Algorithmic nudge: Using XAI frameworks to design interventions

Prerna Juneja
University of Washington
Seattle, WA,USA
prerna79@uw.edu

Tanushree Mitra
University of Washington
Seattle, WA,USA
tmitra@uw.edu

ACM CHI Workshop on Operationalizing Human-Centered Perspectives in Explainable AI, May 8–9, 2021, Virtual
ACM .

Abstract

In this position paper, we propose the use of existing XAI frameworks to design interventions in scenarios where algorithms expose users to problematic content (e.g. anti vaccine videos). Our intervention design includes *facts* (to indicate algorithmic justification of what happened) accompanied with either *fore warnings* or *counterfactual explanations*. While fore warnings indicate potential risks of an action to users, the counterfactual explanations will indicate what actions user should perform to change the algorithmic outcome. We envision the use of such interventions as 'decision aids' to users which will help them make informed choices.

Author Keywords

explainability, bias, transparency, intervention

AI algorithms play an important role in governing and shaping our lives. From recommending what websites to browse, what movies to watch, and what books to read, to informing decisions about defendants in the criminal justice system, these black-box algorithms play a crucial role in several low and high stake tasks. As the algorithmic systems become more pervasive, there has been a widespread concern about their role in amplifying or reinforcing various biases. Thus, researchers and scholars have pushed for making algorithms more accountable and transparent. This

push has propelled the field of explainable AI (XAI) where the goal is to provide an explanation on how the machine learning algorithm reached a particular decision [26]. The literature on XAI is vast (see [1] for a review). A few notable directions include explaining ML classifiers and multi-agent systems [2,4,13,29], designing guidelines for generating explanations [10, 21, 22], defining taxonomy of user needs for AI explainability [23, 24, 25], developing metrics and frameworks for evaluating explanations [7, 17] and identifying various stakeholders of XAI [28]. Another line of scholarly research has also tested the effectiveness of the explanations used by online social media platforms. For example, investigation of Facebook's *Why am I seeing this ad?* feature revealed that users often found the ad explanations misleading and incomplete [3]. There is also a burgeoning field of human-centered XAI where scholars draw from formal theories in HCI to inform the design of explanation interfaces [15, 16, 20, 27] and conduct empirical studies to examine how users interact with explanations [6, 8, 9, 18]. Existing literature reveals that XAI has the capability to not only explain the algorithmic decision making process but also provide a signal of how the algorithm will behave in future [30].

In this position paper, we explore use of existing XAI frameworks in designing persuasive interventions in scenarios where user behaviour can lead to exposure to problematic content. In other words, we want to explore whether feedback from algorithms (designed using XAI frameworks) improve human decision making?

We propose design using (1) facts (what happened and why) accompanied with forewarnings (what could happen) to convey the potential risks of an action in a comprehensible manner. For example, consider a situation where a user searches for query "election fraud proof" on YouTube. The message should then forewarn users about the risk of their future video recommendations getting polluted from election misinformation content and ask them to re-think about their actions. (2) facts (what happened and why) accompanied with counterfactual explanations (what needs to change for another outcome to appear). This design would include message informing users why they are seeing a problematic content along with instructions of how could they sanitize their own content to not see that problematic content. For example, consider a situation when an anti-vaccine video appears in user's homepage. The message explains why such an occurrence occurred *"This happened because..."* and remedies in a counterfactual tone *"For this not to happen do....."*. In this particular example, counterfactual solutions could include suggesting users to delete the video from their search and watch history to remove its influence on future recommendations [14] or to click on "not interested" to signal to YouTube that you are not interested in seeing this video.

Scholars have argued that explanations are not always necessary or desirable and explaining everything in every situation [19] and can lead to information overload [11]. Keeping this principle in mind, we propose our interventions in high stake scenarios where user actions have or could lead to more exposure of problematic content. We envision the use of such explanations as *decision aids* to help users make better choices. It is important to note that defining problematic content is out of scope for this position paper. For the purpose of this proposal, we consider two examples of problematic content. First where algorithmic output has a partisan bias and second where users are presented with misinformation. We include partisan bias in this category because scholars have argued that *selective exposure* to information from a specific ideology could lead to fragmented society [5].

