of awareness of algorithmic processes such as search mechanisms can cause problems. For instance, where users are not aware of ranking processes operate they may have a false view of how far the top result in a search reflects reality – as opposed to being the result of a calculation that might be affected by a number of factors. Similarly, users may have an over high feeling of trust in what they see online in general – expecting it to be authoritative and truthful.

The actions of platforms in keeping these processes largely opaque and unclear was also highlighted as a problem. As was the lack of pluralism online. Our online experience is dominated by a few giant companies and we often have little option to go elsewhere; this kind of environment, according to some of our stakeholders, allows echo chambers to thrive.

Our stakeholders also highlighted the particular vulnerability of children online. As digital natives, they have a social imperative to use social media but lack protection. For instance, terms and conditions of platforms are deliberately not written in ways they can understand and they may not yet have developed the critical thinking skills to read between the lines of what they see online or understand why they are being shown particular adverts or being asked to input particular information about themselves.

A desire for change to improve the user experience of algorithm driven online platforms

Our youth panel participants were highly articulate in expressing a desire for change. Whilst they did at times express resignation and feelings of powerlessness they were also clear at indicating that they wanted change and stating what it could look like. This is illustrated in the quantitative findings of the youth jury sessions:

> 93.8% of youth jury participants came up with some ideas during the sessions about how the Internet could be made better for young people.

Additionally, they:

- wanted more influence on how digital technologies and services are run (65.6% agree, 10.3% neutral agree)
- wanted more control over their personal data online (84.3% agree, 3.5% neutral agree)
- felt that young people would be listened to about the internet they want (44.8% agree, 19.4% neutral agree)

However, they felt that they couldn't control what happens to them and their personal data when they use digital technologies (40.7% agree, 11.9% neutral agree)

Youth panel participants

"More control is definitely nice like for example my Facebook, whenever I open my account it's like full of junk I don't want to see! This algorithm basically fails me. I don't want to see like who is in love or something like that! It just fails me. I think to a great extent I agree that we need more control, so actually we can control how the algorithm works like disable those filter options".

"I think that an issue that needs to be raised in parliament or whatever is we should have more knowledge or accessibility of knowing who's got the information and why they have it and If we can manage that."

"I think they should have something like in schools and other projects outside of schools for other people as well, they should teach them because I know in some schools they do citizenship but there they speak more about moral issues that might not necessarily affect everybody, well, a couple in a class, but if they did do something on like terms and conditions and how the internet is used it would change."

Similarly, our professional stakeholders were clear that change is desirable and described different ways it could be achieved. They pointed to changes in areas around education, regulation and the practices of the platform themselves. They highlighted the importance of key principles: the transparency of algorithms, the accountability of platforms that use algorithmic systems, the auditability of the algorithms.

Our stakeholders were very keen that this was a complex matter and cannot be resolved in simple terms. We may need to make trade-offs: if we want to keep our free access to platforms we might have to give up some data and privacy, or take responsibility for our own behaviours and choices. However, they did also point out that we can learn from other sectors and bring in processes that have worked elsewhere.

Stakeholder participants

"I think having some sort of regulatory framework and some tests that all algorithms must be submitted to where…the algorithm can be assessed to see whether it represents a fair decision based on the data".

"…but as a society we need to get ready, we need to understand, we need to appreciate that in order to get stuff for our personalisation and diversity we might need to pay for it."

"So casinos, whether they are online or in the real world are obliged to monitor their customers to see if they exhibit the traits of addictive problem gamblers…So there is regulation on that and that might be a model that could potentially be applied to something like Facebook".

"So one possible solution was to change the legal framework and to think of platforms under the public service broadcast model in which they have to take responsibility for some of the content that is broadcast on their platforms. Or we could have things such as a kite mark for trusted content…"

Opportunities for change

In our project work we identified the existence of concerns about the operation of algorithm driven online platforms, and their capacity to introduce unjustified bias and produce unfair outcomes. We also identified a desire for change to address this unfairness. Our next step as a project was to consider opportunities for positive change to occur. What mechanisms might mitigate unfairness and promote a fairer user experience online? In this final section of our key findings report we characterise 3 important themes in relation to promoting positive change:

- any change must acknowledge and address the complexity of the issues involved
- value of education and engagement
- specific UnBias outputs and recommendations

Any change must acknowledge and address the complexity of the issues involved in algorithmic bias, 'fairness' and regulation. We need to recognise (as identified in our Workpackage 3 user experience observations) that users employ their own agency and sense making when browsing online. Algorithms do shape the user browsing experience but this is also filtered through users' own practical reasoning and is not straightforwardly deterministic. Methodologically it is also very difficult to trace exactly how users are affected by algorithmic processes and what consequences this has for individual and group behaviours.

