Young People's Policy Recommendations on Algorithm Fairness

Elvira Perez Vallejos
Horizon Digital Economy Research Institute, University of Nottingham
Triumph Road Nottingham, UK NG7 2TU
Elvira.perez@nottingham.ac.uk

Ansgar Koene
Horizon Digital Economy Research Institute, University of Nottingham
Triumph Road Nottingham, UK NG7 2TU
Ansgar.Koene@nottingham.ac.uk

Virginia Portillo
Horizon Digital Economy Research Institute, University of Nottingham
Triumph Road Nottingham, UK NG7 2TU
Virginia.Portillo@nottingham.ac.uk

Liz Dowthwaite
Horizon Digital Economy Research Institute, University of Nottingham
Triumph Road Nottingham, UK NG7 2TU
Liz.Dowthwaite@nottingham.ac.uk

Monica Cano
Horizon Digital Economy Research Institute, University of Nottingham
Triumph Road Nottingham, UK NG7 2TU
Monica.Cano@nottingham.ac.uk

ABSTRACT

This paper explores the policy recommendations made by young people regarding algorithm fairness. It describes a piece of ongoing research developed to bring children and young people to the front line of the debate regarding children's digital rights. We employed the Youth Juries methodology which was designed to facilitate learning through discussions. The juries capture the deliberation process on a specific digital right, the right to know how algorithms govern and influence the Web and its users. Preliminary results show that young people demand to know more about algorithms, they want more transparency, more options, and more control about the way algorithms use their personal data.

CCS CONCEPTS

• **Social and Behavioral Sciences** → **Psychology** • **Computers and Education** → Computer and Information Science Education; *Literacy* • **Computers and Society** → Public Policy Issues.

KEYWORDS

Youth jury; algorithm fairness; youth opinion; deliberation; digital literacy; digital citizenship, privacy; policy.

WebSci'17, June 25-28, 2017, Troy, NY, USA.

ACM 978-1-4503-4896-6/17/06...$15.00.
DOI: http://dx.doi.org/10.1145/3091478.3091512

1 INTRODUCTION

The UnBias Youth Juries are similar to focus groups designed around the principles of deliberation [1-3]. In this paper we define the deliberation process as a series of steps that allow young people to receive and exchange information, to critically examine an issue, and to come to an agreement which will inform decision making. While there is a considerable amount of literature that documents the efficacy of deliberation within educational settings [4], hardly any systematic research has been conducted on the ways in which children and young people deliberate about their digital rights. Even more scarce are engaging educational interventions which aim to promote digital literacy, aside from the current e-safety programs that have broadly been introduced at primary and secondary schools.

To work in equal partnership with children and young people has been crucial in developing the youth juries. Co-production ensures that the scenarios (i.e., stimuli or prompts) represent real issues and experiences that young people can relate to. As a consequence, scenarios are idiosyncratic and sensitive to cultural differences as they should represent a specific and distinct point in time, avoiding universalistic terms. As smart phone applications, computer games and the lexicon around technologies rapidly evolve with time, the scenarios developed for this first wave of UnBias Youth Juries will differ from those that will be developed in the near future. Working with young people as equal partners is also important to guarantee that the language used to facilitate

the juries resonates with their vocabulary and expressions.

The structure and content of these juries is dynamic and changes from jury to jury to accommodate the uniqueness of each group but, as constant variables, the juries usually include an ice-breaking exercise and a group exploration around the definition of algorithms; what are they? (e.g., a series of steps and rules, a predictive mathematical formula), why are they useful? (e.g., to rank or filter large amounts of data), any benefits? (e.g., objectivity) any risks? (e.g., biased inferences). Once the context of the jury has been set up, the facilitator introduces some balanced facts about the way algorithms can affect Web users, for example, by making decisions on the user's behalf (e.g., ranking of newsfeed content) and how personalisation algorithms may influence outcomes from specific search engines (e.g., DuckDuckGo vs. Google). These prompts generate discussions among the jurors, providing opportunities for sharing of personal experiences and learning through conversation [5]. A series of different scenarios are then presented to the jurors as evidence of how algorithms can affect Web users. For example, one of the scenarios illustrates how algorithms are being used in the criminal justice system to predict patterns such as the likelihood of reoffending after release (e.g., Northpointe [6]). The data that feeds this algorithm can include personal data such as postcode, ethnicity, income, and so on. A second scenario describes the role of algorithms linked to Facebook's News Feed and how they track each user's online actions to serve them the posts they are most likely to engage with. These scenarios are packed with dilemmas that trigger discussions and reflections. We are interested in understanding the process of deliberation and opinion formation (e.g., argument and counterargument) and how the jurors may arrive at a consensus.