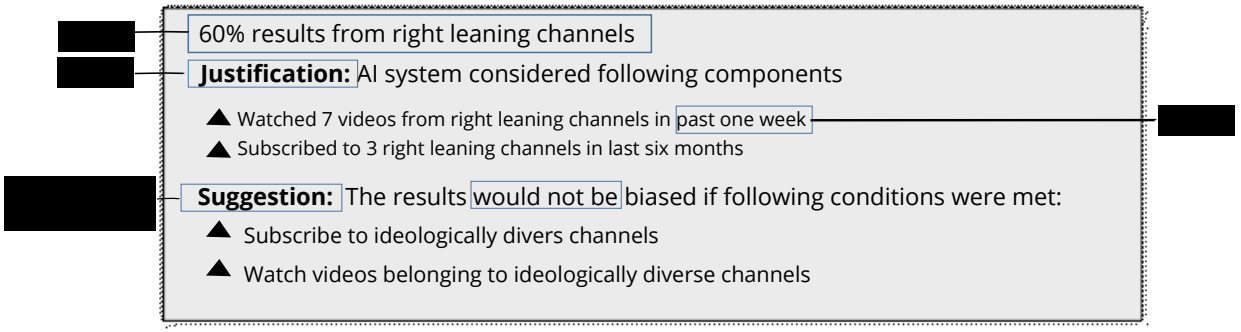

Figure 1: Scenario 1: When a user searches for a political query in YouTube and is presented with biased search results. Apart from algorithmic justification, user is also presented with actions to change the algorithmic output in future.

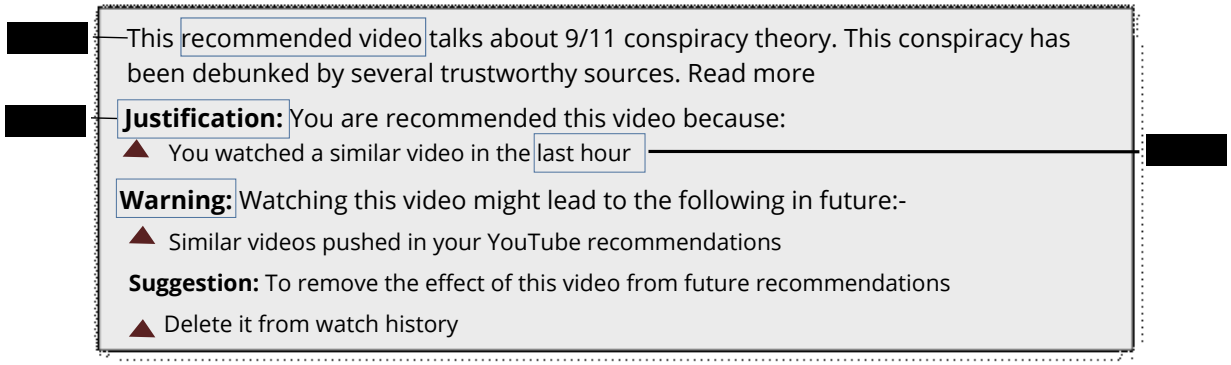

Figure 2: Scenario 2: When a video about 9/11 conspiracy theory appears in user's homepage. Here, apart from algorithmic justification, user is also warned about the consequences of watching this video in future.

We make use of two frameworks and theories to design our interventions. First we, use XAI design framework suggested by Ehsan et al based on the principles of Social Transparency that suggests use of design features reflecting the What, Why, Who, and When of user interactions with AI systems [11]. In our design, 'what' is conveyed by indicating a problematic behaviour and 'when' is expressed by a timestamp, 'why' is indicated in the algorithmic justification of 'what'. 'Who' is the user receiving these interventions, thus we do not explicitly mention that in the design. Next we make use of Fogg's behaviour change model (FGB) that has been used to design persuasive technologies [12]. The model states that for a user to change behaviour, they must be (1) motivated, (2) have the ability to perform the change and should be triggered to perform the change [12]. Based on this model, we provide 'explanations' as triggers, 'bias indicators' as motivation to change behaviour and also 'state the actions' that are required to change behaviour. We demonstrate the design of our XIA based interventions via few example scenarios in Figures 1 and 2. We plan to test the effectiveness of such a design using user studies.

There are several open questions that our proposed design does not address. What are the various high-stake problematic scenarios that demand algorithmic interventions? How frequently should such interventions appear? Where should these interventions appear? What granularity of algorithmic justification should appear in the design? What role will algorithm skepticism play in users acceptance or rejection of these interventions? We hope to discuss these questions during the workshop.

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Lim, and Mohan Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. 1–18. DOI:http://dx.doi.org/10.1145/3173574.3174156

[2] Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze. 2020. Evaluating saliency map explanations for convolutional neural networks: a user study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 275–285.

