Summary report of outcomes
1st UnBias Stakeholder workshop

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

Workshop date: 3 February 2017

Report publication: 30 April 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 February 3rd 2017 the first UnBias project stakeholder engagement workshop took place at the Digital Catapult centre in London, UK. The workshop brought together participants from academia, education, NGOs and enterprises to discuss fairness in relation to algorithmic practice and design. At the heart of the discussion were four case studies highlighting fake news, personalisation, gaming the system, and transparency.

On the issue of defining algorithm fairness, most stakeholders rated as a good or reasonable starting point: "*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*". To this, additional suggestions were made focusing on criteria relating to *social norms and values*, *system reliability*, and *the non-interference with user control/agency* (see page 7). When thinking of issues related to algorithm design, participant recommendations focused on *transparency* and *duty of care to society as well as target users/customers* (see page 9).

Fake news case study: regarding fake news, the focus of the discussion was on the nature of fake news, not a new phenomenon and lack of evidence of actual impacts, paired with a focus on education, critical reading skills, trustmark/branding and breaking the link with financial profit as main solutions. To the extent that algorithms might play a role, it was noted that market research is suggesting that people don't want personally tailored news (see page 10).

Personalisation algorithms case study: regarding personalisation, it was suggested that services marketed as personalisation should really be called *task based channelling* since the 'user type categorizations' don't really address personal goals. It was proposed that the use of personalisation can be useful for object and commercial purposes but that it is not appropriate when applied to dissemination of socially important information like news coverage. This led to discussion about levels of control that users should have (see page 10).

Gaming the system case study: the impact of gaming of systems like search rankings was closely linked to a general lack of awareness about how such rankings are determined and how to interpret the meaning of a ranking. It was highlighted that regulation regarding censorship focuses on removal of content, but not on placement within a ranking even though a very low ranking can often mean that something is effectively removed from peoples' awareness. In terms of solutions one idea that came up was the use of greater user input, through ratings feedback, to help signal the difference between search ranking and content validity (see page 10).

Algorithm transparency case study: discussion on his topic started from observations about need to clarify meaningful transparency (posting source code vs. understanding the process). This in turn requires clarity about the purpose of transparency, is it about fairness or trust? The importance of the data as integral part in determining algorithm bias was raised as well as the need to understand the users. Avoiding 'gaming the system' while still providing transparency was discussed with solutions focusing on intermediate auditing organisations and certification (see page 11).

Plenary discussion and conclusions: discussion of the four case studies raised a number of recurrent key points that were returned to in the plenary discussion that closed the workshop. Debate amongst the participants was very productive and highlighted that: 1) the 'problem' of potential bias and unfairness in algorithmic practice is broad in scope, has the potential to disproportionately

affect vulnerable users (such as children), and may be worsened by the current absence of effective regulation and market pluralism of online platforms; 2) the problem is also nuanced as the presence of algorithms on online platforms can be of great benefit to assist users achieve their goals; so it is important to avoid any implication that their role is always harmful; and 3) that finding solutions to the problem is highly complex. The effective regulation of algorithmic practice would appear to require accountability and responsibility on the part of platforms and other agencies combined with the meaningful transparency of the algorithms themselves.

The **next steps** for the UnBias stakeholder engagement work-package are to run further workshops. These will increasingly focus on issues of regulation and how it might be possible to identify practices to support algorithmic fairness that are both technically feasible and socially, ethically and legally valid. We will also run a parallel online questionnaire panel to seek the informed opinion of stakeholders who are unable to attend the workshops in person.

Introduction

This report summarises the outcomes of the first UnBias project stakeholder engagement workshop that was held at the Digital Catapult centre in London, UK, on February 3rd, 2017. The aim of these workshops is to bring together individuals from a range of professional backgrounds who are likely to have differing perspectives on issues of fairness in relation to algorithmic practices and algorithmic design. The workshops are opportunities to share perspectives and seek answers to key project questions such as:

- What constitutes a fair algorithm?
- What kinds of (legal and ethical) responsibilities do Internet companies have, to ensure their algorithms produce results that are fair and without bias?
- What factors might serve to enhance users' awareness of, and trust in, the role of algorithms in their online experience?
- How might concepts of fairness be built into algorithmic design?

For more information on the workshops, see also the call for participation in Annex 1.

Workshop reporting, procedure & programme

To facilitate an open discussion all stakeholder engagement activities are run under Chatham House rule – meaning that views expresses can be reported back elsewhere but that individual names and affiliations cannot, unless explicit consent is given. The outcomes described in this report combine the data obtained from the audio recordings and the notes made during the discussions. Prior to publication, the report was circulated for approval and/or amendment by all participants.

In preparation for the workshop the participants filled in a pre-workshop questionnaire [see Annex 2] on the topic of a possible definition of algorithm fairness, the participant's own experience with using and/or designing algorithm driven systems and their suggestions/concerns regarding the design of fair system. The purpose of the pre-workshop questionnaire was to obtain insights into individual perspectives prior to group interaction.

The workshop itself focused on four case studies [see Annex 3] concerning key current debates around algorithmic fairness and was run in two parts. The case studies relate to: 1) gaming the system – anti-Semitic autocomplete and search results; 2) news recommendation and fake news; 3) personalisation algorithms; 4) algorithmic transparency. In the first part of the workshop participants separated into four groups (of 5-7 participants), one group per case study. In second part the results of the separate case study discussions were presented and opened for a plenary discussion.

After the workshop participants were asked to fill in a post-event questionnaire [see Annex 4].

The workshop was attended by 15 participants (+ 4 facilitators) representing 3 SMEs, 2 Innovation hubs/charities, 1 consultancy, 11 academic institutions/groups, 5 NGOs/not-for-profits and 2 schools. A list of the participants who consented to be named is included in Annex 5.

Background

What is the UnBias project about?

The UnBias project [http://unbias.wp.horizon.ac.uk] 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?

Our project investigates the user experience of algorithm driven Internet services and the processes of algorithm design. We focus in particular on the perspectives of young people 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.

Stakeholder engagement activities

As part of our project we are running a series of stakeholder engagement activities, combining co-present workshops with online Delphi panel surveys. We invite stakeholders from academia, education, government/regulatory oversight organisations, civil society organisations, media, industry and entrepreneurs to join us in exploring the implications of algorithm-mediated interactions on online platforms especially in relation to access and dissemination of information to users. The activities will take place over a two-year period (2017 – 2018) and will seek to identify relevant perspectives and concerns as well as provide feedback on our project activities. They will also produce:

1. a set of policy and design recommendations for enhanced transparency and global fairness in information control algorithms;
2. a 'fairness toolkit' consisting of three co-designed tools
    i. a consciousness raising tool for young internet users to help them understand online environments;
    ii. an empowerment tool to help users navigate through online environments;
    iii. an empathy tool for online providers and other stakeholders to help them understand the concerns and rights of young internet users.

Responses to pre-workshop questionnaire: Algorithm fairness

The 'working definition of fairness' provided in the introduction of the pre-workshop questionnaire 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 working definition of fairness as either "good" or a "reasonable starting point". When asked how to change and improve such a definition, they highlighted the following interesting points:

Criteria relating to social norms and values

- Sometimes disparate outcomes are acceptable if based on individual choices to pursue lifestyles (lifestyle choices over which people have control).
- Ethical precautions are more important than higher accuracy.
- 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

- Results must be balanced with due regard for trustworthiness.
- There needs to be independent system evaluation and system monitoring over time

Criteria relating to (non-)interference with user control/agency

- 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.
- Individuals should be able to limit the data collection about them and its use. Inferred personal data is still personal data. Any meaning that is assigned to the data must be explicitly communicated to the user with justification for why this meaning is assigned.
- 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.
- If the algorithm is not indispensable to the task, the system should provide the ability to opt-out of the algorithm but still use the other components of the service.
- Users must have freedom to explore algorithm effects, even if this would increase the ability to "game the system".
- 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 "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.

Interestingly, the working definition of fairness and the alternative ones provided by the stakeholders can be applied to scenarios like that of search engines and news recommendation systems, in which, given a definition of fairness and a notion of optimal user outcome, the algorithm's goal is to provide the optimal and fairest outcome to each user independently. However, considering each user separately is not always an option. In this case, identifying a collectively approved definition of fairness is even more challenging, as we show in the next section.

The complete list of the recommendations proposed by the participants is provided in Annex 6

Responses to the question "*[t]hinking as an end user of a system where an algorithm is mediating the information you see; how would you describe the key properties that would define a fair algorithm*" fell into two main categories:

- Demands for transparency of decision criteria/process
- Demands that the algorithm provide a broad range of responses so as not to limit agency of user

and a couple of separate suggestions covering asking for:

- Treatment of minorities; not influenced by adverts/paying parties, or at least very explicit about this.; not influenced by previous user behaviour/location/time
- Provenance/trustworthiness of the source with regard to factual content is a key factor in hierarchisation of results displayed; it is unfair to provide unreliable content to users.
- Non-learning algorithms
- One which guards me according to subject matter rather than political preferences/views

The detailed list of responses to this question is provided in Annex 7

Likert scale rating responses to questions about current experiences with algorithm fairness are shown in Annex 8, averaging the results over all participants as compared to the average for the 'techy' and 'non-techy' participants.

Recommendations for the algorithm design process

Two of the questions in the questionnaire asked participants to specifically think about the issues of in terms of the algorithm design process.

> Question 7 asked: "*Thinking as a designer of an algorithm based information providing system, what would be the main issues to include in the design process?*"
>
> Question 8 asked: "*Still thinking as a designer, what would be the main issues to include in order to make sure the system is fair?*"

Due to the overlap in responses we both are summarised together here. The details response list is provided in Annex 9.

