UnBias*

Emancipating Users against Algorithmic Biases for a Trusted Digital Economy

Ansgar Koene†
Horizon Digital Economy
University of Nottingham
Nottingham, UK
ansgar.koene@nottingham.ac.uk

Liz Dowthwaite
Horizon Digital Economy
University of Nottingham
Nottingham, UK
liz.dowthwaite@nottingham.ac.uk

Giles Lane
Proboscis
London, UK
giles@proboscis.org.uk

Helena Webb
Computer Science
University of Oxford
Oxford, UK
helena.webb@cs.ox.ac.uk

Virginia Portillo
Horizon Digital Economy
University of Nottingham
Nottingham, UK
virginia.portillo@nottingham.ac.uk

Marina Jirotka
Computer Science
University of Oxford
Oxford, UK
marina.jirotka@cs.ox.ac.uk

ABSTRACT

UnBias is a collaboration between the Horizon Digital Economy Research Institute at the University of Nottingham, the Human Centred Computing group at the University of Oxford and the Centre for Intelligent Systems and their Applications (CISA) at the University of Edinburgh, with creative studio Proboscis. It is funded under the Trust, Identity, Privacy and Security (TIPS) programme of the UK's Engineering and Physical Sciences Research Council (EPSRC). In this paper we introduce one of our key project outputs; a fairness toolkit which aims to promote awareness and stimulate a public civic dialogue about how algorithms shape online experiences. It will also prompt reflection on possible changes to address issues of online unfairness.

CCS CONCEPTS

• Information systems ~ Internet communications tools • Information systems ~ Social networks

KEYWORDS

Algorithms, Social Media, Bias, Fairness

ACM Reference format:

Ansgar Koene, Liz Dowthwaite, Giles Lane, Helena Webb, Virginia Portillo, and Marina Jirotka. 2018. UnBias: Emancipating Users Against Algorithmic Biases for a Trusted Digital Economy. In *KDD 2018. ACM, New York, NY, USA, 2 pages.* https://doi.org/10.1145/1234567890

*Article Title Footnote needs to be captured as Title Note
†Author Footnote to be captured as Author Note

KDD 2018, August, 2018, London UK

1 Introduction

UnBias is a collaboration between the HORIZON Digital Economy Research institute at the University of Nottingham, the Human Centred Computing group at the University of Oxford and the Centre for Intelligent Systems and their Applications (CISA) at the University of Edinburgh, with creative studio Proboscis. It is funded under the Trust, Identity, Privacy and Security (TIPS) programme of the UK's Engineering and Physical Sciences Research Council (EPSRC).

The project consists of four interconnected Work Packages (WPs) aligned with our four action objectives:

1. production of citizen education materials about information filtering/recommendation algorithms;
2. development of software tools for identifying biases in filtering/recommendation algorithms;
3. design recommendations for 'fair' algorithms that are subjectively experienced as unbiased;
4. policy briefs for an information and education governance framework capable of responding to the changing landscape in social media usage.

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

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

WP3 [led by Oxford] studies design requirements for filtering/recommender algorithms that satisfy subjective criteria of bias avoidance based on interviews and observation of users' sense-making behaviour when obtaining information through social media;

WP4 [co-led by Oxford and Nottingham] explores policy dimensions, including information and education governance frameworks for responding to the demands of

the changing landscape in the usage of algorithmic decision and recommendation systems. These are developed through broad stakeholder focus groups including representatives of industry, regulators, third-sector organizations, educators, lay-people and young people (a.k.a. "digital natives").

2 The Fairness Toolkit

One of our project outputs is a "Fairness Toolkit" which aims to promote awareness and stimulate a public civic dialogue about how algorithms shape online experiences and to reflect on possible changes to address issues of online unfairness. The tools are not just for critical thinking, but for *civic thinking* – supporting a more collective approach to imagining the future in contrast to the individual atomising effect that such technologies often cause. The toolkit has been co-created with young people and stakeholders and consists of three main parts (Fig. 1):

2.1 Awareness Cards

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

2.2 Trustscapes

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

2.3 Metamaps

This is a poster for stakeholders in the ICT industry, policymaking, regulation, public sector and research to respond to the young people's TrustScapes. By selecting and incorporating a TrustScape from those shared online, stakeholders can respond to the young people's perceptions. MetaMaps will also be captured and shared online via UnBias social media to enhance the public civic dialogue, and demonstrate the value of participation to young people in having their voice listened and replied to.
The Fairness Toolkit also includes "value perception" worksheets to help participants and stakeholders critically assess and evaluate the value of participation.

3 Conclusion

Going forward the UnBias project will continue to develop and refine our tools for engaging the technical and non-technical community on the issues of unintended/unjustified bias in algorithmic decision systems, and work with our stakeholder group to produce updated recommendations for the design and regulation of such systems. Input and engagement with the broad stakeholder community involved in the design, development and deployment of algorithmic systems remains a key component of our work.

ACKNOWLEDGMENTS

This work forms part of the UnBias project, funded by EPSRC grant EP/N02785X/1 and based at the Horizon Digital Economy Research Institute, University of Nottingham.

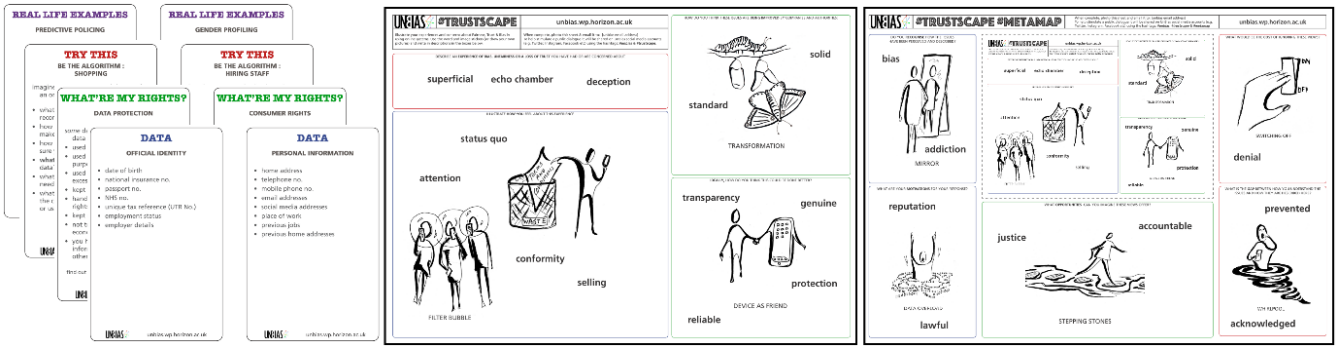

Figure 1: Illustration of the three tools in the UnBias fairness toolkit