Explainable Artificial Intelligence: Human-centered Perspective

Hana Kopecká
hana.kopecka@kcl.ac.uk
King's College London
London, United Kingdom

CCS CONCEPTS

• **Social and professional topics → Cultural characteristics**; • **Human-centered computing** → *Collaborative and social computing*; • **Computing methodologies → Artificial intelligence**.

KEYWORDS

explainable artificial intelligence, human-centered artificial intelligence, socio-cultural background

ACM Reference Format:
Hana Kopecká. 2021. Explainable Artificial Intelligence: Human-centered Perspective. In *Proceedings of WomENcourage.* ACM, New York, NY, USA, 2 pages.

1 INTRODUCTION

Explainable artificial intelligence (XAI) has attracted a lot of attention in the recent years as the artificial intelligence (AI) community recognises a need for more transparent and human-centered systems. At the same time, AI field is facing a 'diversity crisis' in terms of demographics of AI researchers [11]. Could this diversity crisis result in explainable AI systems, that are geared towards a certain community and leaving out others, that are underrepresented or missing in the community of AI designers?

In our work, we bring together insights from cognitive psychology, cognitive sociology and intersectional feminism to investigate if indeed one's social and cultural background, which is known to constitute one's perceptual and cognitive tendencies, also influences one's explanation needs in human-AI interaction. Our aim is to contribute to designing more inclusive and human-centered AI systems.

2 EXPLAINABLE ARTIFICIAL INTELLIGENCE

The field of Artificial Intelligence (AI) have made some impressive strides in the last decade in many areas. Notably in computer vision and natural language processing due to deep neural networks and the availability of big data. However, coupled with these spectacular achievements, some AI application caused serious concerns for exhibiting racists, sexist and otherwise ethically problematic behaviour. For example, the COMPAS software which was supposed to aid criminal justice professional in predicting offender's likelihood of recidivism, has been shown to identify significantly more black offenders as high risk, and white offenders as low risk at exhibiting recidivism [5]. On many occasion, such harmful AI systems, including COMPAS have been in use for many years without their detrimental outputs being known. Among other reasons, it was possible because these systems lacked explainability, hence their users could not access, monitor and oversee the internal decision making mechanism by which these systems reached their output.

The need for AI applications to be explainable is now widely accepted by the AI community and vigorous research is being conducted. There are several reasons for developing explainable AI systems, and these will depend on *who* is seeking the explanation [9]. Mohseni and colleagues [9] provide a comprehensive overview of different XAI design goals for three communities interacting with AI systems; novice users, data experts and AI experts. There is an understanding that users with different relationship with the AI systems would seek explanations for different reasons. For example, an AI expert might seek an explanation to debug the system, while AI novice might be interested in assessing whether the system is unbiased and ethical to prevent situations such as deployment of the COMPAS software. Considering a novice user, the design goals of explainable AI system are (1) AI transparency, (2) building user trust, (3) mitigating the risk of bias and to help the user to (4) assess their data privacy [9]. In our work, we focus on novice users and we are interested in designing systems for the first two goals; that is a systems that is transparent and helps the user with understanding how it works, which improves the user interactions and user trust in the system.

The current literature on explainable AI recognises, that explanation is both a product and process, or in fact, two processes; a cognitive and social one [8]. The congnitive process is undertaken by the explainer, who determines the explanation for the event at hand by identitfying the causal chain leading to the event and selecting the important elements to be presented as explanandum. The explanandum is then the product of the cognitive process, which is transferred to the explainee (the recipient of the explanation) in a social process.

We argue, that there is another process which is crucial to consider in order to design useful and understandable AI explanations, and this is the cognitive process undertaken by the explainee, in which they receive and process the explanation [6] and create or refine their mental model of the AI system. We believe that by putting the user in the centre and consider their needs first, we can better inform how to design the explanandum and the social process of communicating the explanandum. In order to do that, we have to better understand the users needs and natural cognitive tendencies.

Now, we are not the first ones to suggest starting from the user, there is an existing XAI research based on understanding how people engage in explanations [7, 8] following scholarship in social sciences, and the HCI community. We, however, take a novel approach in designing human-centered AI explanations. The insights from social sciences the AI community is building upon and often inspired by an influencial paper by Tim Miller [8], which provides

WomENcourage, 22-24 September 2021, Prague
2021.

and overview of basic principles of human explanations, which are assumed to be universal. Our goal is to is to go beyond these universal principles and explore more nuanced differences in explanations between different communities.

Our research is informed by two traditions. (1) Cognitive psychology and cognitive sociology suggest that perception and cognition are shaped by ones culture and as a result, user's AI explanation preferences might be influenced by their cognitive tendencies. We also follow (2) critical social theory and intersectional feminist tradition in trying to understand the structures of power, in which explainable AI systems are embedded and through which power is realised and reproduced.