Transparency related recommendations:

- Making sure that experts can understand what is happening inside the 'black box', with an active system of expert audit for the most important algorithms.
- Public transparency in functionality and process: translating how it works to end users in a way that allows them to get a sense of what is going on, including insight into data which is collected (or purchased) and used in curation for individual users
- Transparent clarity about the balance of risk/benefit trade-offs affecting different groups, with special attention to 'edge cases'.

Duty of care to society as well as target users/customers:

- Measure disparate impacts on what perspectives appear and how they differ by groups and context.
- Devise spectrums of harm for different contexts.
- 'Sensitive information' must be included in training/testing data to enable detecting discriminating behaviour.
- Traceability of data and features to allow identifying thresholds that trigger changes in the outcome of the decision process.
- The need for user testing and diverse forms of evaluation including oversight of data coding, cleaning and collection.
- Give users partial control over the algorithm, e.g. the attributed that the determining is based on, to allow for subjective differences in perception of fair/biased performance.
- Ensuring the algorithm is testable against contextually appropriate standards of fairness, including assessment of the algorithm's resilience to changes in socio-cultural values underpinning fairness.
- Interaction between the algorithm with external factors, social & technical systems and data.
- Results should aim for relevance to the user, not general popularity or link to advertising
- News/reference searches must not be shaped by commercial factors.
- Targeted marketing on commercial (e.g. shopping) systems needs ethical considerations about exploitation of personal weaknesses.

One of the participants provided a link to a demo illustrating the concept of interactive visualisation of features that are extracted by a personalisation algorithm, with ability for data amendment by the end user: https://predictiveworld.watchdogs.com . As described by the participant, "the demo system makes algorithmic predictions and then when [the] user changes variables, the impact of changes on other things are visualised. For example, changing extraversion or neuroticism impacts longevity or well-being prediction based on academic research on the relationships between them.

Case study discussions

Fake news discussion – key points

- Misleading content has always existed but social media means it can spread more quickly and at the moment is organised – e.g. the content is being generated and managed in particular ways and for particular purposes, such as to generate money via advertising and destabilise Western democracies
- There is no evidence at the moment to suggest fake news has shaped the outcomes of elections. But we can see it does have a highly emotional element – for instance in stirring up hatred.
- Young people and mental health/welfare – young people are vulnerable if they cannot read between the lines, need to be protected? Especially as there is a social imperative for young people to use social media
- People don't want tailored news? Market research suggests people are fearful of invasions of privacy and their vulnerabilities being targeted
- Content can often be misleading but not fake. E.g. Political campaigns – UKIP pledge on the bus during the Brexit campaign. As citizens do we have a right to mislead? c.f the right to offend in relation to freedom of speech debates
- Solutions: education, public service broadcasting model for platforms, kite marking for trustworthy content, breaking the link between false content and making money.

Personalisation discussion – key points

- How do define personalisation? Alternative descriptions such as task based channelling. Helping people to reach their particular goals vs just putting people into broad groups. The assumptions platforms make about you can be patronising and inaccurate.
- Is nice for objects and commercial purposes but not for information due to problem of echo chambers and missing out on news that is 'good' for you.
- Need to achieve a 'fair' balance between being able to access very useful platforms for free and being advertised to.
- What level of control should users have over information that is collected about them and how it is used?
- Need for content based diversity to engage users and keep them up to date with important news etc. Diversity to be balanced with personalisation.

***Gaming the system*: key points**

- Google as dominant search engine- around 90% of market. It's search algorithms operate differently to other platforms
- Gaming the system as a practice that does occur and can be legitimate – c.f. search engine optimisation services
- Users tend to have a low awareness of how the ranking does or does not reflect reality and how it might have been gamed? Unlikely to look beyond first few results but may ignore ads
- Regulation – difference between censorship (removal of content) and alteration of ranking. Platforms have a responsibility to act on potentially illegal content when it is flagged to them but otherwise do not have a public service charter
- What can users/self-governance do? Value of information and digital literacy or the crowd effect to shape content. Users could have a button to press to rate content and prompt others to think about its validity etc.

Transparency discussion – key points

- Difference between transparency and meaningful transparency i.e. posting source code is not real transparency if most people are unable to understand it. What is transparency for? How will it be used? Parallels to other industries? E.g. listing of ingredients and E numbers etc. on food packaging. Trust is more important that openness to achieve fairness?
- Transparency of an algorithmic process does not overcome bias if the data the algorithm works on is biased.
- Wariness that transparency can lead to 'gaming the system', including by auditors. E.g., auditors may look to input factors rather than the outputs of an algorithm.
- if you are using transparency as a means to work out whether an algorithm/outcome is fair or not, you probably need to know more about the users. There is a tension there between minimizing what data you collect to work out whether your algorithm is fair, versus privacy where you want to collect as little as possible.
- Solutions: 1) An intermediate organisation to audit and analyse algorithms, trusted by end users to deliver opinion about what the operation of an algorithm means in practice? 2) Certification for algorithms – showing agreement to submit to regular testing, guarantees of good practice etc. 3) Voluntary certificates of fairness etc. e.g. Fairtrade.
- Transparency needs to be accompanied by responsibility and accountability

***Plenary discussion*: key points**

- Transparency needs to be meaningful. It doesn't just relate to algorithms themselves but also the transparency of decisions made about the operation of the algorithms. It also needs to be combined with accountability and responsibility. Transparency can have negative effects – e.g. users gaming the system; would this mean the system was not very robust in the first place?
- Children as particularly vulnerable online. Need for more effective child protection legislation but also a nuanced approach. Current tools – e.g. for flagging inappropriate searches being made by children – can be very blunt; they do not recognise context and can force action even when there is no real risk of harm.
- Who are vulnerable users? Lack of education can make a user vulnerable but it is also about ability to think critically and reflect on what is seen online (e.g. in particular reference to fake news). Risks and biases are not necessarily unique to single groups.
- Truth as a philosophical question. Distinction between facts and evidence – facts as canonical but evidence built on argumentation. Users expect the Internet to tell the truth?
- Lack of pluralism on social media (compared to traditional media) and its impacts. Lack of pluralism helps echo chambers to thrive, gives us no alternative to turn to. We can make judgements about the Daily Mail etc. based on its outputs and choose an alternative if we don't like it. We can make judgements about Google based on its outputs (i.e. without being able to see its algorithm) but have less choice to go elsewhere
- Personalisation is less problematic if it makes product recommendations compared to content ones? Is helpful if it can make it easier for you to reach your goals.
- Regulation – can we take lessons from elsewhere? E.g. casinos are required to monitor for customers exhibiting problem traits. Option for users to pay if they do not want personalised ads and content (but how fair would this be?). How about an algorithm ombudsman scheme we can go to query an automated decision that has been made about us? A touch button screen on a platform – how was this decision made about me? A daily digest of algorithm news and explanation?

Annex 1: Call for participation

UnBias: Emancipating users against algorithmic biases for a trusted digital economy

Invitation for stakeholder engagement

We would like to invite you to contribute to our ongoing research study by taking part in a small number of stakeholder engagement workshops. These workshops will involve professionals from various groups and will explore the implications of algorithm-mediated interactions on online platforms. They provide an opportunity for relevant stakeholders to put forward their perspectives and discuss the ways in which algorithms shape online behaviours, in particular in relation to access and the dissemination of information to users. The workshops will lead to the production of reports and policy recommendations as well as the design of a 'fairness toolkit' for users, online providers and other stakeholders.

The workshops will be audio recorded and transcribed. We might quote extracts from them in our project publications but we will take great care to anonymise all our data. This means that, unless you explicitly request otherwise, your taking part will be kept confidential.

The rest of this invitation provides further information about the project and the stakeholder workshops. You can also contact ansgar.koene@nottingham.ac.uk or helena.webb@cs.ox.ac.uk if you have any queries.

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?

Our project investigates the user experience of algorithm driven internet services and the processes of algorithm design. We focus in particular on the perspectives of young people 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.

What are the stakeholder engagement workshops about?

As part of our project we are running a series of stakeholder engagement workshops. We invite stakeholders from academia, education, government/regulatory oversight organisations, civil society organisations, media, industry and entrepreneurs to join us in exploring the implications of algorithm-mediated interactions on online platforms especially in relation to access and dissemination of information to users. The workshops will take place over a two-year period (2016 – 2018) and will seek to identify relevant perspectives and concerns as well as provide feedback on our project activities. They will also produce:

3. a set of policy and design recommendations for enhanced transparency and global fairness in information control algorithms;
4. a 'fairness toolkit' consisting of three co-designed tools
    iv. a consciousness raising tool for young internet users to help them understand online environments;
    v. an empowerment tool to help users navigate through online environments;
    vi. an empathy tool for online providers and other stakeholders to help them understand the concerns and rights of young internet users.

What does participation as stakeholder involve?

We invite each participant to take part in a series of workshops– likely to be between 4 – 6 events over the two-year period. We understand that it might be difficult for the same individual to attend each time, so alternatively we hope that each participating organisation will send a representative to each workshop. This will help us to ensure continuity across events. Participants will be asked to:

- participate fully in the workshop discussions. These will be audio recorded but we will take care to anonymise potentially identifying details (such as individual names and names of organisations) in all project outputs and publications. We also understand that participants may not be able to divulge commercially sensitive or confidential information.
- complete brief questionnaires to provide feedback on summary reports after each workshop (reports will be no more than 4 pages long). This will help us to ensure we represent all views accurately.
- contribute to the production of policy and design recommendations and the co-design of the 'fairness toolkit' through input provided both within the workshop sessions and after sessions, as necessary.

