Explainable AI Reloaded: Challenging the XAI Status Quo in the Era of Large Language Models

Upol Ehsan
Georgia Institute of Technology
USA

Mark O. Riedl
Georgia Institute of Technology
USA

ABSTRACT

When the initial vision of Explainable (XAI) was articulated, the most popular framing was to open the (proverbial) "black-box" of AI so that we could understand the inner workings. With the advent of Large Language Models (LLMs), the very ability to open the black-box is increasingly limited. Especially when it comes to non-technical end-users. In this paper, we challenge the assumption of "opening" the black-box in the LLM era and argue for a shift in our XAI expectations. Highlighting the epistemic blind spots of an algorithm-centered XAI view, we argue that a human-centered perspective can be a path forward. We operationalize the argument by synthesizing XAI research along three dimensions: explainability outside the black-box, explainability around the edges of the black box, and explainability that leverages infrastructural seams. We conclude with takeaways that reflexively inform XAI as a domain.

CCS CONCEPTS

• **Human-centered computing**; • **Computing methodologies** → **Artificial intelligence**;

KEYWORDS

Explainable AI, Large Language Models, Generative AI

ACM Reference Format:
Upol Ehsan and Mark O. Riedl. 2024. Explainable AI Reloaded: Challenging the XAI Status Quo in the Era of Large Language Models. In *Halfway to the Future (HTTF '24), October 21–23, 2024, Santa Cruz, CA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3686169.3686185

1 PROVOCATION

With the advent of Foundation Models & Large Language Models like ChatGPT, is "opening the black-box" still a reasonable and achievable goal for Explainable AI (XAI)? Do we need to shift our perspectives?

In February of 2023, Nadeem (pseudonym), a relative of the first author of this article, asked if ChatGPT could be used to do homework. Nadeem is a high-schooler who shared that he was always under a tight deadline and needed to be more "efficient" with his homework. He heard from his friends that ChatGPT can help summarize papers or books, which can make Nadeem more "productive" with his homework.

Before responding to Nadeem, ChatGPT was taken for a test drive. It was prompted to summarize an academic paper (Figure 1 similar to how Nadeem might use it – as someone who was not an AI researcher or experienced prompting Large Language Models (LLMs). Fortunately, ChatGPT generated a coherent response. ChatGPT gave the names of the authors of the paper and details about the paper's publication at ACM FAccT 2020. Unfortunately, the names of the authors and publication details were made up! The confabulated author names were immediately obvious because we wrote the paper that was prompted to be summarized [20]. However, the rest of the details was extremely plausible – the paper very well could have appeared at that conference, but did not. The first author of this paper almost missed another detail in ChatGPT's summary. The original paper described a framework with two dimensions: social and technical. The generated summary claimed the framework described three dimensions: social, technical, and cultural, which, while wrong, was plausible enough that even the very author of the paper almost missed that crucial inaccuracy!

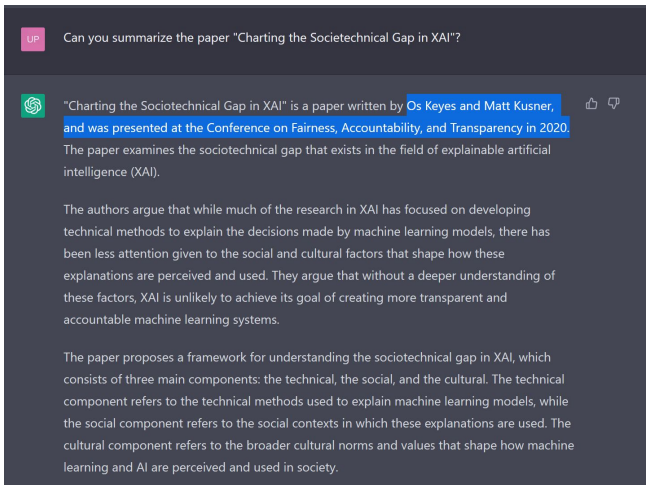

Figure 1: Screenshot of ChatGPT's summary of a paper with hallucinated or confabulated content.

1.1 Separating Fact from Fiction

The true story above demonstrates the effortful process required to disentangle fact from fiction in GPT's output, even from someone knowledgeable of the source material. Even more notably, there was no way for our protagonist, an expert in Explainable AI, to "open" the black-box of ChatGPT and understand why it produced what it produced or where it might be faithful to the facts or prone to confabulation (also called hallucination). On the one hand, he lacked access to the internal details such as the parameters of the model. On the other hand, even if one did have access to the internal parameters of the model, given the scale and complexity of the

HTTF '24, October 21–23, 2024, Santa Cruz, CA, USA

ACM ISBN 979-8-4007-1042-1/24/10
https://doi.org/10.1145/3686169.3686185

neural architecture of such a large language model, interpreting it is unlikely to produce human-understandable and actionable information.