3 CULTURE, COGNITION AND PERCEPTION

In the previous section, we established that we can break the process of explanation into three processes; cognitive process on the part of the explainer, social process of transferring the information between explainer and explainee and a cognitive process of the explainee. In our work, we focus on the last process, which is the user's cognitive process. In particular, we focus on variability in perceptual and cognitive style as a result of socialisation. We explore, to which extend these differences affect user's interaction with AI systems and explanation preferences.

There is ample evidence suggesting that the mode of perceiving and thinking is shaped by one's culture, which determines how people experience and interpret the world around them [2, 4, 10], but also how they use language [1] to communicate with others. These theories tell us that our upbringing in a particular social and cultural context instills perceptual and cognitive dispositions that equip us with lenses, through which we access the world and which determine what aspects of reality we 'see' and how we act upon them.

These theories that we mention focus on different manifestations of perceptual and cognitive style, such as the use of language [1], tastes [2] and casual attribution [10], but they all suggest that perceptive and cognitive tendencies are not universal, which is pertinent to both how we design the explanandum so that it fits well with the users cognitive tendencies, but also how we design the communication process of the information interchange between the AI agent and the user in a way that is easily accessible to the user.

It is therefore of a paramount importance to understand, how do these cultural differences play out when users of different background interact with AI systems.

4 INTERSECTIONAL FEMINISM

Now, there is a trend in XAI community to consider who the user is in order to provide a suitable explanation, but this effort does not go deeper beyond the level of expertise with AI systems or the domain of deployment [9]. Differences in cognitive and perceptual tendencies go mostly unacknowledged, which might rise questions regarding ethics and fairness of XAI systems, that do not recognise cognitive style variability and might be unknowingly optimised for specific users, while being suboptimal for some communitites. This is especially concerning given that the field of AI and leading AI companies in particular are infamous for their lack of diversity among their employees and even more so in their leadership. The paper 'Discriminating systems: Gender, Race, and Power in AI' published by the AI Now Institute reports that in 2019, only 20% of AI professionals globally were women, while these figures are even lower in leading tech corporations. Only 15% and 10% of AI researchers in Facebook and Google respectively are women and these companies fare even worse in terms of racial diversity; in Google, only 2.5% of full-time workers are black and 3.6% are latinx [11]. The lack of diversity in terms of the AI research community, but more importantly in terms of the imagined community of users AI explanations are designed for, we borrow tools from intersectional data feminism [3] to interrogate and understand the power relations AI systems are embedded in and reproduced by and to challenge the power by working towards more equitable XAI systems.

5 CONCLUSION

In our work, located on the intersection of cognitive psychology, cognitive sociology, intersectional feminism and artificial intelligence, we investigate the role of user's socio-cultral background on their cognitive style and ultimately, on their AI explanation needs.

6 ACKNOWLEDGMENTS

ACKNOWLEDGMENTS

This work was supported by UK Research and Innovation [grant number EP/S023356/1], in the UKRI Centre for Doctoral Training in Safe and Trusted Artificial Intelligence.

REFERENCES

[1] Basil Bernstein. 1958. Some Sociological Determinants of Perception : An Enquiry into Sub-Cultural Differences. *The British Journal of Sociology* 9, 2 (1958), 159–174.
[2] Pierre Bourdieu. 1984. *Distinction : a social critique of the judgement of taste.* Routledge & Kegan Paul London. xiv, 613 p. : pages.
[3] Catherine D'Ignazio and Lauren F. Klein. 2020. *Data Feminism.* The MIT Press, Cambridge.
[4] Geert H. Hofstede. 2001. *Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations* (2nd and enlarged ed.). Sage, Thousand Oaks, CA. xx, 596 pages.
[5] Julia Angwin Jeff Larson. 2021. How We Analyzed the COMPAS Recidivism Algorithm. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm [Online; accessed 28. Apr. 2021].
[6] Hana Kopecká and Jose Such. 2020. Explainable AI for Cultural Minds. In *European Conference on Artificial Intelligence (Workshop on Dialogue, Explanation and Argumentation for Human-Agent Interaction).* Accepted.
[7] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2019. A Grounded Interaction Protocol for Explainable Artificial Intelligence. Aamas (2019). arXiv:1903.02409 http://arxiv.org/abs/1903.02409
[8] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007 arXiv:1706.07269
[9] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2018. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. 1, 1 (2018), 1–37. arXiv:1811.11839 http://arxiv.org/abs/1811.11839
[10] Michael W. Morris and Kaiping Peng. 1994. Culture and Cause: American and Chinese Attributions for Social and Physical Events. *Journal of Personality and Social Psychology* 67, 6 (1994), 949–971. https://doi.org/10.1037/0022-3514.67.6.949
[11] Sarah Myers-West, Meredith Whittaker, and Kate Crawford. 2019. Discriminating systems: Gender, Race and Power in AI. AI Now Institute. https://ainowinstitute.org/discriminatingsystems.pdf