The workshops form a central part of our project. We hope that our participants will find them interesting and enjoy having an opportunity to put forward their professional perspectives and to shape our policy recommendations and other outputs. The workshops also provide an opportunity for participants to network with others in relevant fields.

Where, when and how will the workshop discussions take place?

Where: The workshops will generally take place in the UK. We will select locations that best suit the majority of participants so expect that they will often be held in London and other large UK cities. It may be possible for stakeholders to participate through tele-conferencing and we are also exploring the possibility of co-locating some workshops with existing events that particular stakeholder groups are likely to attend, e.g. a session at EuroDIG (http://www.eurodig.org/) for Internet Governance related stakeholders.

For further information about the project see http://unbias.wp.horizon.ac.uk/ or contact helena.webb@cs.ox.ac.uk or ansgar.koene@nottingham.ac.uk

When: The first set of workshops will take place in January 2017, with the exact date to be determined in coordination with the participants. We expect subsequent workshops will take place at intervals of approximately 4 months, placing the second set of workshops in May or early June 2017.

How: We are seeking participants from a range of stakeholder groups - academia, education, government/regulatory oversight organisations, civil society organisations, media, industry and entrepreneurs. In the first round of workshops we will run separate events for each stakeholder group so that we can explore the particular priorities of each group in turn. We anticipate that these workshops will have up to 12 participants. Some later workshops will likely be larger, as we will combine some or all stakeholder groups to allow participants from different fields to engage with each other. The specific plans for this will be developed in response the feedback following the first set of workshop.

Each workshop will address key topics relating to the implications of algorithm-mediated interactions on online platforms – for example; policy and practice, technical matters, and understandings of fairness. No formal preparation will be necessary but in case participants would like time to think through the issues involved beforehand, we will email round the main topics to be discussed at least a week before each event takes place.

Each workshop will be a half-day event, typically taking place over three hours in the afternoon. A sample workshop schedule is:

- 5 min Welcome
- 15 min Updates on UnBias project work and preliminary findings.
- 5 min Introduction to key workshop topic
- 10 min Questionnaire/ written individual views regarding the key topic
- 15 min Coffee break / assigning into subgroups for first round of discussion
- 30 min subgroup based discussion
- 5 min outcomes gathering break
- 10 min Reporting on outcomes of subgroup discussions
- 20 min Combined discussion reflecting on outcomes from the subgroups
- 15 min Coffee break
- 50 min Open discussion in issues raised by participants
- Post workshop drinks at a local pub

Privacy/confidentiality and data protection

All the workshops will be audio recorded and transcribed. This in order to facilitate our analysis and ensure that we capture all the detail of what is discussed. We will use quotations from the discussions in our project reports and other outputs but will take great care to anonymise them. We will remove or pseudonymise the names of participating individuals and organisations as well as other potentially identifying details. We will not reveal the identities of any participants (except at the workshops themselves) unless we are given explicit permission to do so. We will also ask all participants to observe the Chatham House rule – meaning that views expressed can be reported back elsewhere but that individual names and affiliations cannot.

We take data protection very seriously. All our data (audio recordings and de-identified transcripts) will be encrypted and stored securely in compliance with the data protection policies of the University of Nottingham and University of Oxford. The data will only be handled by researchers working on the UnBias project.

For more information about the UnBias project see http://unbias.wp.horizon.ac.uk/ and follow us on Twitter @UnBias_algos.

Annex 2: Pre-workshop questionnaire

Stakeholder Engagement: Algorithm Fairness

This questionnaire forms part of the first round of stakeholder engagement workshops for the UnBias project [http://unbias.wp.horizon.ac.uk/]. The purpose of this questionnaire is to record individual experiences and thoughts/concerns about the issues of algorithm fairness prior to group discussions in order to identify perspectives that might otherwise get lost due to the dynamics of the discussion.

Working definition of fairness – 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.

Question 1: On a scale of 0 to 5 rate how much experience you have with
[0=none; 1=very little; 2=little; 3=moderate amounts; 4=a lot; 5=this is a core activity of my work]

Q1a: using systems with embedded algorithms [ ]

Q1b: evaluating the behaviour of algorithmic systems [ ]

Q1c: creating new algorithms [ ]

Q1d: applying algorithms within systems for practical use [ ]

Q1e: deploying algorithmic system to end users [ ]

Question 2a: How would you rate our working definition of fairness, as applied to algorithms? [0=useless; 1=way off; 2=incomplete; 3=reasonable starting point; 4=good; 5=excellent]?
[ ] (more than one answer allowed)

Q2b: When judging algorithm fairness, is it more important to focus on the process (how it works) or the outcomes? [1=process only; 2=mostly process; 3=equal weight; 4=mostly outcomes 5=outcomes only] [ ]

Q2c: How would you change the definition of fairness?

.....

.....

.....

.....

Question 3: Thinking about the way in which issues of fairness by algorithms is being reported in the media, how accurately do you think these issues are presented, on a scale of: [0=not aware of media reports; 1=grossly distorted; 2=a bit distorted; 3=neutral; 4=well presented; 5=excellently presented]? [ ]

Q3 supplemental: Which news source does your answer primarily relate to (name all that you feel are relevant to your previous response)

.....

.....

.....

Question 4: Thinking about your **own experience** when using algorithm driven systems that mediate your access to online information, rate how fair you think the results were when using: [0=never used such a system; 1=very unfair; 2=somewhat unfair; 3=no opinion; 4=reasonably fair; 5=very fair]

Q4a: search engines [ ]

Q4b: product recommender systems [ ]

Q4c: news recommending systems [ ]

Question 5: Thinking about your level of concern regarding algorithm driven systems that mediate your access to online information, rate how concerned you are about the fairness of the results: [0=never used such a system; 1=not concerned; 2=slightly concerned; 3=reasonably concerned; 4=fairly concerned; 5=very concerned]

Q5a: search engines [ ]

Q5b: product recommender systems [ ]

Q5c: news recommending systems [ ]

Question 6: Thinking as an end user of a system where an algorithm is mediating the information you see; how would you describe the key properties that would define a fair algorithm

.....

.....

.....

.....

Question 7: Thinking as a designer of an algorithm based information providing system, what would be the main issues to include in the design process?

.....

.....

.....

.....

.....

.....

Question 8: Still thinking as a designer, what would be the main issues to include in order to make sure the system is fair?

.....

.....

.....

.....

.....

.....

Annex 3: Case-studies

UnBias project stakeholder engagement workshop 3rd February 2017

Case studies for discussion

Introduction to UnBias

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?

Our project investigates the user experience of algorithm driven internet services and the processes of algorithm design. We focus on the interest of a wide range of 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. This EPSRC funded project, full title, "UnBias: Emancipating Users Against Algorithmic Biases for a Trusted Digital Economy" runs from September 2016 to August 2018. It will provide policy recommendations, ethical guidelines and a 'fairness toolkit' that will be co-produced with stakeholders.

Aims of stakeholder workshops

Our UnBias stakeholder workshops bring together individuals from a range of professional backgrounds who are likely to have differing perspectives on issues of fairness in relation to algorithmic practices and algorithmic design. The workshops are opportunities to share perspectives and seek answers to key project questions such as:

- What constitutes a fair algorithm?
- What kinds of (legal and ethical) responsibilities do internet companies have to ensure their algorithms produce results that are fair and without bias?
- What factors might serve to enhance users' awareness of, and trust in, the role of algorithms in their online experience?
- How might concepts of fairness be built into algorithmic design?

The workshop discussions will be summarised in written reports and will be used to inform other activities in the project. This includes the production of policy recommendations the development of

a fairness toolkit consisting of three co-designed tools 1) a consciousness raising tool for young internet users to help them understand online environments; 2) an empowerment tool to help users navigate through online environments; 3) an empathy tool for online providers and other stakeholders to help them understand the concerns and rights of (young) internet users.

The case studies

We have prepared four case studies concerning key current debates around algorithmic fairness. These relate to: 1) gaming the system – anti-Semitic autocomplete and search results; 2) news recommendation and fake news; 3) personalisation algorithms; 4) algorithmic transparency.

The case studies will help to frame discussion in the first stakeholder workshop on February 3rd 2017. Participants will be divided into four discussion groups with each group focusing on a particular case study and questions arising from it. There will then be an opportunity for open debate on these issues. You might like to read through the case studies in advance of the workshop and take a little time to reflect on the questions for consideration put forward at the end of each one. If you have a particular preference to discuss a certain case study in the workshop please let us know and we will do our best to assign you to that group.

Definition

To aid discussion we also suggest the following definitions for key terms:

Bias – unjustified and/or unintended deviation in the distribution of algorithm outputs, with respect to one, or more, of its parameter dimensions.

Discrimination (should relate to legal definitions re protected categories) – unequal treatment of persons on the basis of 'protected characteristics' such as age, sexual identity or orientation, marital status, pregnancy, disability, race (including colour, nationality, ethnic of national origin), religion (or lack of religion). Including situations where the 'protected characteristics' is indirectly inferred via proxy categories.

Fairness – 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.

Transparency – the ability to see into the workings of the algorithm (and the relevant data) in order to know how the algorithm outputs are determined. This does not have to require publication of the source code, but might instead be more effectively achieved by a schematic diagram of the algorithm's decision steps.

CASE STUDY 1: Gaming the system -anti-Semitic autocomplete and search results

4 December 2016: Journalist Carole Cadwalladr reports in The Observer[1] that when typing 'are jews' into a Google search bar the platform autocompletes the search request with suggestions including 'are Jews evil'. Selecting this query leads to a page of responses; 9 out of the 10 first responses are from anti-Semitic sites (e.g. Stormfront) that 'confirm' that Jews are evil. Cadwalladr reports that searches beginning 'are women' and 'are muslims' also autocomplete with suggestions for '…evil' with the results of searches listing websites confirming they are evil.