Furthermore, we also need to recognise that the 'problems' of algorithmic bias and unfairness have multiple roots and causes, and that there are competing perspectives over how they can be resolved. It might be that we need to think on a case by case basis and make trade-offs to find workable solutions. As our participants remind us, these questions are grounded in wider societal issues.

Stakeholder participants

"We want our algorithms in a sense to follow higher values, moral values that we think are more important than giving an exact reflection of the world…algorithms are inherently politicised."

"Well if you want to have a discussion about fairness, you have to go into a causal analysis to try and work out your context and what actually is important. You cannot have a one size fits all rule."

Furthermore, they remind us that whilst the aim of the UnBias project is to promote fairness online, identifying exactly what fairness is, is a complex task We conducted two rounds of a survey[1] that presented participants with a scenario about the allocation of limited resources within a particular context. We also gave them a set of 5 algorithms that could distribute the limited resources and asked them to select their preferred and least preferred one. We wanted to test out whether we present algorithms to lay audiences in a way that they would find meaningful and be able to make an informed choice about.

We were able to achieve this. Our participants engaged enthusiastically with the task we gave them and displayed an understanding of the how the algorithms would operate in the context and the implications they might have. We also wanted to see whether our participants would select the same algorithms as most or least preferred: there was no consensus and selections of most and least preferred were spread across all 5 algorithms. All of our participants spoke of wanting to select an algorithm that was fair but drew on different understandings of what was fair when making their selections – sometimes it was about giving everyone more or less the same, sometimes it was about making the most people as possible happy, sometimes it was about finding ways to measure what people deserved and so on. It was much easier for our participants to articulate what was not fair than what was fair. Their selections were also highly context dependent with changes in selection being made when the scenario changed slightly. Some of our results suggested that participants' interpretations of what was fair in a context had systematic connections to the cultural background the participants were in – this is something we are keen to assess further.

Education and engagement are valuable mechanisms to prompt and promote change. This is something we have been observing and practising across the project. We have developed educational materials including a video animation to help users understand algorithmic processes and the effects of personal data collection etc. These have often been co created with end users – in particular young people – to draw on their understandings and perspectives, and meet their needs.

We have run activities with university students – such as hackathon competitions and group projects – that challenged them to (re)design fair social networks and build systems that can empower users online. We have taken part in various public engagement activities and created a set of tools and activities that help online users of all ages to increase their understanding and reflect on their own

[1] Round 2 of the survey can be seen here: https://oxford.onlinesurveys.ac.uk/unbias-preference-survey

perspectives regarding algorithmic bias and fairness. More details about all of our engagement and education materials are on our project website: http://unbias.wp.horizon.ac.uk/

The other great benefit of our engagement activities in the project is that they provide a ground-up approach to identifying specific opportunities for change. We have been able to elicit from our participants the forms of improvement they would like to see – for instance in relation to more accessible platform terms and conditions, the incorporation of regulatory mechanisms from other industries, and the requirement for meaningful transparency rather than 'just' making algorithms visible. Bringing together stakeholders from different sectors has been particularly useful in building up a nuanced perspective on what changes might be possible and feasible. These perspectives are then incorporated into our further project outputs.