2 TENSIONS: XAI AND LARGE LANGUAGE MODELS

The field of Explainable AI (XAI) is concerned with developing techniques, concepts, and processes that can help stakeholders understand the reasons behind the AI system's decision-making [21, 34].

For our purposes, we adopt a design lens in XAI that is sociotechnically-informed [12, 19, 34] and adopt the broad definition that an explanation is an answer to a *why*-question [11, 30, 35]. Given AI systems exist in sociotechnical settings [33, 45], it takes more than just algorithmic transparency to make them explainable [23, 35]. Thus, explaining what is happening "inside the black box" often requires us to also understand things "outside the black box" [12, 16, 32], requiring us to consider the entire AI lifecycle (vs. just the algorithm). For instance, why a facial recognition system disproportionately misclassified women of color [8] can be explained by looking at demographic compositions in the training data. A sociotechnically situated view of XAI expands the concept of explainability beyond the bounds of the algorithm [16] and positions it as a relational and audience-dependent construct instead of a model-inherent one [4, 5, 35, 36]. Emerging work [27, 40, 41] showcases how a broader XAI perspective can potentially address criticisms of popular algorithm-centered XAI techniques, which can be ineffective [3, 39, 48] and potentially risky [29, 44].

When we consider a service such as ChatGPT, GPT-4, Microsoft Copilot, Google Gemini, Claude, or Meta AI, what prospects are there for "opening" the black-box of AI? These models have hundreds of billions of parameters, all acting in conjunction to generate a distribution over possible words to choose from to build a response, word by word. If we had access to all the weights, could we interpret and explain the model? If we had access to the parameters of a model and the activation values for an input could we interpret and explain the model? In the case of the above large language models the point is moot. All these models run on servers behind APIs that do not allow inspection of the neuron activations and weights. However, even if we could access this information, the raw values of weights and activations are meaningless to most people without synthesizing some visualization or text summarization that provides a lay-understandable analysis of the internal operations of the system and how the results were generated by the system. Consider OpenAI's work on interpreting the patterns that cause individual neurons to activate [7]. How would knowing what causes neuron #2142 to activate have helped Nadeem, a non-AI expert, know how to better use ChatGPT to complete his homework? What actionable information from this neural activation pattern can a non-AI expert use meaningfully?

LLMs are increasingly being incorporated as components in systems that chain multiple processes together. In particular, Retrieval Augmented Generation (RAG) combines LLMs with web search such that a web retrieval module first retrieves relevant documents, which are then used to inform an LLM [31]. While opening the black-box LLMs may not yield actionable explanations, modular architectures afford the ability to inspect and explain how data is changed going in and out of black-box modules.

3 IS EXPLAINABLE AI DOOMED TO FAIL?

Despite the commendable progress in algorithm-centered approaches in XAI, there are significant deficiencies. Studies examining how people actually interact with AI explanations have found popular XAI techniques to be ineffective [3, 39, 48], potentially risky [29, 44], and even obsolete in real-world deployed contexts [32]. XAI developers tend to design explanations *as if* people like them are going to use their systems, earning an infamous reputation of "inmates running the asylum" [35]. In fact, a majority of current deployments serve AI engineers instead of end-users whose needs are ignored [6]. This creates a gap between design expectations and reality— how developers envision the designed AI explanations to get interpreted and how users actually perceive those explanations in reality.

As Large Language Models (LLMs) become prominent, is Explainable AI – a research area in flux and its infancy – doomed to fail? No. There is hope. Before we throw in the towel, there are a few things to consider.

3.1 AI systems are Human-AI assemblages

First, the techno-centric, algorithm-centered, discourse of XAI fails to appreciate the sociotechnical reality of AI systems. When we say "AI systems," what we very often mean to say is "Human-AI assemblages," where the "human" part of the Human-AI assemblage is often implicit [16]. No real-world AI systems work in a vacuum. Black-boxes *by themselves* do not do the work – humans *with* black-boxes do the work [19]. Even if the human contribution to the work is to just provide an input, this is a significant contribution because AI systems are useful to people as tools. Thus, the explainability of AI systems entails explainability of the Human-AI assemblage, which has at least two components: the human (or humans) and the AI [16, 20]. Thus, how can we achieve the explainability of the Human-AI assemblage by just focusing on the explainability of the AI model? *XAI is therefore not just technical, it is sociotechnical.* It requires more than just algorithmic transparency – more than being able to open the black box.