Cadwalladr suggests that some users in the alt-right community are able to 'game' Google's PageRank algorithm in order to drive traffic to anti-Semitic sites and reinforce bigotry[2].

5 December 2016 A Google spokesperson states: *"We took action within hours of being notified on Friday of the autocomplete result. Our search results are a reflection of the content across the web. This means that sometimes unpleasant portrayals of sensitive subject matter online can affect what search results appear for a given query."* Reports suggest[3] that the auto complete terms mentioned by Cadwalladr relating to women and Jews have been removed but not those relating to Muslims. Search results have not been altered[4].

11 December 2016 Carole Cadwalladr reports[5] that typing 'did the hol' into a Google search bar leads to an auto-complete suggestion 'did the Holocaust happen' and that the first item returned from this search is an article from Stormfront 'Top 10 reasons why the Holocaust didn't happen.' She argues that this is enabling the spread of racist propaganda.

16 December 2016 The Guardian (sister paper to The Observer) reports[6] that it has found 'a dozen additional examples of biased search results', stating that its autocomplete function and search algorithm prioritise websites that claim climate change is a hoax, the Sandy Hook shooting did not happen and that being gay is a sin.

17 December 2016 Cadwalladr reports[7] that she has been able to replace Stormfront's article at the top of search results by paying to place an advertisement at the top of the page. She concludes that Google are prioritising commercial gain over fact.

20 December 2016 Search Engine Land report[8] that Google have now made alterations to its algorithm so that Holocaust denial sites are no longer the first results from a search for 'Did the Holocaust really happen?' It quotes Google as saying:

> *"Judging which pages on the web best answer a query is a challenging problem and we don't always get it right. When non-authoritative information ranks too high in our search results, we develop scalable, automated approaches to fix the problems, rather than manually removing these one-by-one. We recently made improvements to our algorithm that will help surface more high quality, credible content on the web. We'll continue to change our algorithms over time in order to tackle these challenges."*

The denial sites remain in search returns but are lower down the order.

27thDecember 2016 Search Engine Land reports that after a few days on the second page, the Stormfront site has returned to the first page of results for the search 'Did the Holocaust happen'.

Elaine's Idle Mind[9] blog site notes that Stormfront still tops searches with slightly different terms such as 'Did the Holocaust really happen' and 'Did the Holocaust ever happen'

"Google can't optimize its search algorithm to account for the 0.000001% of problematic queries that people might type in. There are infinite combinations of words that will inevitably offend somebody somewhere. In this case, Google put forth the minimum possible effort to get people to shut up and go away…"

Questions for consideration

1. How easy is it for a) individual users and b) groups of users to influence the order to responses in a web-search?

2. How could search engines weight their search results towards more authoritative results ahead of more popular ones? Should they?

3. To what extent should web search platforms manually manipulate their own algorithms and in what instances? NB Google has made a number of adjustments re anti-Semitism etc. and places a suicide help line at the top of searches about how to kill oneself.

4. To what extent should public opinion influence the ways in which platforms design and adjust their autocomplete and search algorithms?

5. What other features should and should not have a role in influencing the design of autocomplete and search algorithms?

CASE STUDY 2: News recommendation and fake news

May 3rd 2016 Gizmodo.com[10] reports on the role of human curators in determining what appears in Facebook's 'Trending news' section.

> *"We choose what's trending," said one. "There was no real standard for measuring what qualified as news and what didn't. It was up to the news curator to decide."*

May 9th 2016 Gizmodo.com[11] reveals that a former member of the trending news team has told them that "*Facebook workers routinely suppressed news stories of interest to conservative readers from the social network's influential "trending" news section"* and that "*they were instructed to artificially "inject" selected stories into the trending news module, even if they weren't popular enough to warrant inclusion—or in some cases weren't trending at all*"

26 August 2016 Facebook announces[12] that its trending feature will become more automated and no longer require people to write descriptions for trending topics. It emerges that all human members of the team have been fired[13].

29th August 2016 Facebook's fully automated trending news module is referred to as a 'disaster'[14] which has been 'pushing out' false news, controversial stories and sexually provocative content.

Early November 2016 Following Donald Trump's victory in the US presidential elections, various news[15] outlets begin speculating on the role of 'fake news' in his success. The New York Times publishes an article 'Donald Trump won because of Facebook'[16] It states:

> *"The most obvious way in which Facebook enabled a Trump victory has been its inability (or refusal) to address the problem of hoax or fake news. Fake news is not a problem unique to Facebook, but Facebook's enormous audience, and the mechanisms of distribution on which the site relies — i.e., the emotionally charged activity of sharing, and the show-me-more-like-this feedback loop of the news feed algorithm — makes it the only site to support a genuinely lucrative market in which shady publishers arbitrage traffic by enticing people off of Facebook and onto ad-festooned websites, …The valiant efforts of Snopes and other debunking organisations were insufficient; Facebook's labyrinthine sharing and privacy settings mean that fact-checks get lost in the shuffle. Often, no one would even need to click on and read the story for the headline itself to become a widely distributed talking point, repeated elsewhere online, or, sometimes, in real life…Tens of millions of people…were served up or shared emotionally charged news stories about the candidates, because Facebook's sorting algorithm understood from experience that they were seeking such stories."*

10th November 2016 Facebook CEO Mark Zuckerberg calls claims that fake news on the platform influenced the election 'pretty crazy'[17].

12th November 2016 The New York times reports[18] that despite Zuckerberg's public statement, senior members at Facebook are worried about the platform's role in the outcome of the election and are taking confidential steps to investigate it. On the same day Mark Zuckerberg posts an update on his own Facebook page stating that *"Of all the content on Facebook, more than 99% of what people see is authentic".*

Throughout **November 2016** news media report potential solutions to the fake news 'problem' on Facebook. The Guardian lists[19] three kinds of solution:

> *"Most of the solutions fall into three general categories: the hiring of human editors; crowdsourcing, and technological or algorithmic solutions."*

Bloomberg technology[20] suggests that:

> "Facebook could tweak its algorithm to promote related articles from sites like FactCheck.org so they show up next to questionable stories on the same topic in the news feed"

Vox.com[21] suggest that Facebook are exercising caution as they fear being seen as biased against conservatives following on from the algorithm controversy in spring 2016.

15th December 2016 Facebook unveils plans to stop the spread of fake news[22]. Changes include:

- new ways for users to report suspected fake news stories – which will then be sent to fact checking organisations. Some stories will then carry the label 'disputed'
- adjustment to the News Feed algorithm to decrease the reach of fake news stories – e.g. shares when only a headline have been read will receive less prominence than when a full article has been read.

Questions for consideration

1. How effective will Facebook's new plans be in addressing the spread of fake news?

2. Should social media platforms accept responsibility for the spread of fake news on their platform? If so, are they also obliged to take steps to attempt to address it?

3. Can an algorithm alone prevent the spread of fake news?

4. Why might social media platforms be reluctant to address the problems of fake news? E.g. due to political, commercial reasons etc. Are any of these reasons justifiable?

5. To what extent is the spread of fake news on social media similar to or different from the offline spread of rumour, conspiracy theories, propaganda etc.?

CASE STUDY 3: Personalisation algorithms

Online personalisation mechanisms are designed to sift through data in order to supply users with content that is apparently most personally relevant and appealing to us. These algorithm driven mechanisms curate and shape much of our browsing experience – for instance the results of a Google search may be influenced by past searches we have made; the content and order of items in our personal Facebook newsfeed will be shaped by what Facebook's algorithms have calculated is of most interest to us and Amazon shows us products we might like based on our past purchases and searches on the platform.

Personalisation can be seen as helpful to online users as it avoids them having to sort through the vast amounts of content that are available online and instead directs them towards what they might find most useful or interesting[23]. It also brings many advantages to internet companies as it can increase user numbers and drive up purchasing and/or advertising revenues[24]. However, concerns have been raised around the 'gatekeeping' role played by personalisation algorithms. These concerns are exacerbated by the opaque nature of most personalisation algorithms and the lack of regulation around them[25]. Issues include:

1) **The creation of online echo chambers**. On a social network such as Facebook personalisation algorithms ensure that we are more likely to see content similar to what we have previously 'like' or commented on. This can mean that we repeatedly see content that reaffirms our existing views and we are not exposed to anything that might challenge our own thinking[26].Recent political events such as the election of Donald Trump to the US presidency have led to much debate over the role in echo chambers in modern democratic societies[27].

2) **The results of personalisation algorithms may be inaccurate and even discriminatory.** Despite the sophisticated calculations underpinning them, the algorithms that recommend or advertise a purchase to us or present us with content we might want to see, might not in fact reflect our own interests. This can be an annoyance or distraction. More seriously, algorithms might alternatively curate content for different users in ways that can be perceived as discriminatory against particular social groups[28]. For instance researchers at Carnegie Mellon University[29] ran experimental online searches with various simulated user profiles and found that significantly fewer female users than males were shown advertisements promising them help getting high paid jobs. A member of the research team commented "Many important decisions about the ads we see are being made by online systems. Oversight of these 'black boxes' is necessary to make sure they don't compromise our values.[30]"

3) **Personalisation algorithms function to collate and act on information collected about the online user.** Many users may feel uncomfortable about this, for instance feeling that it constitutes a breach of their privacy[31]. The impact of this perception can be seen in the emergence of options to opt out of personalisation advertisements on platforms such as Google[32] and the growth of platforms that claim not to track you[33].