We have produced a wide range of **project outputs**, including:

- **Contributions to sets of policy recommendations:** European Parliament Science Technology Options Assessment; UK ICO Age Appropriate Design Code;
- **Contributions to Standards:** Chairing development of IEEE P7003 Standard for Algorithmic Bias Considerations; Participation (through BSI) in work on Standards for Artificial Intelligence by ISO/IEC
- **24+ Public engagement engagement events**
- **Education materials:** Online Educational Resources; Fairness Toolkit; Video animation; Outreach sessions and events with schools.
- **Contribution to 7 Government enquiries:** Impact of social media and screen use on young people's health; House of Lords Artificial Intelligence committee; Algorithms in decision-making; Fake News; The Internet: to regulate of not to regulate?; Children and the Internet; Role and priorities of UKRI interim chair
- **Academic publications** in journals relating to the social sciences, ethics in computing, human factors in computing, social media, AI
- **Media engagement:** television and radio interviews, podcasts, online articles in The Conversation, mentions in print media
- **45+ Workshops and conferences**
- **8 Collaborations and follow on activities:** 5Rights (ICO Code); DefendDigitalMe (evaluation of educational apps); HumanRightsNetwork/RightToPrivacy.org (Telling Tales of Engagement videos); ThirdWolrdNetwork (eCommerce trade negotiations – algorithm briefings); ISOC-England (User Trust); IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (P7003 standards + ad hoc presentations); RevealingReality (ICO AgeAppropriateDesign Code); Impact Exploration Grant (UnBias Fairness Toolkit - Usage and Impact);

More details about all of these outputs can be found on our project website http://unbias.wp.horizon.ac.uk/ or by contacting helena.webb@cs.ox.ac.uk or ansgar.koene@nottingham.ac.uk

In the final part of this report we highlight two particular UnBias outputs.

1) A governance framework for algorithmic accountability and transparency

As part of our policy and stakeholder engagement work, members of the UnBias team are co-authoring[2] a Science Technology Options and Assessment report on algorithmic accountability and transparency for the European Parliament. In this report we provide an overview of possible legal, technical and operational frameworks for governing algorithmic accountability and transparency. Beyond providing an up-to-date and systematic review of current and potential governance frameworks for algorithmic accountability and transparency, the report aims to provide recommendations for an algorithmic impact assessment metric for assessing the degree of regulatory scrutiny of an algorithmic system that would be appropriate/necessary for a particular context of use.

The report is still being finalised but our recommendations will include:

Awareness raising: education, watchdogs and whistleblowers

- The provision of "algorithmic literacy" that teaches core concepts such as: computational thinking, the role of data and the importance of optimisation criteria.
- The introduction of standardised notification practices to communicate the type and degree of algorithmic processing involved in decisions.
- The provision of computational infrastructure and access to technical experts to support the data analysis and algorithm reverse engineering efforts of "algorithmic accountability journalists".
- Whistleblower protection (expanding the current EC proposal to include any violation of Human Rights) and protection against prosecution on grounds of breaching copyright or Terms of Service when doing so served the public interest.

Accountability in public sector use of algorithmic decision making

- Mandating of Algorithmic Impact Assessments as part of public sector use and procurement of algorithmic systems, involving:
    - Publication of public authority's definition of "algorithmic system".
    - Public disclosure of purpose, scope, intended use and associated policies/practices, self-assessment timeline/process and potential implementation timeline of the system.

[2] Full list of authors

Ansgar Koene,	University of Nottingham
Nozha Boujemaa,	Inria
Chris Clifton,	Purdue University
Yohko Hatada,	EMLS RI
Jacob LaViolette	University of Oxford
Caio Machado	University of Oxford
Menisha Patel,	University of Oxford
Rashida Richardson	AI Now Institute
Dillon Reisman	AI Now Institute
Helena Webb	University of Oxford
David Weinberger	Harvard University

    - Performing and publishing of self-assessment of the system with focus on inaccuracies, bias, harms to affected communities, and describes mitigation plans for potential impacts.
    - Publication of plan for meaningful, ongoing access to external researchers to review the system once it is deployed.
    - Public participation period
    - Publication of final Algorithmic Impact Assessment, once issues raise in public participation have been addressed.
    - Renewal of AIAs on a regular timeline
    - Opportunity for public to challenge failure to mitigate issues raised in the public participation period or foreseeable outcomes.