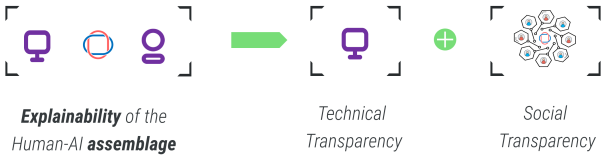

Figure 2: Illustrating how the explainability of the Human-AI assemblage is more than just technical (algorithmic) transparency

Second, what we mean by "AI" is evolving. Compared to AI systems even five years ago, the Deep Learning systems in the Foundation Model era, such as LLMs, are much more complex, have orders of magnitude more parameters, and are running at unprecedented scales. Thus, AI as a *design material* is tricky and is evolving [15, 20, 47]. Our understanding and expectations of what

it means for AI-as-design-material to be explainable should also evolve. Further, XAI techniques that focus solely on the algorithm or the model face a new challenge: it is getting increasingly hard to open the black box! As AI systems are increasingly end-user facing, those that need the explanations the most are on the other side of an AI or user interface. This is the case for the most popular Large Language Models and chatbots, and it is also the case for other types of consumer-facing systems. When the initial vision of XAI was articulated, a popular framing was to "open" the (proverbial) "black-box" of AI [9, 37], so that we could see inside of it, figure out what it was doing, why it was doing it, and if it was doing it correctly. With the advent of large language models, that ability to open the black-box is increasingly limited due to the sheer complexity of the models and the increased prevalence of models behind restrictive APIs. And even if we did manage to "open" it, we will not understand what we see.

4 HUMAN-CENTERED EXPLAINABLE AI (HCXAI): BEYOND ALGORITHMIC TRANSPARENCY

Given AI systems are bounded by their training data, by construction, they cannot incorporate the real-world dynamics "outside" the black-box. Thus, an algorithm-centered view of XAI is–by construction–a limiting view, one that handicaps the XAI system from doing what we want to do– solve real world problems. We need a paradigm that can accommodate an expansion of the epistemic canvas– an increase of the aperture of the viewing lens– to include the sociotechnical dynamics in which XAI systems are embedded so that we can do what we set out to do – solve real world problems.

This is where the domain of *Human-Centered Explainable AI (HCXAI)* [19] can help. HCXAI is a holistic vision of AI explainability, one that is human-centered and sociotechnical in nature. Situated as a Critical Technical Practice [1, 2], it draws its conceptual DNA from critical AI studies and HCI (e.g., reflective design [13, 14, 42], value-sensitive design [25]). HCXAI encourages us to critically reflect and question dominant assumptions and practices of a field, such as algorithm-centered XAI. It also adopts a value-sensitive approach to both users and designers in the development of technology that challenges the status quo of a field. HCXAI encapsulates the philosophy that not everything that is important lies inside the black box of AI. Critical insights can lie outside it. Why? *Because that's where the humans are.*

Thinking *outside* the black box of AI can help us meet our goals of helping people understand and calibrate their trust in AI systems. Even if we cannot meaningfully open the black box or interpret its complexities, there are a lot of things we can do to understand and explain the system *around* the black box. Increasing the aperture of XAI can help us focus on the most important part: *who* the human(s) is (are), what they are trying to achieve in seeking an explanation, and how to design XAI techniques that meet those needs. Indeed, explanations of the sociotechnical system can offer us an important affordance: *actionability* [18, 28, 43].

At its core, actionability is about what a user can do with the information in an explanation [18]. An actionable XAI system empowers the user by increasing the space of possible informed actions to achieve their end goals. This could be understanding how to change the inputs, contesting a decision, or learning when and how to use the system more appropriately. Actionability also addresses another important question: how do we know if an XAI system is useful? There are an increasing number of reports of XAI systems that are deployed and fail to have any measurable impact on their users [3, 44]. Many of these systems failed because the XAI systems were not designed with user needs in mind, such as by providing users with information they could already intuit themselves, by providing information that was onerous to verify, or by providing information that users could not use. In other words, the explanations generated by the systems were not actionable.

5 THE WAY FORWARD

With the reframing around human-AI assemblages and XAI systems that place the human as the central concern, and armed with actionability as the metric for success, we now lay out three possible paths forward. This list is not meant to be exhaustive or prescriptive. It is meant to be generative by providing emerging evidence for how Human-Centered XAI (HCXAI) can address the growing needs for understanding our increasingly AI-infused world.

5.1 Explainability outside the black-box: Social Transparency

Most consequential AI systems are embedded in organizational environments where groups of humans interact with it. These real-world AI systems, as well as the explanations they produce, are socially-situated [22, 32]. Therefore, the socio-organizational context in which these systems are used is key. Why are we not incorporating socio-organizational contexts into how we think about explainability in AI? How can we tackle the explainability of Human-AI assemblages?