Questions for consideration

1. What is your response to this comment from Mark Zuckerberg to explain the value of personalisation on the platform? *"A squirrel dying in front of your house may be more relevant to your interests right now than people dying in Africa"*[34]

2. What (legal or ethical) responsibilities do internet platforms have to ensure their personalisation algorithms are not inaccurate or discriminatory?

3. To what extent should users be able to determine how much or how little personal data internet platforms collect about us?

4. To what extent are personalisation algorithms necessary or useful for efficient internet browsing?

5. To what extent would algorithmic transparency help to address concerns raised about the negative impacts of personalisation algorithms?

CASE STUDY 4: Algorithmic transparency

In June 2008 the social news aggregator and discussion site Reddit announced that it was going open source. This meant "the code behind Reddit is available to the public for download, and we're inviting the public to submit code to help improve the site."[35]. It became possible to understand how algorithms on the site function – for instance, how individual posts and comments can be up or down voted and the consequences this has for where they are placed on the page[36].

Reddit positioned this decision as a commitment to the 'open source world' and something its users deserved[37]. Commentators at the time also suggested this was a move against a commercial competitor[38], that it would appeal to marketers (as they could learn how to ensure their content would achieve prominence)[39] and that individual users would work out how to 'game' the system to promote their own posts[40]. In December 2016 Reddit announced an overhaul of its voting rules to attempt to 'mitigate cheating and brigading.[41].

The kind of transparency exercised by Reddit is relatively rare amongst internet companies. In most cases, the algorithms used to derive their platforms are treated as proprietary and commercially sensitive. However, in recent times there have been increasing calls for algorithmic transparency[42]. It is argued that since algorithms have the capacity to impact many areas of people's lives, it is important that we are able to know how they function and hold them accountable[43] Transparency of this kind might help us to challenge apparent (discriminatory) bias in the outcomes of search engine queries[44]. or work to address the negative consequences of online echo chambers[45]

Reservations have also been expressed about algorithmic transparency. In addition to commercial concerns, it is possible to argue any positive impact of transparency would be limited.

> *"Even if a company were to release a proprietary algorithm to the public, the task of understanding and reacting would be extremely difficult. Consumers and policymakers are unlikely to understand what an algorithm says or means, it would likely undergo continuous change over time or in reaction to new data inputs, and it would be difficult to decide how to measure unfairness — whether by looking at inputs, outputs, decision trees, or eventual effects. These challenges may leave even companies that care deeply about avoiding discrimination unsure as to what best practices really are[46]"*

In December 2015 the Commissioner of the US Federal Trade Commission[47] noted the potential for algorithmic calculations to have effects that might be unfair, unethical or discriminatory but also acknowledge the difficulties involved in making those algorithms transparent to the public. Consequently, she called for companies to carry out proactive and thorough internal auditing of their own data use.

In December 2016 Elizabeth Denham delivered her first speech as Information Commissioner[48]. This speech included the comment: "*Wherever you are in the world the themes of good data protection are the same – that consumers have the right to know what's happening with their information combined with business transparency and accountability."*

Questions for consideration

1. What do large companies such as Google and Facebook have to gain from adopting algorithmic transparency?

2. What might individual users stand to gain from algorithmic transparency?

3. To what extent are concerns over commercial sensitivity a valid argument against algorithmic transparency?

4. Is there any point to algorithmic transparency? Will it just mean that the majority won't understand the algorithms they are using and the minority will be able 'game' the platform they are using?

5. What might be some useful alternatives to algorithmic transparency?

Annex 4: Post-event questionnaire

Stakeholder Engagement: Algorithm Fairness

This questionnaire forms the final part of the first multi-stakeholder engagement workshop for the UnBias Project.

The purpose of this short questionnaire is to allow individuals to reflect on issues and their perspectives surrounding algorithm fairness, given their participation and discussion in the workshop.

This is just a reminder of our working definition of "fairness" that was used as a foundation for discussion in the workshop:

Working definition of fairness – 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.

Question 1: what did you find most useful about the workshop? Did you learn anything new about current concerns around algorithms on internet platforms?

.....

.....

Question 2: has your perspective on algorithm fairness been altered by the workshop? Please elaborate on your answer.

.....

.....

Question 3a: which case study were you involved in discussing?

.....

Question 3b: if you were given the opportunity to resolve the problematic issues that were discussed in relation to your case study, how would you go about this? (you can imagine a hypothetical situation where there are no limitations to the solutions you are able to offer)

.....

.....

Question 4: what do you think are the key ethical issues facing the use of algorithms now and in the coming years?

.....

.....

Thank you very much for taking the time to complete this questionnaire.

If you would like more information about the UnBias project, visit our website at http://unbias.wp.horizon.ac.uk/ .

You can also contact helena.webb@cs.ox.ac.uk or ansgar.koene@nottingham.ac.uk if you have any questions about the project or the stakeholder workshops.

Annex 5: List of workshop participants

Name	Affiliation
Robert van der Pluijm	Bibblio.org
Adam Galloway	HIVE
Daniele Orner	Iota
Rene Arnold	WIK-Consult GmBH
Anat Elhalal	Digital Catapult
Teresa Macchia	Digital Catapult
Harry Armstrong	Nesta
Sofia Ceppi	University of Edinburgh
Michael Veale	UCL STEaPP
Lachlan Urquhart	U. Nottingham Horizon DER
John Northam	University of Winchester
Mark Leiser	University of Strathclyde
Uta Kohl	U. Aberdeen
Rob Procter	Warwick U. and Alan Turing Institute
Reuben Binns	U. Oxford SOCIAM project
Eric Meyer	Oxford Internet Institute and ATI
Anna Schneider	Fresenius Hochschule
Vesselin Popov	U. Cambridge Business School
Virginia Dignum	Faculty of Technology Policy and Management, TU Delft
Laura James	Doteveryone
Raimondi Andrea	Facebook.tracking.exposed
Stephane Goldstein	InformAll
Yohko Hatada	EMLS RI
Jen Persson	defenddigitalme
Irmantas Baltrusaitis	Cressey College
James Wilton	Repton Boardingschool

Annex 6: Participant recommendations for algorithm fairness

[Titles summarizing each participant suggestion were added by the UnBias team, as was the grouping into three main types of recommendation]

---- ---- *Criteria relating to system reliability* ---- ----

Avoid recursive reasoning
e.g. recommendation ranking based on clicks of ranked recommendations (users tend to click on first entries)

System monitoring over time
Adaptive systems can change their behaviour with time. It is not sufficient to evaluate if the system is fair (e.g. non-discriminating against protected groups) at time of first release.

Informative labelling of methods, beyond buzz-words
Many systems that are advertised as AI supported adaptive systems are actually simple 'click counting' systems.

Need to think in hierarchies of factor relevance
It seems hard to see how clearly 'task irrelevant factors' can be delineated in a 'context dependent evaluation'. Wouldn't most factors need to be hierarchised rather than rejected as 'task irrelevant'?

Need for information integrity
Does the algorithm distinguish between 'alternative' and real facts when processing input information? Particularly with regard to search engines and news providers, building into algorithms a sense that they offer balanced results, with due regard for trustworthiness (and possibly challenge rather than echo) is important.

Transparency of decision making process
If an algorithm is to make decisions on your behalf, you should be aware of the decision making process it is following, and of the data upon which its process is founded.

Independent system evaluation
Who is doing the 'fairness' evaluation of the system? Are they independent?

---- ---- *Criteria relating to (non) interference with user control/agency* ---- ----

Ability to opt-out of the algorithm but still use the service
Can you continue to live your life independently and effectively without having to pass through an algorithm? That is has the algorithm become the only route to do x or y or z? Can you choose between algorithms? After awareness of an algorithmic function is raised, there should be ways for a user to adapt or reject some or all of its functions.

Ability to appeal/negate algorithm decisions
What power do you have to appeal second guess and negate a decision made by algorithm?

Ability to set individual limits on data collection and use
Consumers want to have a voice in deciding what data may be collected and due to which reasons, they should have possibility to set individual limits to what is collected and what it is used for.

Non-interference with user agency
Do not impair our capacity as knowers, i.e to gain knowledge by telling/being told, and do not impair our capacity as agents, i.e. to make sense of the social world.

User centred ownership of interaction
Users are increasingly 'productised' entities that feed hungry algorithms & data needs, losing ownership. How is society informed of the real impact and how can it influence or define ethics around algorithm use?

No re-use of personal data without explicit permission
The algorithm should not use personal data which was provided by the data subject for another purpose or in a different context without an explicit new permission.

Inferred personal data = personal data
Personal data inferred by or created by an algorithm is subject to Data Protection law.

Ethical limits on taking advantage of detected personal weaknesses
eg an algorithm that detects someone has an eating disorder and sends them adverts for diets. Some such processing operations should be banned.

Transparency of data used by algorithm
Transparency is often connected to fairness as consumers report it as unfair that they do not even know what kind of personal data is collected to feed algorithms

Ethical precautions more important than higher accuracy
We must recognise that even algorithms with very positive outcomes for society or the individual may, in the process of arriving at their conclusion, go through computational steps that are inherently discriminatory on protected traits or that lack in transparency to the extent that they cannot be adequately explained to the end user, or to those whose data has been used to train the algorithm. A sole focus on outcomes is therefore inappropriate. Similarly, an over emphasis on the precise machine learning technique being used to generate results can also lead practitioners astray, with a race for higher accuracy negating the importance of taking proper ethical precautions in algorithm design and communicating those processes to citizens and stakeholders.