2 UNBIAS YOUTH JURIES

2.2 Methods

All participants (12 females and 14 males; average age: 16) were self-selected and were recruited via the website of the National Video Game Arcade (i.e., GameCity) in Nottingham, UK, where the juries took place. A large community room in GameCity was hired. The room layout included cabaret style for the first half of the session and circle or round style for the second half with the view to promote participation. In total, two jury sessions were organized, one in the morning and one in the afternoon. Each session had a duration of 2 hours including a break for refreshments. These sessions were part of a feasibility study designed to capture preliminary data to inform a posterior pilot study. Juries were audio recorded and data transcribed for subsequent qualitative analysis. Thematic analysis [7] was carried out by the author VP and transcripts reviewed alongside the analytical frame by all team members to ensure consistency of analysis. Ethical approval was granted by the Ethics Committee at the School of Computer Science at the University of Nottingham, UK. The youth jury methodology is described in detail elsewhere [8, 9].

3 RESULTS AND DISCUSSION

After the ice-breaking exercises, open-ended questions about algorithms were presented to the group to collectively arrive at a consensus about the meaning given to the label 'algorithm'. The qualitative analysis showed that, even though young people had a basic understanding about the conceptual meaning of algorithms:

"[…] a series of steps you take to derive an output from an input"

"a mathematical formula"

"a piece of code"

Jurors were less aware about the constant presence of algorithms controlling the Web and their influence on how search results were displayed. It is also important to highlight that the distinction between 'search engines' and 'browsers' was not clear, often referring to both as 'Google'.

The qualitative analysis generated five themes; online identity, security, privacy, concerns due to lack of transparency and control. In order to protect their online identity many jurors declared the use of pseudonyms, browse in private with incognito mode or carefully design the privacy settings of certain online services to make a distinction between their public vs. private profile.

In general, jurors were aware that likes, clicks and text entered in search engines and social media platforms could influence personalisation algorithms (i.e., recommender systems), however, most of them were unaware of the scale of personal data sharing (e.g., third parties) or the type of other information that could influence recommendation systems (e.g., likes from friends).

"Well, I've noticed with Twitter, like my friends are politically opinionated, and I get recommendations from

politicians and stuff they like [...] so probably they [Twitter] see what I follow and [...] I get recommendations based on that"

In general, jurors were aware of their privacy and anonymity options when accessing Web services and, as boyd and Marwick have already pointed out [10], young people care about their online privacy and develop strategies for managing privacy in public spaces in an attempt to assert control over their personal data. Some participants expressed location as being a more sensitive form of personal data that often was not disclosed. In general, jurors were less aware of the extent in which companies owned and traded with their personal data. While some jurors would prefer that there were no tracking systems at all, others felt that exceptions for surveillance would be justified to minimise crime when there are risk indicators:
"I generally don't use my actual name if I do so it would be pretty hard to find me [...] It was kind of a way to stop my friends bugging me."

"I don't really post anything publicly. I would privately do stuff so that not just anybody can see what I've been doing."

"I never really post 'right now I am here'... I wouldn't post where I am all the time"

"…. as long as you don't say certain things like: I live in Nottingham…., I am actually at…, you can say things to people and you don't generally need to worry about it being tied up to your name, ….you don't know who each other is"

"[…] I don't want people exploiting my personal enjoyment for the sake of making money."

In general, jurors appreciated the usefulness and need for algorithms:

"I do understand why people do it [use algorithms], because otherwise you have to see thorough lots of information".

"I think they [personalisation algorithms] are useful because it can keep you updated on things you like".

"[…] within the entertainment sector and stuff they [algorithms] are not that bad […]".

However, sometimes jurors found personalisation "annoying" and tended to ignore outcomes from recommender systems if they were inaccurate or far from their own preferences:

"[…] when I am searching for my music I can't ever find anything I want to find because they [service providers] are trying to tailor it to me but they are getting it wrong, totally wrong, so it's a bit annoying".

"[…] if you watch Netflix a lot for example, and you like certain shows it will give you recommendations of things you should like, but my brother went into my Netflix profile and added a bunch of Anime on to it I didn't watch and ruined my profile… I don't know now I only get Anime recommendations and it would take me ages to undo it so my recommendations don't mean anything anymore."

Most jurors acknowledged not knowing how recommender systems or search engines ranked their results and expressed concerns about the potential for censorship and bias. In general, algorithms were seen as neutral tools but there were concerns about the hidden purpose or intention (i.e., outcome value) and the consequences of 'echo chamber' or 'filter bubble' effects:

"I think that using an algorithm to sort things isn't necessarily a problem. What would be a problem is well, it's what it's actually being selected for."

"It's not just a matter of… what do you think they [industry] are going to do with it [personal info], it's also the information they give back to you. It's probably tailored to your interest. So, you might not see certain things and that could be pretty damaging too."

Jurors demanded plausible solutions to provide users with more options and control over their personal data. In general, jurors expressed a preference for data security over privacy and preferred signing up to Web services that were perceived as more secure and less likely to be hacked (e.g., Google vs. Yahoo) at the cost of their personal data (aware or inadvertently) being shared with third parties. Participants also deliberated on the effects that data protection breaches, from established companies, could have on their corporate reputation and anticipated that large businesses would continue trading with users' personal data without worrying about it or upsetting users.