Enter Social Transparency (ST) a sociotechnically-informed perspective that incorporates the socio-organizational context into explaining AI-mediated decision-making [16]. Social transparency allows us to augment the explainability of a human-AI assemblage without necessarily changing anything about the AI model. Social transparency allows one to annotate an output or behavior from an AI system with the 4W **who** did **what**, **when** and **why**. These annotations are shared between others using the system. They allow users to see whether and why others have accepted or rejected an AI's output. Social transparency does two important things: first, it challenges the dominant narrative of algorithm-centered notions of XAI; second, it expands our understanding of XAI beyond technical transparency by illustrating how adding social context can help people make better, more actionable decisions with AI systems.

Imagine the following scenario (Figure 3): Aziz is a software seller trying to use a powerful AI-based pricing tool to do something consequential: offer the right price to a client company. The AI suggests a price. Moreover, its suggestion has technical transparency – it explains its recommendation by showing Aziz the top features it considered, such as sales quota goals, comparative

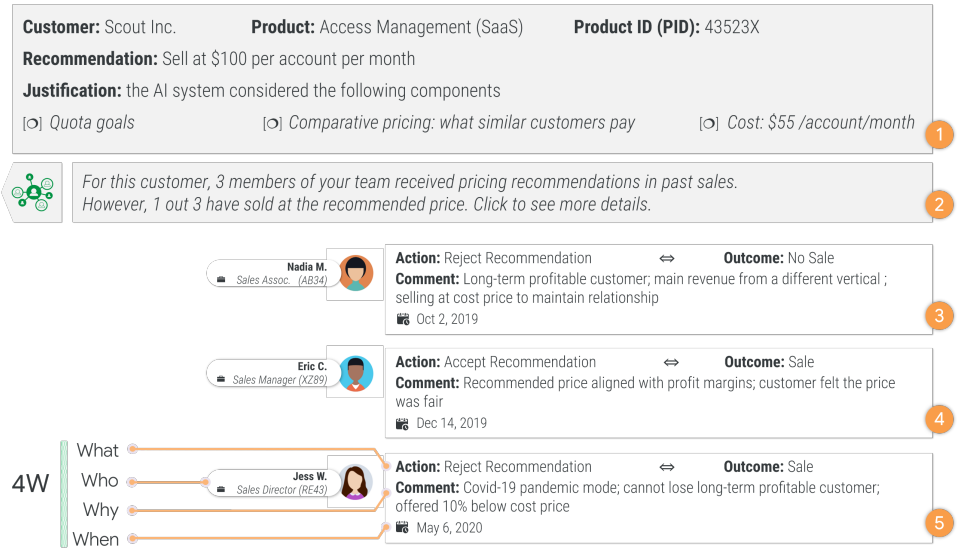

Figure 3: Sales scenario with Social Transparency (ST) used in [16] (reproduced with permission from authors). The labeled blocks are: (1) Decision information and model explanation: Information of the current sales decision, the AI's recommended price and a "feature importance" explanation justifying the model's recommendation, inspired by real-world pricing tools; (2) ST summary: Beginning of ST giving a high-level summary of how many teammates in the past had received the recommendation and how many sold at the recommended price; (3-5): ST blocks with "4W" features containing the historical decision trajectory of three other users.

pricing with other clients, and costs. Confident with the AI's recommendation, Aziz makes a bid, but the client finds the price too high and walks out.

Despite an accurate AI model and the presence of technical transparency, why did the bid fail? There could be algorithmic reasons for it. But might also be relevant contextual factors outside the box that can help explain why the bid failed. Perhaps the history between Aziz and the client that was not honored? Or maybe there were external events that happened since the model was trained, such as a pandemic-induced budgetary crisis.

Now imagine that Aziz could see that more than 65% of his peers rejected the AI's pricing recommendation (Block 2 in Fig. 3). Or, what if Aziz knew that Jess, a director in the company, sold the product at a loss due to pandemic-related budgetary cuts?(Block 5 in Fig. 3)

This peripheral vision of *who did what, when and why* – called the **4W** – are the constitutive design elements of Social Transparency that can encode relevant socio-organizational context. The benefit of taking a holistic approach to explainability is clear: a study of real-world AI users in sales, cybersecurity, and healthcare found that social transparency, in the form of the 4W, helped people calibrate their trust in the AI's performance, provide actionable information for AI contestability and robust decision-making, and the organizational context made visible enabled better collective actions in the organization and strengthened the human-AI assemblages [16].

By incorporating the socio-organizational context, Social Transparency makes our understanding of XAI more holistic, representing the Human-AI assemblage more realistically than a purely algorithm-centered XAI view. We should note that Social Transparency is agnostic to whether an AI system is black-boxed or not. As long as there is an AI-based recommendation or decision, we can attach 4W – the socio-organizational context – to it. In a completely black-boxed AI system, there will not be any technical transparency. However, the 4W can add transparency to the social side of the Human-AI assemblage.