---- ---- *Criteria relating to social norms and values* ---- ----

Need to clearly justify meaning assigned to data
The manner in which an algorithm interprets words needs to be fair; loading words and phrases with meaning because of common user preoccupations may be fair, but may be corrosive.

Openness to scrutiny – process & output
The necessary weighting is given to each of the relevant socio-cultural factors. The processes as well as the output can be shown to be fair and where the whole is subject to scrutiny.

Accountable, Comprehensible, Testable

Fairness of algorithm depends on objectives of users
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.

Explain-ability to data subject
The reasoning and behaviour of the algorithm as applied to the facts of the data subjects case can be demonstrated and explained and understood by the data subject

Transparency to end-users of underlying values

Freedom to explore algorithm effect
Can you experiment with the algorithm in a non binding way? That is can you input a variety of responses or inputs to check the effect the algorithm will have?

Contextually appropriate transparency of algorithm mechanics
Exposing the underlying mechanics of algorithms in a contextually appropriate manner (eg both audience and environment) to enable interrogation and measurement against relevant rules and standards (eg legal frameworks, social norms, ethical standards)

Context of algorithm use as factor in redress/appeal
consideration of outcomes on different groups, particularly minorities who may suffer disproportionately from bad outcomes. Evaluation of context of algorithm use, including routes for redress/appeal

Balancing individual values and socio-cultural values
Evaluation of the algorithm processes and outcomes against individual values instead of socio-cultural values as there may be fundamental differences between those concepts (for example in Turkey, Saudi Arabia)

Allowing for pluralism of voices/values
include treatment of minority opinion; non-monotonicity

Values imbedded in algorithms must match cultural context of use
If algorithms are rated in terms of socio cultural values, what values exactly are taken as a reference? Thinking of “European” socio-cultural Values in comparison to for example values in other regions of the world makes it very difficult to estimate whether there will be a common sense of what is “fair” or “unfair”, “good” or “legal”...

Equal treatment of population groups
Everyone is treated/processed equally without prejudice or advantage due to task irrelevant factors (examples: gender, age, religion).

Algorithm should be tune-able by individual users
fairness is subjective

Disparate outcomes for some lifestyle choices can be OK
Sometimes disparate outcomes are acceptable if based on individual choices to pursue lifestyles which we feel do not deserve special protection or compensation (e.g. extreme sports)

Annex 7: Key properties that would define a fair algorithm

Pre-workshop survey question 6: Thinking as an end user of a system where an algorithm is mediating the information you see; how would you describe the key properties that would define a fair algorithm.

Transparency of decision criteria/process

- Increasing the accountability of the process underpinning why the information I see is being presented to me. Like with newspapers, there should be baseline standards, such as an editorial code, and then I can know minimum considerations algorithm designers have reflected on. Overtly stating their approach could be useful too (i.e. we trained this algorithm using these datasets, hence there is a higher risk of bias for these factors…race, health, political values) The EU GDPR special categories of personal data could be a good starting point as high risk information where increased accountability around processing by algorithms is needed.

- It would allow me not just to get what I wanted, but to understand at least at a basic level how it sorted through the options to present me that result.

- Focus on transparency of bias and plurality of biases.

- Transparent in system goal and profile they generated on me, gives me control over output, I have the ability to kill personalisation, I want system to include notion of diversity/serendipity.

- The user should have the same level of knowledge or more, than the 'system' about what information is being used "you liked this, you may also like this" - users should be able to draw on a report to see order to personalise the algorithm, that data should be made available to the user. ("If behind the scenes about the individual. If information is used in what past purchasing is influencing the current display selection.)

- A fair algorithm should:

- Ask user permission for a specific prediction to be made using a specific set of data
- Explain what value will be gained from giving permission
- Give end user chance to alter and interact with prediction (e.g. use my social media data but not my IP, this bit but not that bit, adjust predicted age and gender, etc.)
- Show the results of the prediction (i.e. end user should know whatever system does)
- Let user easily opt out of prediction once it's been made
- Internally, it should not discriminate based on protected traits (e.g. gender, ethnicity, sexual or political orientation) unless the user has explicitly requested that such information be used to improve the prediction.

Provide a broad range of responses so as not to limit agency of user

- Its ability to produce a good range of results, preferably contrasting ones, that are able to reflect different viewpoints, types of sources, etc.

- Personalized but not oblivious of the world, customizable, somehow interactive in the decision making process.

- Difficult...when we spoke to consumers recently, they told us, that fairness means full access to literally "everything" except illegal content. They want to be free, but with a guardian angel

protecting them from harm. So Fairness in Germany is perceived as all “decent” users being treated the same

- Depending on the specific purpose of the algorithm key properties may change, but overall I would appreciate if the algorithm driven system took all information on the subject into account and presented me with a ranking suited to my request.

-’Fair’ seems the wrong word when thinking as an individual, as it is about different effects on difference groups, and ability of different views to appear in my information stream

- combines personalisation and diversity

Other

- Treatment of minorities; not influenced by adverts/paying parties, or at least very explicit about this.; not influenced by previous user behavior/location/time

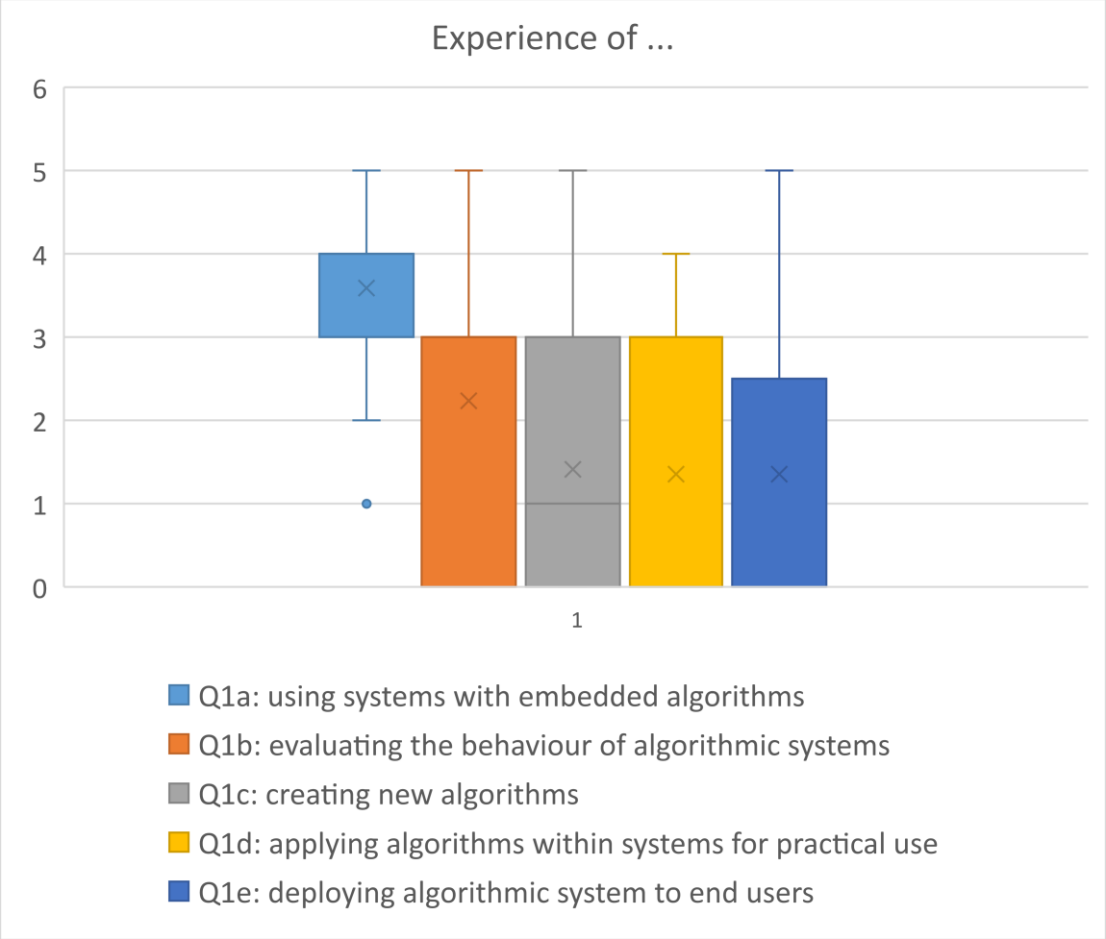
- a non-learning algorithm

- One which guards me according to subject matter rather than political preferences/views

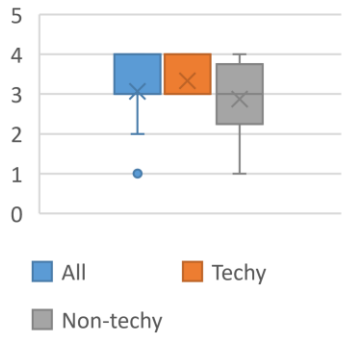
- Provenence/trustworthiness of the source with regard to factual content is a key factor in hierarchisation of results displayed; it is unfair to provide unreliable content to users. Geographical/cultural breadth of reference also a factor. Potential to create algorithms which deliberately ‘challenge’ the recipient of information rather than feed them what they want/expect. Linguistic connotations should be given a balanced assessment ie. The search ‘Did the holocaust’ should not presuppose that the next word will be ‘really happen’ even if that was the most common search term. It ought to be that the more contentious ‘readings’ of words within search terms is represented in the hierarchy of search results displayed. Transparency of the algorithm’s more contentious functions eg. A search engine separating the ‘most popular’ results from the ‘most trusted’ ones. Likewise shopping adverts based on your own previous searches are more ‘fair’ than ones that make assumptions about your needs based on personal data of other kinds. Potential to offer more options for the user so they can affect the algorithm’s function?

