Michael Ridley

Explainable AI

An Agenda for Explainability Activism

If artificial intelligence (AI), particularly generative AI, is an opaque "black box," how are we able to trust it and make the technology accountable? Academic libraries are evaluating, providing, using, and increasingly building AI-based information tools and services. Typically, the underlying models for these intelligent systems are large language models (LLMs) based on generative AI techniques. While many of these systems have shown remarkable advances and advantages, the risks and deficiencies are also widely known and easily demonstrated.[1]

One path to trust and accountability is through explainability: the capacity of AI tools and services to explain their outcomes, recommendations, or decisions. Academic libraries need to adopt a multifaceted approach to explainability to ensure intelligent systems meet (and hopefully exceed) our expectations for authority, credibility, and accuracy.

Why Explainability?

Tania Lombrozo underscores the importance of explanations, noting that explanations "are more than a human preoccupation—they are central to our sense of understanding, and the currency in which we exchange beliefs."[2] To that end, libraries need to be *explainability activists*. Not passive, sceptical, or neutral but instead passionately on the frontlines of AI literacy, AI research and development, and technology policy. Explainability is a challenge for the AI community; it is an imperative for the library community.[3]

Explainable AI

"Explainable AI" (XAI) is the field of computer science "concerned with developing approaches to explain and make artificial systems understandable to human stakeholders."[4] Concerns about transparency and explanations have preoccupied AI since its earliest days.[5] The highly technical nature of XAI focuses on opaque AI algorithms using approaches such as feature engineering and model approximations.[6]

However, critics of XAI recognized a significant deficiency: "AI researchers are building explanatory agents for ourselves, rather than for the intended users . . . the inmates are running the asylum."[7] Taking user-centric approach, Upol Ehsan and Mark Riedl argue that "not everything that is important lies inside the black box of AI. Critical insights can lie outside it *because that's where the humans are.*"[8] This holistic focus is central to the subfield of XAI referred to as "human-centered explainable AI" (HCXAI). If explainability is to meet the values, principles, and policies central to academic libraries and the academy, adopting and promulgating the principles of HCXAI will be important.

Michael Ridley is Librarian Emeritus at the University of Guelph in Guelph, Ontario Canada, email: mridley@uoguelph.ca. ORCID: 0000-0002-2524-7507.

Human-Centered Explainable AI (HCXAI)

Human-centered explainable AI

> puts the human at the center of technology design and develops a holistic understanding of "who" the human is. It considers the interplay of values, interpersonal dynamics, and socially situated nature of AI systems. In particular, we advocate for a reflective sociotechnical approach that incorporates both social and technical elements in our design space.[9]

HCXAI consists of specific techniques, practices, system design features, and policy recommendations.[10] These emphasize the importance of understanding the human context, including who is using the AI, when, and why. It considers the broader environment in which AI operates and how humans interact with intelligent systems. Crucially, HCXAI stresses the need for AI explanations to be actionable and contestable. Users should be able to act on the explanations and challenge them if necessary. Inherent in these characteristics is the priority HCXAI places on user reflection over acquiescence to the system.

Many of the HCXAI recommendations emphasize user empowerment. In supporting the idea of "explanatory systems not explanations,"[11] HCXAI recognizes that a single, static response from an intelligent system is insufficient. Users should be engaged in a clarifying dialogue with the system and provided with additional information (e.g., training data source, model objectives, counterfactuals) so that a user can participate in "active self-explanation." This would allow users to form their own conclusions (explanations) and contextualize the system behavior to their circumstances.

While most contemporary technology strives for a seamless, frictionless experience, HCXAI advocates for a "seamful" experience. Seamful system design makes the limitations and boundaries of the system visible to the user, not hidden or smoothed over. Seamful design is a form of explanation that allows users to understand system weaknesses and modulate their trust appropriately.

HCXAI advocates for the availability of AI performance metrics. Akin to nutritional labeling on products, these user-friendly metrics would alert users to system effectiveness "in the wild" (as opposed to lab tests). It would also facilitate comparisons of different systems with similar objectives.

Perhaps the most radical HCXAI proposal is that "explanatory systems" should be independent of the platforms or AI systems providing the explanations.[12] Explanatory capabilities embedded in specific systems or platforms are at risk of bias or worse, manipulation or coercion. Third-party explanatory systems based on explanation protocols "would push the power and decision making out to the ends of the network [i.e., where the users are] rather than keeping it centralized among a small group of very powerful companies."[13]

An Agenda for Explainability Activists

Advancing explainability requires action in five areas. Some of these occur at the institutional level (library and/or university), while others involve individual commitments.

1. First and foremost, libraries must continue and enhance AI literacy initiatives for library staff and their user communities. Critical information literacy that incorporates AI issues remains the most effective explainability instrument.

2. Institutions should encourage and support research and development in the explainable AI community. Basic XAI research in opening the black box is still a priority as new advances continually redefine the core building blocks of AI.

3. Libraries need to hold its vendors and information providers accountable for explainability. From startups to OpenAI, explainability should be a default in the tools and services they provide. Contracts, user agreements, acceptable use policies, and public persuasion are all opportunities to influence vendor and provider behavior.

4. Explainability must be entrenched in regulation. While the European Union AI Act is making progress, it falls short on robust explainability provisions.[14] Efforts in the US and Canada are even less successful. It is unimaginable that any other such powerful innovation would not receive legislated guardrails and consumer protection. The promulgation of regulatory oversight of AI that enshrines explainability requires advocacy from libraries, universities, professional organizations, and individual librarians.