5.2 Explainability around the Edges of the Black Box: Rationale Generation & Scrutability

If the black box cannot be cracked open in any meaningful sense, there is another possibility: incorporate explainability *around the edges* of the black box to foster a better functional understanding in the user [38] such that it fosters actionability. One of the original formulations of rationale generation [21] postulated that there was no need to know how a black box worked as long as we could learn how to give actionable advice about the black box by looking at its inputs and outputs. It was philosophically grounded in Fodor's work on Language of Thought [24]: how is it that, despite not having a 1-1 neural correlate of thought, humans can effectively communicate by translating their thoughts into words? For Human-AI interaction, even if the exact mechanism of the (artificial) neural correlate of AI's thought was not known to the human, as long as actionable information is present in the explanation from an AI agent, the Human-AI interaction can proceed. In short, explanations that do not directly access the model can still generate actionable information.

In the case of large language models, the actionable information is whether any particular input is likely to produce a reliable

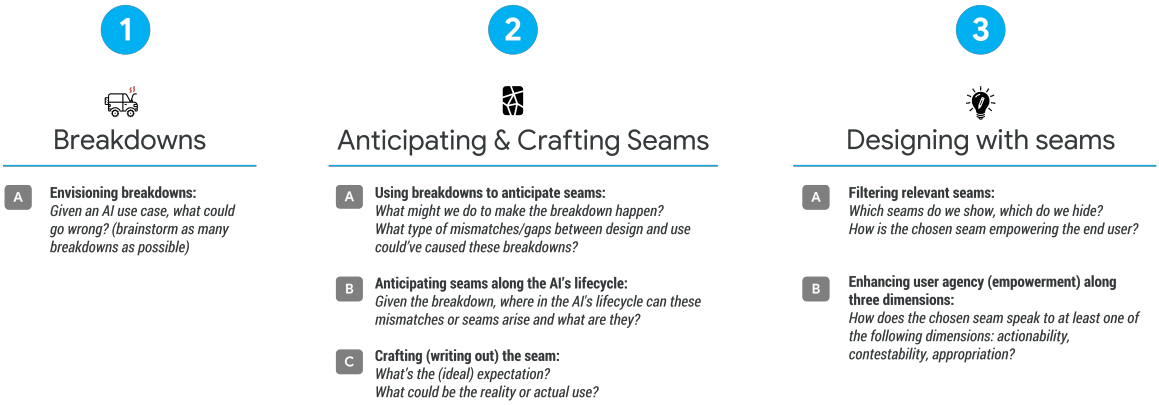

Figure 4: An overview of the Seamful XAI design process used in [17] (reproduced with permission).

response that can be trusted. Large language models might be generally capable at many tasks such as question-answering, they are not infallible, and it is always possible for a user to ask a question that results in a confabulation (also called a "hallucination") that the user is unable to vet. In this case, we can directly use the API to probe how it responds to particular stimuli [46]. It is proposed that an XAI system can decompose the original, human authored question into a series of more fine-grained, related questions that provide more opportunities for the model to confabulate responses if it is not competent at the original question. These sub-questions can be selected to be easier for the user to vet. Generating questions to challenge an LLM has been demonstrated to increase users' ability to determine whether the answer should be trusted or not.

5.3 Explainability by Leveraging Infrastructural Seams: Seamful XAI

No AI system is perfect. Mistakes are inevitable. Breakdowns in AI systems often occur when the assumptions we make in design and development do not hold true when they are deployed in the real-world. For example, an AI system can fail when it is trained on data from North America but deployed in South Asia, especially when the end user is unaware of this infrastructural mismatch. These mismatches between design assumptions and real-world usage are called *seams* [17]. Handling the mistakes from AI systems is hard, especially when the AI's decision-making is hidden or black-boxed. Although black-boxing AI systems can make the user experience *seamless* and easy to use, concealing the seams can lead to downstream harms for end-users, such as uncritical AI acceptance. What can we do differently? How do we move beyond seamless AI? And what can we gain by doing so?

Seamful XAI is a design lens that incorporates the principles of seamful design [10] to augment explainability and user agency. A classic example of seamful design is a "seamful map" of WiFi coverage in your home. If you know the WiFi's dead zones in your home, you will be able to best use it because you can then avoid.Without revealing the seams, users can have reasonable expectations of perfect WiFi. The map makes the seams in the WiFi's infrastructure visible to users, which allows them to recalibrate their expectations and behavior. A seamful design principle asks us to leverage the weakness in opportunistic ways [26].