"Google and Facebook have better infrastructure to handle security than smaller companies"

"Established companies CAN afford to upset users"

When asked about who should regulate fairness policy and ethics guidelines, jurors were unsure of who should take on this important governance work. They argued that a global approach would be ideal, but they were not confident about an international framework due to the need to accommodate so many cultural differences and attitudes to data privacy across countries. They also reasoned that NGOs or independent bodies could lack the resources needed to cope with fast changes, while expressing frustration and lack of trust towards large corporations, which were perceived as powerful entities that could influence countries' economy.

A limitation of the youth jury's methodology is the potential bias that the facilitator inadvertently could introduce either within the examples or facts introduced during the sessions. There are other, more subtle, biases that could also affect the outcome or direction of the discussion including body language and facial expressions that could favor one position or argument more than another. The research team was aware of this limitation and special attention was given to highlighting the neutral nature of algorithms and minimising undesirable effects that could contaminate the juror's arguments and counterarguments illustrating the risks and opportunities embedded in algorithms. How the jury was delivered and implemented was also important to the research team, not only because the juries should be replicable and participants' outputs should not depend on the personal attributes of the facilitator, but because explicit training, guidelines, and processes were in place, and a sense of ownership, responsibility, and care were also part of the training. For example, understanding the current evidence on algorithm fairness and potential sources of biases was important to ensuring that accurate and factual information was discussed during the deliberation process with the view to minimise any potential biases. The same person (lead author EPV) facilitated both juries which resulted in similar outcomes.

4 CONCLUSIONS

Jurors put forward several solutions and recommendations such as plug-ins to add user-friendly interfaces in which users could decide levels of tracking, more control over personal data and ways to influence their outcome (e.g., what exactly is being stored, who is storing it and for how long), or how to combine results from different search engines and compare results depending on users' priorities. There was a consensus that decision-making should not be left entirely to an automated system, especially when the decision had important consequences for the users (e.g., job recruitment). Unanimously, jurors asked for more accessible Terms & Conditions and accessible information about the way algorithms rule the Web. It was agreed that more engaging educational programs and increased knowledge would be beneficial not only for young people but for parents and IT teachers.

The results from this feasibility study have informed a posterior pilot study that, to date, has recruited more than 100 young people. The concerns and recommendations that the jurors of these two preliminary juries have put forward have been very valuable to (1) improve the co-design of scenarios making them more relevant and engaging and (2) identify topics for discussion and reflection (i.e., personalisation through algorithms, autocomplete, search results, fake news and algorithm transparency).

ACKNOWLEDGMENTS

This work was supported by EPSRC Trust, Identity, Privacy and Security grant "UnBias: Emancipating users against algorithmic biases for a trusted digital economy" (EP/N02785X/1)

REFERENCES

[1] Chris Degeling, Lucie Rychetnik, Jackie Street, Rae Thomas, Stacy M Carter 2017. Influencing health policy through public deliberation: Lessons learned from two decades of Citizens'/community juries. *Social Science and Medicine*, 179, 166-171. DOI: http://dx.doi.org/10.1016/j.socscimed.2017.03.003

[2] Stephanie Solomon, Julia Abelson. 2013. Why and when should we use public deliberation? Hastings Cent Rep, 42, 2, 17-20. DOI: 10.1002/hast.27

[3] Julia Abelson, Pierre G. Forest, John Eyles, Patricia Smith, Elisabeth Martin, Francois P. Gauvin. 2003. Deliberations about deliberative methods: issues in the design and evaluation of public participation processes. *Soc. Sci. Med.*, 57, 2, 239–251. DOI http://dx.doi.org/10.1016/S0277-9536(02)00343-X

[4] Timothy J. Shaffer. 2014. Deliberation In and Through Higher Education. *Journal of Public Deliberation*, 10, 1, 10. http://www.publicdeliberation.net/jpd/vol10/iss1/art10

[5] Elvira Perez Vallejos, Deborah Hencker, Dick Churchill. 2016. Video Interactive Guidance (VIG): a reflective pedagogical tool for enhancing learning goals and compassion in the context of Clinical Communication Skills education, 360-362. In Peterkin, A. & Brett-MacLean, P. (Eds.) *Keeping Reflection Fresh*. Kent State Press

[6] Julia Angwin, Jeff Larson, Surya Mattu, Lauren Kirchner, L. 2016. Machine Bias. There's software used across the country to predict

future criminals. And it's biased against balcks. Online article published at Pro Publica. Available https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[7] Virginia Braun, Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3 (2): 77-101.

[8] Stephen Coleman, Kruake Pothong, Elvira Perez Vallejos, Ansgar Koene. 2017. The Internet on Trial: How children and young people deliberated about their digital rights. Report available at http://casma.wp.horizon.ac.uk/wp-content/uploads/2016/08/Internet-On-Our-Own-Terms.pdf

[9] Elvira Perez Vallejos, Ansgar Koene, Chris J. Carter, Ramona Statache, et al. 2016. Juries: Acting Out Digital Dilemmas to Promote Digital Reflections. *ACM SIGCAS Computers and Society* 45, 3, 84-90. DOI 10.1145/2874239.2874252

[10] boyd, danah and Marwick, Alice E., Social Privacy in Networked Publics: Teens' Attitudes, Practices, and Strategies (September 22, 2011). A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society, September 2011. Available at SSRN: https://ssrn.com/abstract=1925128