Annex 8: Likert scale rating responses to questions about algorithm fairness

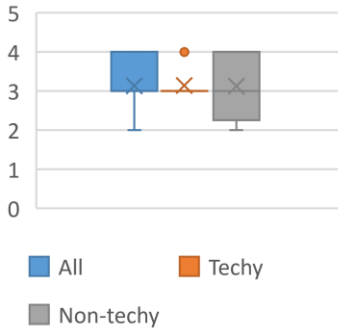
Participants were divided into "Techy" and "Non-techy" based on their responses to Question 1: "On a scale of 0 to 5 rate how much experience you have with [0=none; 1=very little; 2=little; 3=moderate amounts; 4=a lot; 5=this is a core activity of my work]"
Participants who responded 2 or lower for Q1c "creating new algorithms", Q1d "applying algorithms within systems for practical use" and Q1e "deploying algorithmic system to end users" were classes as "non-techy".



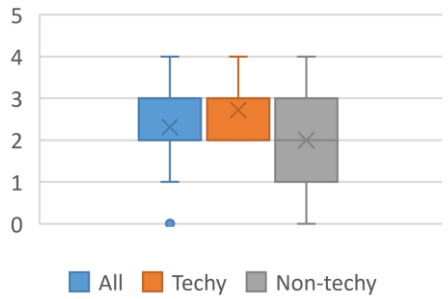
Question 2a: How would you rate our working definition of fairness, as applied to algorithms? [0=useless; 1=way off; 2=incomplete; 3=reasonable starting point; 4=good; 5=excellent]



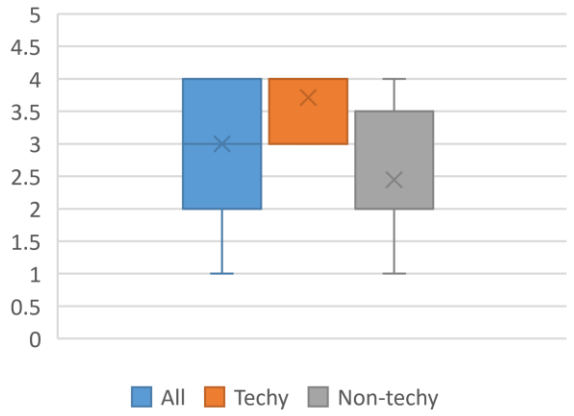
Question 2b: When judging algorithm fairness, is it more important to focus on the process (how it works) or the outcomes? [1=process only; 2=mostly process; 3=equal weight; 4=mostly outcomes 5=outcomes only]



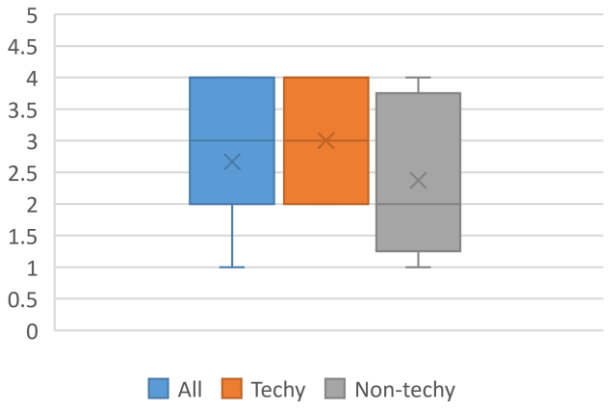
Question 3: Thinking about the way in which issues of fairness by algorithms is being reported in the media, how accurately do you think these issues are presented, on a scale of: [0=not aware of media reports; 1=grossly distorted; 2=a bit distorted; 3=neutral; 4=well presented; 5=excellently presented]



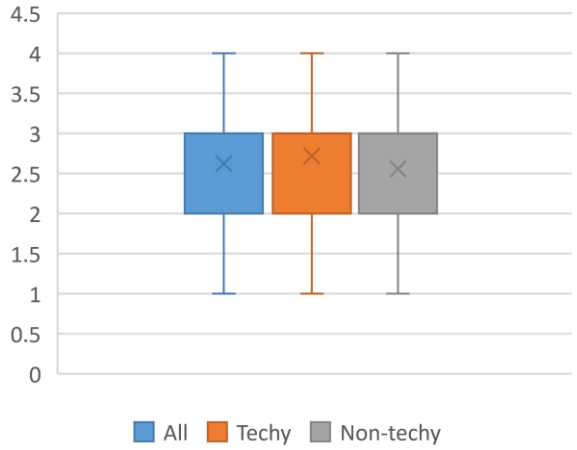
Question 4a: Thinking about your own experience when using search engines, rate how fair you think the results were when using: [0=never used such a system; 1=very unfair; 2=somewhat unfair; 3=no opinion; 4=reasonably fair; 5=very fair]



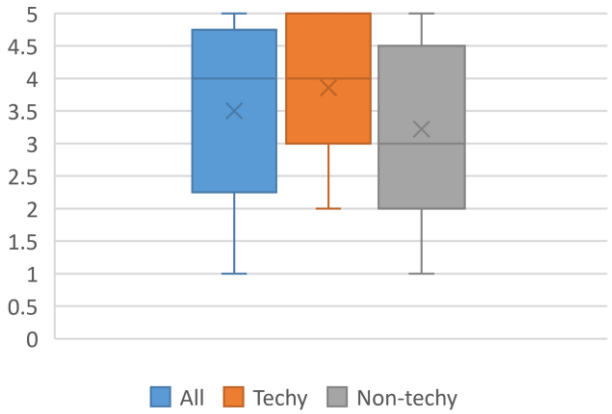
Question 4b: Thinking about your own experience when using product recommender systems, rate how fair you think the results were when using: [0=never used such a system; 1=very unfair; 2=somewhat unfair; 3=no opinion; 4=reasonably fair; 5=very fair]



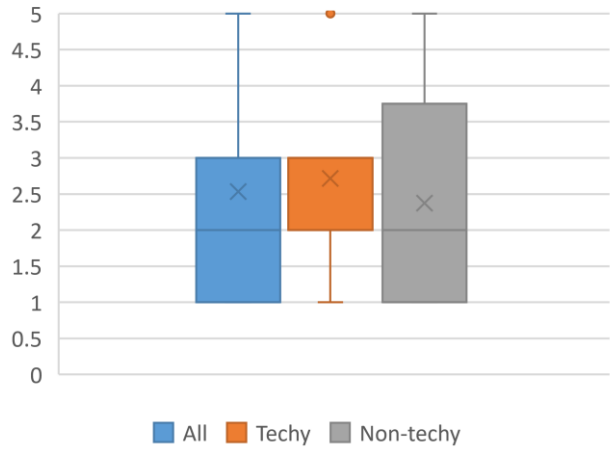
Question 4c: Thinking about your own experience when using news recommending systems, rate how fair you think the results were when using: [0=never used such a system; 1=very unfair; 2=somewhat unfair; 3=no opinion; 4=reasonably fair; 5=very fair]



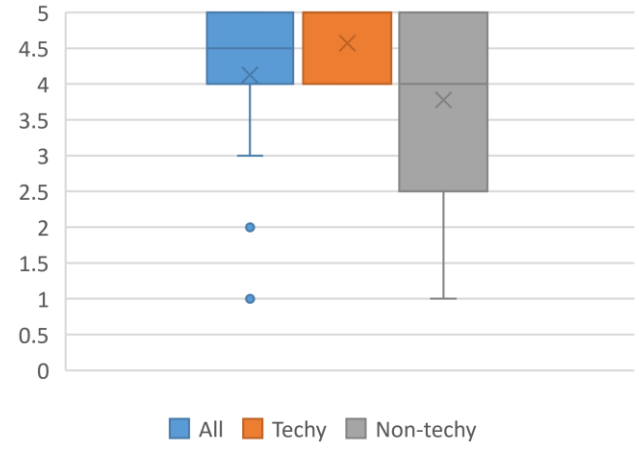
Question 5a: Thinking about your level of concern regarding search engines, rate how concerned you are about the fairness of the results: [0=never used such a system; 1=not concerned; 2=slightly concerned; 3=reasonably concerned; 4=fairly concerned; 5=very concerned]



Question 5b: Thinking about your level of concern product recommender systems, rate how concerned you are about the fairness of the results: [0=never used such a system; 1=not concerned; 2=slightly concerned; 3=reasonably concerned; 4=fairly concerned; 5=very concerned]



Question 5c: Thinking about your level of concern regarding news recommending systems, rate how concerned you are about the fairness of the results: [0=never used such a system; 1=not concerned; 2=slightly concerned; 3=reasonably concerned; 4=fairly concerned; 5=very concerned]



Annex 9: Recommendation for algorithm design

***Pre-workshop survey question 7**: Thinking as a designer of an algorithm based information providing system, what would be the main issues to include in the design process?*

- Ensure the algorithm is testable against contextually appropriate standards of fairness. As socio-cultural values underpinning such assessments of fairness may shift, keep the process of assessing algorithms resilient to change. Devising spectrums of harm for different contexts (i.e. fake news is more of an issue around election time so increased focus on ensuring algorithms provide fair news coverage during the run up to an election could be important…to avoid filter bubbles)

- Increasingly, the algorithms are either so complex or generated by other algorithms that even the designers can fail to understand what is happening inside the black box.

- external factors – not just what I am designing, but other information and systems in play, that either feed into my system or operate in parallel and affect results. This includes both technical and social systems (eg organised up/down voting of certain content). Duty of care to society as well as target users/customers. Stakeholder analysis – who wants what.