Regulatory oversight and Legal liability

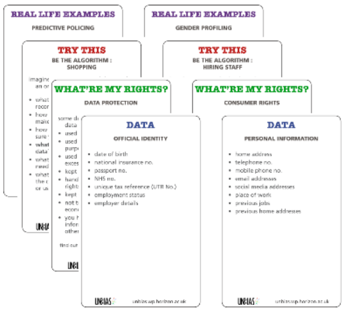
- The creation of a regulatory agency for algorithmic decision making tasked with:
    - Establishing a risk assessment matrix for classifying algorithm types and application domains according to potential for significant negative impact on citizens.
    - Investigating the use of algorithmic systems where there is a suspicion (e.g. evidence provided by a whistleblower) of infringement of Human Rights.
    - Advising other regulatory agencies regarding algorithmic systems as they apply to the remit of those agencies.
- Systems classified as causing potentially severe non-reversible impact are required to produce an Algorithmic Impact Assessment, similar to public sector applications.
- Systems with medium severity non-reversible impact require the service provider to accept strict tort liability, with a possibility of reducing the liability by having the system certified as compliant with (as yet to be determined) impact mitigating standards.

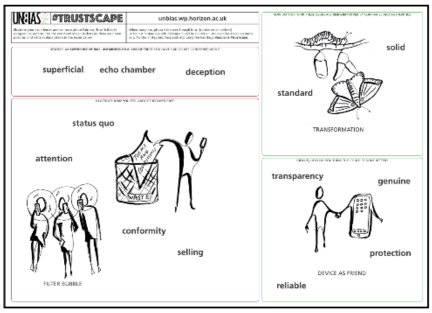
Global dimension of algorithmic governance

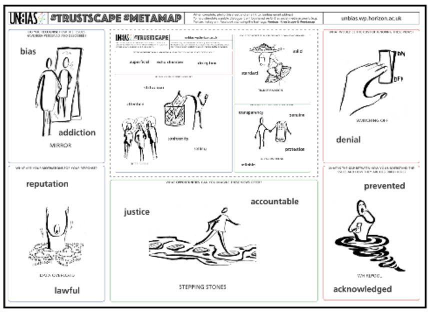
- A strong position in trade negotiations to protect regulatory ability to investigate algorithmic systems and hold parties accountable for violations of European laws and Human Rights.
- Establishing a permanent forum for multi-stakeholder dialog on the ethical development of algorithmic systems, based on the principles of Responsible Research and Innovation.

For more details about the upcoming EP STOA report, contact ansgar.koene@nottingham.ac.uk

2 Fairness Toolkit







Our Fairness Toolkit which aims to promote awareness and stimulate a public civic dialogue about how algorithms shape online experiences and to reflect on possible changes to address issues of online unfairness. The tools are not just for critical thinking, but for *civic thinking* – supporting a more collective approach to imagining the future in contrast to the individual atomising effect that such technologies often cause. The toolkit has been co-created with young people and stakeholders and consists of three main parts:

1. Awareness Cards – a deck of cards designed to help young people devise and explore scenarios that illustrate how bias in algorithmic systems can affect them. The cards are a "peer to peer" tool, enabling young people to collaboratively explore the issues of data privacy and protection, online safety and social justice to create compelling stories and scenarios that help communicate and develop awareness. The cards are designed to be used in both facilitated environments, such as schools and youth groups, but also as a civic thinking tool for anyone to investigate the relationships between data, rights, values, processes and the factors which affect their operation for us as individuals and as a society.

2. TrustScapes – a poster for young people to visualise their perceptions the issues of algorithmic bias, data protection and online safety and what they would like to see done to make the online world fair and trustworthy. Designed to capture both their feelings about the current situation and their dreams and ideals for what the internet could or should be in a dynamic and visual way. The TrustScapes form the first element in the public civic dialogue that UnBias will initiate: images of the TrustScapes will be shared via UnBias social media accounts to articulate young people's visions for the future internet and amplify their voice in the debate on trust and fairness.

3. MetaMaps – a poster for stakeholders in the ICT industry, policymaking, regulation, public sector and research to respond to the young people's TrustScapes. By selecting and incorporating a TrustScape from those shared online, stakeholders can respond to the young people's perceptions. MetaMaps will also be captured and shared online via UnBias social media to enhance the public civic dialogue, and demonstrate the value of participation to young people in having their voice listened and replied to.

The Fairness Toolkit also includes "value perception" worksheets to help participants and stakeholders critically assess and evaluate the value of participation.

For more information about the Toolkit, please contact helena.webb@cs.ox.ac.uk