Opinion

Social construction of XAI: Do we need one definition to rule them all?

Upol Ehsan1,2 and Mark O. Riedl1

1Georgia Institute of Technology, Atlanta GA, USA
2Correspondence: ehsanu@gatech.edu

https://doi.org/10.1016/j.patter.2024.100926

In this opinion, Upol Ehsan and Mark Riedl argue why a singular monolithic definition of explainable AI (XAI) is neither feasible nor desirable at this stage of XAI’s development.

There is a growing frustration among researchers and developers in explainable AI (XAI) around the lack of consensus around what is meant by “explainability”. Do we need one definition of explainability to rule them all? In this paper, we view XAI through the lenses of Social Construction of Technology (SCOT) to explain how diverse stakeholders (relevant social groups) have different interpretations (interpretative flexibility) that shape the meaning of XAI. Forcing a standardization (closure) on the pluralistic interpretations too early can stifle innovation and lead to premature conclusions. We share how we can leverage the pluralism to make progress in XAI without having to wait for a definitional consensus.

Of bicycles & explainable AI

A lack of consensus on the definition of explainability is often portrayed as a problematic aspect in XAI. However, do we need one definition of XAI to rule them all? Considering the current developmental stage of XAI as a field, we argue that it is neither feasible nor desirable to have a singular definition of XAI. To understand why that is the case, we need to first talk about bicycles and analyze their evolution over the last 200 years through the lenses of SCOT, a theory in science and technology studies that explains how human action foundationally gives shape and function to technology (for a full treatise of SCOT, refer to Bijker1).

The word “bicycle” may evoke a prototypical image in your head, one with two wheels, pedals, a handlebar with brake levers, and a seat offset from the rear wheel. Despite being one of the most stable pieces of technology, did you know that bicycles evolved rather drastically? For instance, the first bicycles lacked brakes by design because young men of “means and nerves” wanted that thrill of controlling a dangerous object while rolling down the English countryside. In the 1800s, the famous penny-farthing bicycle had a huge front wheel with a smaller back wheel. Many bicycles had fixed handlebars that prevented turning.2 As we reflect on the evolution of the bicycle, why and how did things evolve the way they did?

We will address this question using three concepts from SCOT. First, we have relevant social groups—stakeholders with skin in the game such as bikers, families of bikers, mechanics fixing bikes, etc. These are the ones who are involved in or affected by a technological development. Different relevant social groups have their own interpretive flexibility—interpretations of what it means to be a bicycle. Different interpretive flexibilities can give rise to different types of bicycles such as mountain bikes, electric bikes, BMX bikes, etc. Finally, we have the notion of closure—over time, some interpretations of the bicycle achieved stability while others withered out (e.g., equal-sized wheels won out over differently sized wheels). This is how social aspects can construct the technology (of bicycles).

How does SCOT help with XAI? Just like bicycles, XAI has its relevant social groups. Many communities of practice are involved in XAI. Let’s consider two relevant social groups: the natural language processing (NLP) and computer visions (CV) communities. Given each group has its own ways of knowing (epistemology), there is interpretive flexibility on how they operationalize the notion of explainability. For instance, in NLP question answering, explanations are often of the form of additional text that justifies the ground truth answer.3 In CV, object recognition can consider saliency maps that show how visual features correlate to a predicted label.4 Given different operationalizations, their evaluations can also be different. This is a major reason behind the current pluralistic notions of XAI. There is also no universal consensus on what methodology should be applied in evaluating XAI methods.5 This is to be expected because, unlike bicycles, we don’t have 200+ years of development to reach clusters of closures yet.

Even though we do not have full closures in our interpretations of explainability, as a community there is a general consensus that our shared goal is to make an AI system’s functioning or decisions understandable by people such that they can take informed actions.6 There is a growing communal awareness that explainability is more than algorithmic transparency. That is, solving XAI challenges may require more than just opening the black box. For example, human-centered XAI (HCXAI) advocates to tackle XAI problems through a sociotechnical view (vs. a purely technical one).7 We need to consider who is opening the box just as much as the algorithmic mechanisms of opening it. Different “whos” need different explanatory facets to understand the AI system’s decision making.7,8 A strength of the SCOT-based view is that definitional relativism is addressed by closure where in due time certain facets of the interpretation get established as the “norm”—e.g., over time, certain interpretations of “mountain bikes” reached closure from an initial state of varied interpretive flexibility from different relevant groups. Whereas a lot of initial focus was on developers and
data scientists as end users of XAI systems, there is a growing recognition that we need to accommodate a diverse set of end users, especially non-AI experts. If we had forced closure on what we mean by explainability too early, a lot of the emerging work would not have the intellectual space to flourish.