- 1) To give users partial control over the algorithm, in particular if the aim is not just to provide users with a fixed set of options among which they can choose. 2) Make the algorithm transparent while taking also into account that the algorithm is an intellectual property of the provider and describing how it works would benefit his/her competitors.

- Making a difference between results that are delivered to consumers proactively and results that are shown based on active search by consumers. It´s OK and part of the business to suggest news and products, but not to embezzle content

- Notion of relevance build into all results – matching query with info - never enforce information because of popularity or advertisement (be transparent and have ways for user to turn this off)

- The manner in which user data is interpreted and the hierarchisation of returned results. News/reference searches must be protected. If the system is a commercial (eg. Shopping) system consideration of the ethics of targeted marketing is important eg. Is it 'fair' to use the fact that on Facebook someone's status is 'engaged' and adverts focus on weight loss or cures for baldness?

- Often crucial to have the sensitive information in the training dataset so that the developer can design the system specifically not to discriminate on these traits i.e. if you don't know that one of the branches in decision tree is 'gay or straight', then you can't prove algorithm does not discriminate. Traceability of data and of features also important, so that the moment of change (or relevant thresholds for decision-making) can be identified. For example, does insurance decision go from a yes to a no when conscientiousness goes from 53 to 52, or is it when it goes from 30 to 29

- Can users determine how the personalisation operates? Control which kinds of attributes the determining is based on?

- User testing; Diverse forms of evaluation; Different measures of performance M&E data coding & cleaning & collection oversight

- empathy

- avoid manifestations of cultural and economic chavinism as well as racial, gender bias etc.

Pre-workshop survey question 8: *Still thinking as a designer, what would be the main issues to include in order to make sure the system is fair?*

- Transparency in functionality: translating how it works to end users in a way that lets them get a sense of what is going on, whilst not being too technically detailed. Striking the balance between the two would be tough.

- Not just transparency (e.g. someone could look at it, but in practice might not ever), but an active system of audit (knowledgeable people will look at it) for the most important algorithms.

- What impacts can we anticipate, how might these change (due toexternal factors) in the future? How are risks and benefits distributed across different groups? What are the 'edge cases' in terms of user groups and needs and how are they affected? (particularly thinking of vulnerable communities and minorities, and also elites – do they disproportionately benefit?) what happens if things go wrong? Are the tradeoffs transparent or visible to stakeholders?

- Giving insight into data which is collected and used in curation for individual user

- Transparency of function and process in layman's terms wherever possible. Sophistication (and responsible handling) of linguistic interpretation by algorithms. Sufficient user control over algorithm's 'reading' of personal data and search history. Breadth of search before returning results, and a balance between 'trust' and 'popularity'.

- Any system using data about me, should be able to show me what it "knows" transparently, in the form of personal data usage reports on demand. ie. if on order to present me personalised news, the system has used purchased data from a data broker, combined with meta data from my system use, and a system user profile, I should be able to call up an on screen view of what data has been used and its source. But in order to do that first I need to know that an algorithm has been used. "Automatic decision making has been used [in this screen display/search result/purchasing suggestion] to personalise this result. To see the data used about you in that process, click here."

- The main issue is to make the algorithm flexible/customisable enough to account for the subjective perception of what is fair and what is biased.

- Interactivity, so features that are extracted can be clearly seen and amended by an end user. Not quite the full solution but we attempted to mock up what this might look like on https://predictiveworld.watchdogs.com/en/. The system makes algorithmic predictions and then when user changes variables, the impact of changes on other things are visualised. For example changing extraversion or neuroticism impacts longevity or well being prediction based on academic research on the relationships between them. Can imagine a similar thing with product or news recommendation algorithms where you alter Google's predictions of your interests and profile.

- Measure disparate impacts on what perspectives appear and how they differ by groups

- empathy

- Apart from the obvious forms of prejudice as outlined above, I would be wary of systems which present political issues as 'closed'. Whilst I have my own political stance I am mindful of the risks of algorithmic processes which 'mirror' my pre-existing views, closing my mind to other possibilities.

[1] https://www.theguardian.com/technology/2016/dec/04/google-democracy-truth-internet-search-facebook
[2] https://en.wikipedia.org/wiki/PageRank
[3] https://www.middleeastmonitor.com/20161207-google-removes-anti-semitic-autocompletes-from-its-search-engines/ and
[4] http://www.telegraph.co.uk/technology/2016/12/05/google-removes-anti-semitic-suggestion-autocomplete-feature/
[5] https://www.theguardian.com/commentisfree/2016/dec/11/google-frames-shapes-and-distorts-how-we-see-world
[6] https://www.theguardian.com/technology/2016/dec/16/google-autocomplete-rightwing-bias-algorithm-political-propaganda
[7] https://www.theguardian.com/technology/2016/dec/17/holocaust-deniers-google-search-top-spot
[8] http://searchengineland.com/googles-results-no-longer-in-denial-over-holocaust-265832
[9] http://elaineou.com/2016/12/27/googles-anti-semitic-search-queries/
[10] http://gizmodo.com/want-to-know-what-facebook-really-thinks-of-journalists-1773916117
[11] http://gizmodo.com/former-facebook-workers-we-routinely-suppressed-conser-1775461006
[12] http://newsroom.fb.com/news/2016/08/search-fyi-an-update-to-trending/
[13] http://uk.businessinsider.com/facebook-fires-trending-topics-team-2016-8
[14] https://www.theguardian.com/technology/2016/aug/29/facebook-fires-trending-topics-team-algorithm and http://arstechnica.co.uk/business/2016/08/facebook-fires-human-editors-algorithm-immediately-posts-fake-news/
[15] https://www.buzzfeed.com/craigsilverman/how-macedonia-became-a-global-hub-for-pro-trump-misinfo?utm_term=.rcx6Rb8Pq9#.wjVNoQK0jb and
https://www.theguardian.com/commentisfree/2016/nov/14/fake-news-donald-trump-election-alt-right-social-media-tech-companies and http://heavy.com/news/2016/11/fake-news-facebook-google-donald-trump-bernie-sanders-wins-election/
[16] http://nymag.com/selectall/2016/11/donald-trump-won-because-of-facebook.html
[17] http://www.theverge.com/2016/11/10/13594558/mark-zuckerberg-election-fake-news-trump
[18] https://www.nytimes.com/2016/11/14/technology/facebook-is-said-to-question-its-influence-in-election.html?_r=2
[19] https://www.theguardian.com/technology/2016/nov/29/facebook-fake-news-problem-experts-pitch-ideas-algorithms
[20] https://www.bloomberg.com/news/articles/2016-11-23/facebook-s-quest-to-stop-fake-news-risks-becoming-slippery-slope
[21] http://www.vox.com/new-money/2016/11/16/13637310/facebook-fake-news-explained
[22] http://www.independent.co.uk/news/uk/politics/facebooks-plan-to-stop-fake-news-revealed-by-mark-zuckerberg-facebook-changes-what-are-they-fake-a7478071.html
[23] https://www.wired.com/insights/2014/11/the-internet-of-me/
[24] http://www.nytimes.com/2011/05/23/opinion/23pariser.html
[25] https://www.oii.ox.ac.uk/should-there-be-a-better-accounting-of-the-algorithms-that-choose-our-news-for-us/
[26] http://www.nytimes.com/2011/05/29/technology/29stream.html
[27] http://www.independent.co.uk/voices/donald-trump-president-social-media-echo-chamber-hypernormalisation-adam-curtis-protests-blame-a7409481.html
[28] https://www.nytimes.com/2015/07/10/upshot/when-algorithms-discriminate.html?_r=0
[29] https://www.cmu.edu/news/stories/archives/2015/july/online-ads-research.html
[30] https://www.cmu.edu/news/stories/archives/2015/july/online-ads-research.html
[31] http://www.ey.com/uk/en/services/specialty-services/the-data-revolt---ey-survey-reveals-consumers-are-not-willing-to-share-data

[32] https://support.google.com/ads/answer/2662922?hl=en-GB
[33] https://duckduckgo.com/
[34] http://briandoddonleadership.com/2011/05/22/is-a-squirrel-dying-in-your-front-yard-more-important-than-people-dying-in-africa/
[35] https://redditblog.com/2008/06/17/reddit-goes-open-source/
[36] https://medium.com/hacking-and-gonzo/how-reddit-ranking-algorithms-work-ef111e33d0d9#.qwf2u2yxr
[37] https://redditblog.com/2008/06/17/reddit-goes-open-source/
[38] http://mashable.com/2008/06/18/reddit-goes-open-source-takes-aim-at-diggs-shady-algorithm/#llJxGrSQsiqL
[39] https://www.branded3.com/blog/reddit-just-made-their-algorithm-open-source/
[40] https://github.com/reddit/reddit
[41]https://www.reddit.com/r/announcements/comments/5gvd6b/scores_on_posts_are_about_to_start_going_up/
[42] https://www.oii.ox.ac.uk/should-there-be-a-better-accounting-of-the-algorithms-that-choose-our-news-for-us/
[43] http://en.unesco.org/news/privacy-expert-argues-algorithmic-transparency-crucial-online-freedoms-unesco-knowledge-cafe?language=en
[44] https://www.nytimes.com/2015/07/10/upshot/when-algorithms-discriminate.html?_r=0
[45] http://bigthink.com/videos/oliver-luckett-on-facebook-algorithms-and-online-echo-chambers
[46] https://iapp.org/news/a/algorithmic-transparency-examining-from-within-and-without/
[47] https://www.ftc.gov/public-statements/2015/12/transparency-trust-consumer-protection-complex-world-keynote-address
[48] https://ico.org.uk/about-the-ico/news-and-events/news-and-blogs/2016/09/transparency-trust-and-progressive-data-protection/