Unlike seamlessness, *seamful design does not aim to hide the infrastructure*. Rather, it puts the infrastructure and all its imperfections front and center. Seamful design helps us recognize and grapple with the complex infrastructures systems reside in. Conversely, seamless design ideals risks making the labor it takes to make the system work invisible (e.g., datawork, ghostwork, maintenance work). And, as invisible work is invariably unaccounted for and unappreciated, workers who conduct this work will feel undervalued or invisible. Seamfulness embraces the imperfect reality of spaces we inhabit and makes the most out of it.

In the context of AI, *seams can be conceptualized as mismatches, gaps, or cracks in assumptions between the world of how AI systems are designed and the world of how AI systems are used in practice*. Seamful XAI seeks to empower users with information that augments their agency by identifying gaps between ideal design assumptions and reality.

At the heart of Seamful XAI are four observations:

(1) Seams are inevitable, arising from the integration of heterogeneous sociotechnical components during technology deployments.
(2) Seams are revealed through system breakdowns.
(3) Instead of treating seams as problematic negatives to be erased, they can be used strategically to calibrate users' reliance and understanding of an AI system.
(4) The goal of this strategic revelation (and concealment) is to support user agency (actionability, contestability, and appropriation).

Seamful XAI Design Process: Let's review the design process proposed by [17].

The *first* step of the process begins with generating "breakdowns." Breakdowns are answers to the question, "what could go wrong when this technology gets deployed?" Answers could include technology failures, unfair treatment of groups, inducing over-reliance, or deskilling.

The *second* step is around anticipating and crafting seams, which is done in three parts. First (2A in the diagram), we ask: "what might

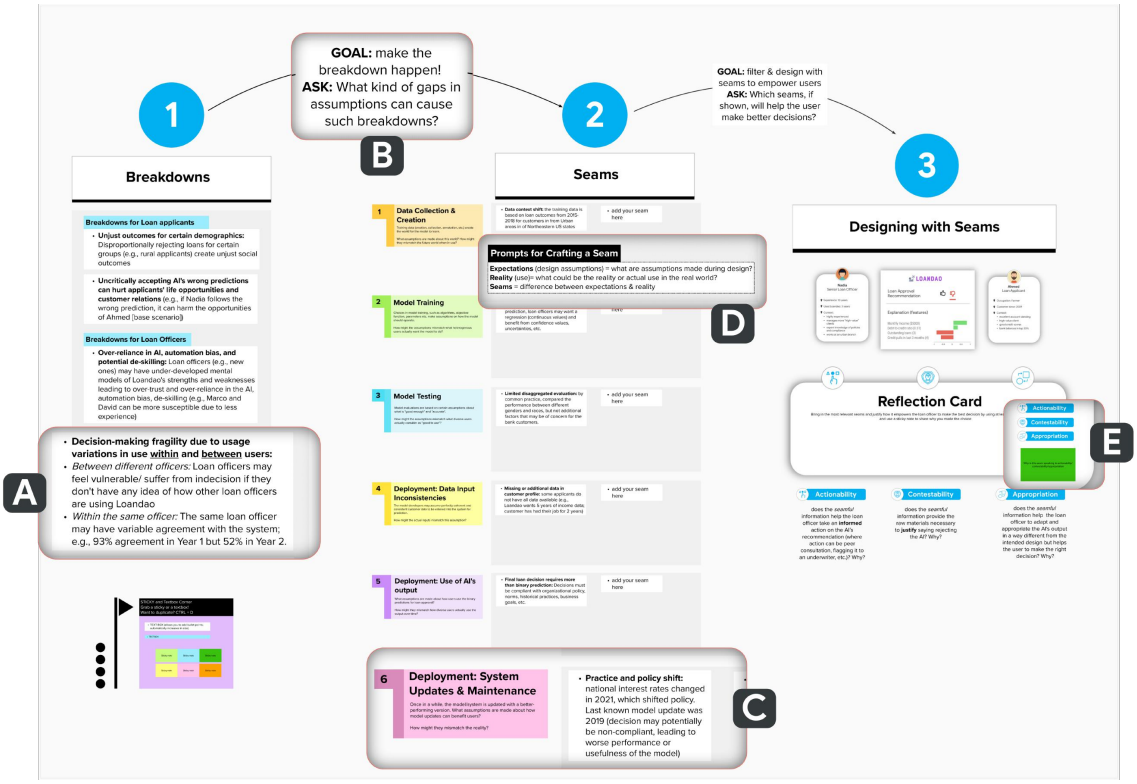

Figure 5: The virtual whiteboard used for the seamful XAI design activity showing key features in [17] (reproduced with permission). Area 1: Envisioning breakdown (Step 1). Participants were provided sample breakdowns (A), which participants could either use directly or get inspiration for their own envisioning. Area 2: Anticipating & crafting seams (Step 2). Fuiding prompts were provided (B) for effectively crafting the seams. Exemplary seams were shared (C) for each stage of the AI lifecycle framework. Area 3: Designing with seams (Step 3). Participants were asked to articulate their reasoning for choosing a seam and tag which user goals the selected seam (E) can support for augmenting user agency.