5. Libraries need to adopt the principles of HCXAI. Whether through library policy and practice, co-development with AI system designers, or advocacy work with professional and civil society, HCXAI represents a way forward to explainability that is consistent with the values and principles that guide the academy.

Conclusion

Latanya Sweeney, director of the Public Interest Tech Lab at Harvard, notes that "technology designers are the new policymakers; we didn't elect them but their decisions determine the rules we live by."[15] Rebalancing this power dynamic is a central concern.

Invoking "explainability activism" is more than a glib phrase. Misinformation, disinformation, bias, hallucinations, deepfakes, privacy violations, and intellectual property protection are just some of the urgent challenges and risks posed by AI.[16] Advancing the trustworthiness and accountability of AI tools and services used by academic libraries requires clear directions and effective practices. Explainability through the principles of human-centered explainable AI (HCXAI) represents an action agenda for libraries that can yield positive impacts.

Notes

[1] Peter Slattery, Alexander Saeri, Emily A. C. Grundy, Jess Graham, Michael Noetel, Risto Uuk, James Dao, Soroush Pour, Stephen Casper, and Neil Thompson, "The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks from Artificial Intelligence," ResearchGate, August 2024, https://doi.org/10.13140/RG.2.2.28850.00968.

[2] Tania Lombrozo, "The Structure and Function of Explanations," *Trends in Cognitive Sciences* 10, no. 10 (2006): 464, https://doi.org/10.1016/j.tics.2006.08.004.

[3] Michael Ridley, "Explainable AI (XAI): Adoption and Advocacy," *Information Technology & Libraries* 41, no. 2 (2022): 1–17, https://doi.org/doi.org/10.6017/ital.v41i2.14683.

[4] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum, "What Do We Want from Explainable Artificial Intelligence (XAI)?—A Stakeholder Perspective on XAI and a Conceptual Model Guiding Interdisciplinary XAI Research," *Artificial Intelligence* 296 (July 2021): 1, https://doi.org/10.1016/j.artint.2021.103473.

5. Edward H. Shortliffe, Stanton G. Axline, Bruce G. Buchanan, Thomas C. Merigan, and Stanley N. Cohen, “An Artificial Intelligence Program to Advise Physicians Regarding Antimicrobial Therapy,” *Computers and Biomedical Research* 6, no. 6 (1973): 544–60, https://doi.org/10.1016/0010-4809(73)90029-3.

6. Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M. Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera, “Explainable Artificial Intelligence (XAI): What We Know and What Is Left to Attain Trustworthy Artificial Intelligence,” *Information Fusion* 99 (2023), https://doi.org/10.1016/j.inffus.2023.101805.

7. Tim Miller, Piers Howe, and Liz Sonenberg, “Explainable AI: Beware of Inmates Running the Asylum” paper, International Joint Conference on Artificial Intelligence, Workshop on Explainable Artificial Intelligence (XAI), Melbourne, 2017, https://doi.org/10.48550/arXiv.1712.00547.

8. Upol Ehsan, Elizabeth Anne Watkins, Philipp Wintersberger, Carina Manger, Sunnie S. Kim, Niels van Berkel, Andreas Riener, and Mark O. Riedl, “Human-Centered Explainable AI (HCXAI): Reloading Explainability in the Era of Large Language Models (LLMs),” in *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems, CHI EA ’24* (Association for Computing Machinery, 2024), 2, https://doi.org/10.1145/3613905.3636311.

9. Upol Ehsan and Mark O. Riedl, “Human-Centered Explainable AI: Towards a Reflective Sociotechnical Approach,” in *HCI International 2020—Late Breaking Papers: Multimodality and Intelligence*, ed. Constantine Stephanidis et al., Lecture Notes in Computer Science (Springer International, 2020), 464, https://doi.org/10.1007/978-3-030-60117-1_33.

10. Michael Ridley, “Human-Centered Explainable Artificial Intelligence: An Annual Review of Information Science and Technology (ARIST) Paper,” *Journal of the American Society for Information Science* (2024): 1–23, https://doi.org/10.1002/asi.24889.

11. Robert R. Hoffman, Timothy Miller, Gary Klein, Shane T. Mueller, and William J. Clancey, “Increasing the Value of XAI for Users: A Psychological Perspective,” *KI—Künstliche Intelligenz* (2023), https://doi.org/10.1007/s13218-023-00806-9.

12. Michael Ridley, “Protocols Not Platforms: The Case for Human-Centered Explainable AI (HCXAI),” 2023, https://cais2023.ca/talk/10.ridley/10.Ridley.pdf.

13. Mike Masnick, “Protocols, Not Platforms: A Technological Approach to Free Speech,” Knight First Amendment Institute at Columbia University, August 21, 2019, p. 6, https://knightcolumbia.org/content/protocols-not-platforms-a-technological-approach-to-free-speech.

14. Luca Nannini, Agathe Balayn, and Adam Leon Smith, “Explainability in AI Policies: A Critical Review of Communications, Reports, Regulations, and Standards in the EU, US, and UK,” in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT ’23* (Association for Computing Machinery, 2023), 1198–1212, https://doi.org/10.1145/3593013.3594074.2023

15. Latanya Sweeney, “How to Save Democracy and the World,” paper, ACM Conference on Fairness, Accountability, and Transparency, New York University, 2018.

16. Gary Marcus, *Taming Silicon Valley: How We Can Ensure That AI Works for Us* (MIT Press, 2024).