Michael Ridley
University of Guelph, Guelph, Ontario, Canada

PROTOCOLS NOT PLATFORMS: THE CASE FOR HUMAN-CENTERED EXPLAINABLE AI (HCXAI)

Abstract

Currently explainable AI (XAI) is embedded in large, centralized platforms such as Facebook, Google or TikTok. These platforms control the nature and extent of explanations for their recommendations, decisions, and predictions raising the possibility of manipulation or deception. Human centered XAI (HCXAI) promotes explanatory systems not merely explanations as part of a set of principles supporting the non-expert, lay public. This position paper proposes moving from platform enabled HCAXI to protocol based HCXAI to facilitate user focused, independent explanatory systems more conducive to building and sustaining user trust and system accountability.

1. Introduction

Machine learning systems are complex, powerful, opaque, and ubiquitous. They are part of the "digital everyday" (Kant, 2020). More importantly, they are consequential. Their recommendations, decisions, and predictions have a material effect on our lives. The opacity of machine learning, the "black box" effect, has led to the rise of "explainable AI" (XAI) as a means of establishing validity, trust, and accountability (Adadi & Berrada, 2018). XAI research and development has been dominated by a focus on technical issues and the needs of system developers. Only recently has human centered XAI (HCXAI) emerged with a focus on the needs of the non-expert, lay population (Ehsan & Riedl, 2020; Haque et al., 2023).

The principles of HCXAI call for robust "explanatory systems" not merely isolated explanations (Mueller et al., 2021). A neglected issue is where, and under whose control, will such explanatory systems exist? Currently XAI, human centered or otherwise, is embedded in the platform (e.g., Facebook, Google, TikTok).

In 2019, Masnick proposed a solution to the challenges of free speech on social media: focus on protocols, not platforms (Masnick, 2019). Introducing protocols "would push the power and decision making out to the ends of the network, rather than keeping it centralized among a small group of very powerful companies" (Masnick, 2019, p. 6). Protocols can be applied to XAI and specifically to the objectives of HCXAI. Developing and promulgating HCXAI protocols would create user focused, independent explanatory systems better able to promote trust and accountability in machine learning systems.

2. Protocols

Protocols are "common tools designed for controlling information transfer between computer systems. They are made up of sequences of messages with specific formats and meanings"

(Pouzin & Zimmermann, 1978, p. 1346). Their development and proliferation during the 1970s were notable because “for the first time the design of computer systems puts the emphasis not on the internal management of resources, but on communication between resources of different systems” (Pouzin & Zimmermann, 1978, p. 1368).

Early protocols include Usenet (NNTP), Terminals (Telnet), and File Systems (FTP) (Khare, 1998), and enduring protocols such as internet email (SMTP) (Partridge, 2008). The divisive “protocol wars” regarding internet networking eventually resulted in the emergence of TCP/IP as a standard (Russell, 2014). TCP/IP allowed for the development of the HTTP protocol and the web (Berners-Lee, 1999). New information protocols are being discussed and developed. A recent example is the debate over protocols for the Internet of Things (IoT) (Silva et al., 2019). An emerging protocol, the Authenticated Transfer Protocol (ATP), is a social network protocol to enable federated social networking (“The AT Protocol,” 2022). Tokenized systems (such as Blockchain) have been proposed as general-purpose exchange protocols (Masnick, 2019; Xiao et al., 2020).

While focused on different applications for different purposes, each of these protocols created or proposed a standard for interchanges that enabled an open environment facilitating access, diversity, and competition. In addition to standardization, protocols offer guidelines for system development, modular components to support flexibility (e.g., for different languages, contexts, and domains), customization opportunities at both the client and server ends, and a governance structure to manage the evolution of the protocol. The design, technologies, and decision-making processes of these protocols offer lessons and templates for the development and promulgation of HCXAI protocols.

3. HCXAI

While the European Union’s General Data Protection Regulation (GDPR) introduced, and codified to a certain extent, the “right to explanation” (European Union, 2016; Goodman & Flaxman, 2017), the GDPR is now six years old, and most subsequent legislation and regulation has focused on limited, high-risk contexts rather than overall explanatory needs. However, as recognized by the recently proposed US AI Bill of Rights (White House Office of Science and Technology Policy, 2022), it is with everyday consumer-facing systems, low risk but still consequential, that most users engage with machine learning and might want or expect explanations.

Consumer-facing recommender systems, platforms like Facebook, Google, TikTok, and many others, provide explanations using a variety of XAI techniques and strategies. The platform determines the nature and conditions of the explanations provided, largely discounting or ignoring the needs and preferences of the user. The platform’s control of the explanations raises questions about possible manipulation, deception, and a general withholding or concealment of information. These concerns are key issues in user trust and system accountability.

HCXAI principles for system design emphasize “explanatory systems, not explanations,” the importance of context and user expertise, and the recognition that an explanation is a process never a “one-off” event (Mueller et al., 2021). While these principles are user focused, they still assume platform implementations. The power and agency regarding explanations remains with the platform.

Putting the user in more control is the central idea behind focusing on protocols rather than platforms. Given the user focus of HCXAI, shouldn't the recommended explanatory systems be independent of the platforms?

4. HCXAI Clients and Servers

HCXAI protocols enable a client-server architecture allowing both the client and server software to be customized to the needs and requirements of the user and the platform. An HCXAI server could be shared by multiple platforms, or it could be implemented for a specific platform. A customizable client would allow a user to

> create their own set of rules—including which content do they not want to see and which content would they like to see promoted. Since most people would not wish to manually control all of their own preferences and levels, this could easily fall on any number of third parties—whether they be competing platforms, public interest organizations, or local communities. Those third parties could create whatever interfaces, with whatever rules, they wanted" (Masnick, 2019, p. 17).