we (as developers, designers, researchers, etc.) do to make the breakdown happen?" While this question might seem counter-intuitive, it allows us to systematically prevent breakdowns by understanding their causes. This step inverts the problem and makes it a goal directed task, which is important to generate concrete outcomes instead of open-ended problems. Next (2B), we try to anticipate the reasons for the breakdown (the seams) in the appropriate stage in the AI's lifecycle (the colored boxes numbered 1-6 in Fig. 5). Finally (2C), we craft the seam by thinking about the gap between the ideal expectation and the reality of use.

The *final* step involves using the seams generated in step 2 in a way to empower user agency and explainability. Here (3A), we ask: given our end goal, which seams do we show and which do we hide (e.g. strategic revelation and concealment)? The revealed seams (3B) should empower users through better explainability. This step of the Seamful XAI process is a major differentiator from other Responsible AI processes: unlike most processes that stop at identifying gaps, this one goes beyond. It not only uncovers the gaps but also utilizes them as avenues to support users (for more details, refer to [17]).

A co-designing study [17] with 43 real-world AI users found three beneficial elements of Seamful XAI:

- It **enhances explainability** by helping stakeholders reveal the AI's blind spots, highlight its fallibility, and showcase the strengths and weaknesses of the system, which can calibrate reliance in AI systems.
- It **augments user agency** by providing peripheral vision of the AI's blind spots. Seamful information expands the action space of what users can do. Information in seams can convert "unknown unknowns" to "known unknowns," which can empower users to know "where" to start an investigation.
- It is a resourceful way to not just reveal seams but also **anticipate and mitigate harms from AI systems**.

6 TAKEAWAYS

We began with the provocation: With the advent of Foundation Models & Large Language Models like ChatGPT, is "opening the black-box" still a reasonable and achievable goal for XAI? Do we need to shift our perspectives?

Yes. The proverbial "black-box" of AI has evolved, and so should our expectations on how to make it explainable. As the box becomes more opaque and harder to "open," the human side of the Human-AI assemblage remains as a fruitful space to explore. In the most extreme case, *the human side may be all there is left to explore.* Even

if we can open the black box it is unclear what actionable outcomes would become available.

There are four important lessons from Human-centered XAI that can inform the shift in our XAI expectations.

(1) First, the human-centered XAI perspective takes a pragmatic and resourceful view of explainability, especially if black boxes are expected to persist. By considering the actions afforded to the user by the explanations, HCXAI centers the focus on the human, ensuring AI augments human abilities rather than replace them.
(2) Second, explainability is not only achieved by looking inside the black box through mechanistic descriptions of how an algorithm works. Actionability can be achieved by exploring explainability outside and around the edges of the black box because human-centered XAI takes a more expansive view of what it means to provide insights into a black box that can afford a wider range of actions.
(3) Third, explicitly treating AI systems as human-AI assemblages means focusing on explainability of the assemblage, not just the AI. This widened perspective opens up avenues for not just factoring in who is interacting with the black box, but also how human teams can work together — directly or indirectly — to contextualize a dynamically changing real-world AI behavior.
(4) Fourth, seamful XAI turns the disadvantages and weaknesses of an AI system into advantages. The gaps between user expectations and AI capabilities are exactly the gaps that explanations address. Instead of hiding those gaps to create seamless experiences, seamful XAI leverages these gaps in an opportunistic manner to augment explainability and user agency.

As we reload our expectations on XAI, we invite you to do what HCXAI asks us to do: centering the design and evaluation around the human. This positioning can reveal unmet needs that must be addressed while avoiding the costly mistake of building XAI systems that do not make a difference. While there have been many examples of XAI systems that have failed to have the intended impact of users, it is often the case that these tenets of HCXAI were overlooked. XAI is a relatively young field of research that has yet to find its footing, even as the landscape of black box AI systems is rapidly evolving. It is not yet time to give up hope on XAI. Instead, we invite you to adopt critical reflection and value-sensitivity into XAI research and evaluation, making it human-centered.

Will Human-centered XAI solve all our problems? No, but it will help us ask the right questions.