[3] Athanasios Andreou, Giridhari Venkatadri, Oana Goga, Krishna Gummadi, Patrick Loiseau, and Alan Mislove. 2018. Investigating ad transparency mechanisms in social media: A case study of Facebook's explanations. In *NDSS 2018-Network and Distributed System Security Symposium*. 1–15.

[4] Plamen Angelov and Eduardo Soares. 2020. Towards explainable deep neural networks (xDNN). *Neural Networks* 130 (2020), 185–194. DOI: http://dx.doi.org/https://doi.org/10.1016/j.neunet.2020.07.010

[5] Mahmoudreza Babaei, Juhi Kulshrestha, Abhijnan Chakraborty, Fabrício Benevenuto, Krishna P Gummadi, and Adrian Weller. 2018. Purple feed: Identifying high consensus news posts on social media. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 10–16.

[6] Carrie J Cai, Jonas Jongejan, and Jess Holbrook. 2019. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 258–262.

[7] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. 2019. Machine learning interpretability: A survey on methods and metrics. *Electronics* 8, 8 (2019), 832.

[8] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. 2019. Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.

[9] Jonathan Dodge, Q Vera Liao, Yunfeng Zhang, Rachel KE Bellamy, and Casey Dugan. 2019. Explaining models: an empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 275–285.

[10] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[11] Upol Ehsan, Q Vera Liao, Michael Muller, Mark O Riedl, and Justin D Weisz. 2021. Expanding Explainability: Towards Social Transparency in AI systems. *arXiv preprint arXiv:2101.04719* (2021).

[12] BJ Fogg. 2009. A Behavior Model for Persuasive Design. In *Proceedings of the 4th International Conference on Persuasive Technology (Persuasive '09)*. Association for Computing Machinery, New York, NY, USA, Article 40, 7 pages. DOI: http://dx.doi.org/10.1145/1541948.1541999

[13] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51, 5 (2018), 1–42.

[14] YouTube Help. accessed in 2020. Manage your recommendations and search results. (accessed in 2020). https://support.google.com/youtube/answer/6342839?co=GENIE.Platform%3DAndroid&hl=en

[15] Robert R Hoffman and Gary Klein. 2017. Explaining explanation, part 1: theoretical foundations. *IEEE Intelligent Systems* 32, 3 (2017), 68–73.

[16] Robert R Hoffman, Shane T Mueller, and Gary Klein. 2017. Explaining explanation, part 2: Empirical foundations. *IEEE Intelligent Systems* 32, 4 (2017), 78–86.

[17] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).

[18] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M Drucker. 2019. Gamut: A design probe to understand how data scientists understand machine learning models. In *Proceedings of the 2019 CHI conference on human factors in computing systems*.

[19] Beena Ammanath James Guszcza, Michelle A. Lee and Dave Kuder. 2020. Human values in the loop: Design principles for ethical AI. (2020). https://www2.deloitte.com/content/dam/insights/us/articles/6452_human-values-in-the-loop/DI_DR26-Human-values-in-the-loop.pdf

[20] Gary Klein. 2018. Explaining explanation, part 3: The causal landscape. *IEEE Intelligent Systems* 33, 2 (2018), 83–88.

[21] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*. 126–137.

[22] Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. 2016. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1675–1684.

[23] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.

[24] Brian Y Lim and Anind K Dey. 2009. Assessing demand for intelligibility in context-aware applications. In *Proceedings of the 11th international conference on Ubiquitous computing*. 195–204.

[25] Brian Y Lim, Anind K Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2119–2128.

[26] Zhong Lin, Mohammad Javad Shafiee, Stanislav Bochkarev, Michael Jules, Xiao Wang, and Alexander Wong. 2019. Explaining with Impact: A Machine-centric Strategy to Quantify the Performance of Explainability Algorithms. (10 2019).

[27] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.

[28] Alun Preece, Dan Harborne, Dave Braines, Richard Tomsett, and Supriyo Chakraborty. 2018. Stakeholders in explainable AI. *arXiv preprint arXiv:1810.00184* (2018).

[29] Avi Rosenfeld and Ariella Richardson. 2019. Explainability in human–agent systems. *Autonomous Agents and Multi-Agent Systems* 33, 6 (2019), 673–705.

[30] Matt Turek. accessed in 2020. Explainable Artificial Intelligence (XAI). (accessed in 2020). `https://www.darpa.mil/program/explainable-artificial-intelligence`