The rules would allow the client to negotiate explanations from the HCXAI server implementation. These rules could include the nature, extent, and complexity of explanations, different presentation preferences (e.g., text, visualizations), and even not requiring an explanation at all from fully trusted sites. A history of the user's interactions with explanations could be maintained by the client allowing for adjustments as the user's algorithmic literacy increased.

One option for an HCXAI client is to incorporate the protocol into an internet browser, something that might align with the privacy and public interest objectives of browsers such as Firefox and Brave. Since the HCXAI client would contain personal and confidential information, data security would be important. This need suggests using the emerging protocol-based Solid Pods (www.inrupt.com/solid) that Berners-Lee is developing for secure, federated access to personal data.

5. Barriers and Challenges

Masnick acknowledges the difficulty in building and maintaining protocols: "most of the work was done by volunteers, and protocols over time were known to atrophy without attention" (Masnick, 2019, p. 23). While the pressure for protocols can come from government, consumer advocates, and industry itself, it is industry groups that typically define the protocol, and manage its support and maintenance.

Barriers to HCXAI protocols come from both users and platforms. The HCXAI client may be perceived as too complicated or bothersome for users. Platforms may be concerned that HCXAI protocol requirements will expose or compromise IP or trade secrets. While protocols create a level playing field for new entrants to the marketplace, existing and new providers might complain that protocols inhibit innovation (Mignano, 2022). However, clearly the power and entrenchment of existing platforms will be the single biggest challenge to adopting HCXAI protocols.

6. Conclusion

Effective HCXAI is an essential aspect of consumer-facing machine learning systems. The embedding HCXAI in platforms perpetuates the control of explanations by centralized systems. Moving to HCXAI protocols would enable independent explanatory systems insulated again possible manipulation or deception by the powerful platforms. User preferences, specific contexts, and levels of algorithmic literacy could be built into client software facilitating the user focus central to the HCXAI principles.

The "protocol wars" of the early days of the internet demonstrated the importance of protocols to an open, vibrant network upon which was implemented a global information infrastructure. With respect to HCXAI protocols, perhaps battles and skirmishes can be replaced with collaboration and cooperation enabling the same openness that will facilitate trust and accountability in the ubiquitous and consequential machine learning systems of our everyday lives.

7. References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Berners-Lee, T. (1999). *Weaving the Web: The original design and ultimate destiny of the World Wide Web by its inventor*. Harper.

Ehsan, U., & Riedl, M. O. (2020). Human-centered explainable AI: Towards a reflective sociotechnical approach. *Proceedings of HCI International Conference on Human-Computer Interaction*. http://arxiv.org/abs/2002.01092

European Union. (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016*. http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision making and a "right to explanation." *AI Magazine*, *38*(3), 50–57. https://doi.org/10.1609/aimag.v38i3.2741

Haque, A. B., Islam, A. K. M. N., & Mikalef, P. (2023). Explainable artificial intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research. *Technological Forecasting and Social Change*, *186*. https://doi.org/10.1016/j.techfore.2022.122120

Kant, T. (2020). *Making it personal: Algorithmic personalization, identify, and everyday life*. Oxford University Press.

Khare, R. (1998). The transfer protocols. *IEEE Internet Computing*, *2*(2), 80–82. https://doi.org/10.1109/4236.670688

Masnick, M. (2019). *Protocols, not platforms: A technological approach to free speech*. Knight First Amendment Institute, Columbia University. https://knightcolumbia.org/content/protocols-not-platforms-a-technological-approach-to-free-speech

Mignano, M. (2022, July 16). The standards innovation paradox. *Medium*. https://mignano.medium.com/the-standards-innovation-paradox-e14cab521391

Mueller, S. T., Veinott, E. S., Hoffman, R. R., Klein, G., Alam, L., Mamun, T., & Clancey, W. J. (2021). Principles of explanation in human-AI systems. *Explainable Agency in Artificial Intelligence Workshop. AAAI 2021*. Explainable Agency in Artificial Intelligence Workshop. AAAI 2021. http://arxiv.org/abs/2102.04972

Partridge, C. (2008). The technical development of internet email. *IEEE Annals of the History of Computing*, *30*(2), 3–29. https://doi.org/10.1109/MAHC.2008.32

Pouzin, L., & Zimmermann, H. (1978). A tutorial on protocols. *Proceedings of the IEEE*, *66*(11), 1346–1370. https://doi.org/10.1109/PROC.1978.11145

Russell, A. L. (2014). *Open standards and the digital age: History, ideology, and networks*. Cambridge University Press.

Silva, J. de C., Rodrigues, J. J. P. C., Al-Muhtadi, J., Rabêlo, R. A. L., & Furtado, V. (2019). Management platforms and protocols for internet of things: A survey. *Sensors*, *19*(3), 1–40. https://doi.org/10.3390/s19030676

The AT protocol. (2022, October 18). *Bluesky*. https://blueskyweb.xyz/blog/10-18-2022-the-at-protocol

White House Office of Science and Technology Policy. (2022). *Blueprint for an AI Bill of Rights: Making automated systems work for the American people*. OSTP. https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf

Xiao, Y., Zhang, N., Lou, W., & Hou, Y. T. (2020). A survey of distributed consensus protocols for blockchain networks. *IEEE Communications Surveys & Tutorials*, *22*(2), 1432–1465. https://doi.org/10.1109/COMST.2020.2969706