Cycling onward & making progress in XAI

So what does all of this mean?

1. XAI is pluralistic. Given the different epistemic cultures co-existing in the space, we cannot expect monolithic conformity at this stage.

2. Pluralism, however, does not mean that anything goes; in fact, it’s the opposite. Instead of definitional relativism, we need to be precise in our articulation of what we mean by explainability when we communicate. Thus, instead of using the term at face value, whenever we write a paper, we should start with “who” the relevant stakeholders are, their explainability needs, and justify how a particular conception of explainability satisfies the shared goals of the relevant social group.

3. We should use SCOT to understand the sociology of XAI—who is saying what, when, and why. To grasp the flavor of explainability in a given context, we need to pay attention to a relevant social group’s interpretation of it and how that informs their operationalization.

4. While the notion of XAI is in flux, we are fortunate to join the conversation at this stage. We have substantial agency in steering the discourse, a privilege we need to exercise responsibly.

A SCOT-based view helps us understand the sociological dynamics behind the pluralism in XAI, empowering us to focus on what XAI can do for us instead of obsessing over definitional supremacy. It affords progress while simultaneously refining our shared understanding of the concept. It does not mean that we are advocating for epistemological anarchy or definitional relativism.

If AI research could come this far without a monolithic consensus on the meaning of “intelligence”, XAI can make progress without being bogged down in a definitional tug of war.

ACKNOWLEDGMENTS

With our deepest gratitude, we thank discussants at the Human-centered XAI workshop at CHI where the initial ideas were developed. We are grateful to members of the human-centered AI lab at Georgia Tech for their feedback on the ideas. We are indebted to Michael Muller, Vera Liao, Elizabeth Watkins, and Samir Passi for their feedback on the early discussions. Special thanks to Rachel Urban for generously providing proofreading feedback. This project was partially supported by the National Science Foundation under Grant no. 1928586.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Bijker, W.E. (1997). Of Bicycles, Bakelites, and Bulbs: Toward a Theory of Sociotechnical Change (MIT press).
2. Hadland, T., and Lessing, H.E. (2016). Bicycle Design: An Illustrated History (Mit Press).
3. Wiegreffe, S., Hessel, J., Swayamdipta, S., Riedl, M., and Choi, Y. (2022). Reframing Human-AI Collaboration for Generating Free-Text Explanations. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 632–638.
4. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision, pp. 618–626.
5. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. (2018). Sanity checks for saliency maps. Adv. Neural Inf. Process. Syst., 31.
6. Ehsan, U., and Riedl, M.O. (2020). Human-centered explainable ai: Towards a reflective sociotechnical approach. In International Conference on Human-Computer Interaction (Springer), pp. 449–466.
7. Ehsan, U., Passi, S., Liao, Q.V., Chan, L., Lee, I.H., Muller, M., and Riedl, M.O. (2021). The who in explainable ai: How ai background shapes perceptions of ai explanations. Preprint at arXiv. https://doi.org/10.48550/arXiv.2107.13509.
8. Dhanorkar, S., Wolf, C.T., Qian, K., Xu, A., Popa, L., and Li, Y. (2021). Who needs to know what, when?: Broadening the Explainable AI (XAI) Design Space by Looking at Explanations Across the AI Lifecycle. In Designing Interactive Systems Conference 2021, pp. 1591–1602.

About the authors

Upoit Ehsan makes AI systems responsible and explainable so that people who aren’t at the table don’t end up on the menu. He is a doctoral candidate at Georgia Tech and an affiliate at Data & Society. His work has pioneered the domain of human-centered explainable AI (HUXAI), receiving awards at top-tier venues like ACM CHI, HCII, and AAAI AIES. Serving on multiple government-level AI task forces, his work has been covered in major media outlets. He led the editorial efforts for the inaugural ACM TiIS Journal on HUXAI and spearheads the organization of the HUXAI workshops at CHI since 2021.

Dr. Mark Riedl is a professor in the Georgia Tech School of Interactive Computing and associate director of the Georgia Tech Machine Learning Center. Dr. Riedl’s research focuses on human-centered artificial intelligence—the development of artificial intelligence and machine learning technologies that understand and interact with human users in more natural ways. Dr. Riedl’s recent work has focused on story understanding and generation, computational creativity, explainable AI, and teaching virtual agents to behave safely.