ACKNOWLEDGMENTS

With our deepest gratitude, we acknowledge the time of all participants of all the studies reported here. Without their input, these projects would not have been possible. We thank reviewers for their valuable input. We also want to thank the organizations, the sites for the case studies, for their cooperation. We are grateful to members of the Human-Centered AI Lab at Georgia Tech whose continued input refined the conceptualizations presented here. We are indebted to Justin Weisz for his editorial feedback that helped scope the project appropriately. This project was partially supported by the National Science Foundation under Grant No. 1928586.

REFERENCES

[1] P Agre. 1997. Toward a critical technical practice: Lessons learned in trying to reform AI in Bowker. *Social science, technical systems, and cooperative work: Beyond the Great Divide* (1997).
[2] Philip E Agre. 1997. *Computation and human experience.* Cambridge University Press.
[3] 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.
[4] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115.
[5] Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. 2019. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012* abs/1909.03012 (2019). arXiv:1909.03012 http://arxiv.org/abs/1909.03012
[6] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José MF Moura, and Peter Eckersley. 2020. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 648–657.
[7] Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. 2023. Language models can explain neurons in language models. https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html.
[8] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.
[9] Davide Castelvecchi. 2016. Can we open the black box of AI? *Nature News* 538, 7623 (2016), 20.
[10] Matthew Chalmers and Ian MacColl. 2003. Seamful and seamless design in ubiquitous computing. In *Workshop at the crossroads: The interaction of HCI and systems issues in UbiComp*, Vol. 8.
[11] Daniel Clement Dennett. 1989. *The intentional stance.* MIT press.
[12] Shipi Dhanorkar, Christine T Wolf, Kun Qian, Anbang Xu, Lucian Popa, and Yunyao Li. 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*.
[13] Paul Dourish. 2004. *Where the action is: the foundations of embodied interaction.* MIT press.
[14] Paul Dourish, Janet Finlay, Phoebe Sengers, and Peter Wright. 2004. Reflective HCI: Towards a critical technical practice. In *CHI'04 extended abstracts on Human factors in computing systems*. 1727–1728.
[15] Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. 2017. UX Design Innovation: Challenges for Working with Machine Learning as a Design Material. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17* (2017), 278–288. https://doi.org/10.1145/3025453.3025739
[16] Upol Ehsan, Q Vera Liao, Michael Muller, Mark O Riedl, and Justin D Weisz. 2021. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–19.
[17] Upol Ehsan, Q Vera Liao, Samir Passi, Mark O Riedl, and Hal Daume III. 2022. Seamful XAI: Operationalizing Seamful Design in Explainable AI. *arXiv preprint arXiv:2211.06753* (2022).
[18] Upol Ehsan, Samir Passi, Q Vera Liao, Larry Chan, I Lee, Michael Muller, Mark O Riedl, et al. 2021. The who in explainable ai: How ai background shapes perceptions of ai explanations. *arXiv preprint arXiv:2107.13509* (2021).
[19] Upol Ehsan and Mark O Riedl. 2020. Human-centered Explainable AI: Towards a Reflective Sociotechnical Approach, In International Conference on Human-Computer Interaction. *arXiv preprint arXiv:2002.01092*, 449–466.
[20] Upol Ehsan, Koustuv Saha, Munmun De Choudhury, and Mark O Riedl. 2023. Charting the Sociotechnical Gap in Explainable AI: A Framework to Address the Gap in XAI. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–32.
[21] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O Riedl. 2019. Automated rationale generation: a technique for explainable AI and its effects on human perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 263–274. https://doi.org/10.1145/3301275.3302316
[22] Upol Ehsan, Philipp Wintersberger, Q Vera Liao, Martina Mara, Marc Streit, Sandra Wachter, Andreas Riener, and Mark O Riedl. 2021. Operationalizing human-centered perspectives in explainable AI. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–6.

[23] Upol Ehsan, Philipp Wintersberger, Q Vera Liao, Elizabeth Anne Watkins, Carina Manger, Hal Daumé III, Andreas Riener, and Mark O Riedl. 2022. Human-Centered Explainable AI (HCXAI): beyond opening the black-box of AI. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–7.

[24] Jerry A Fodor. 1975. *The language of thought*. Vol. 5. Harvard university press.

[25] Batya Friedman, Peter H Kahn, and Alan Borning. 2008. Value sensitive design and information systems. *The handbook of information and computer ethics* (2008), 69–101.

[26] William W Gaver, Jacob Beaver, and Steve Benford. 2003. Ambiguity as a resource for design. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 233–240.

[27] MD Romael Haque, Katherine Weathington, Joseph Chudzik, and Shion Guha. 2020. Understanding Law Enforcement and Common Peoples' Perspectives on Designing Explainable Crime Mapping Algorithms. In *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*. 269–273.

[28] Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. 2019. Towards realistic individual recourse and actionable explanations in black-box decision making systems. *arXiv preprint arXiv:1907.09615* (2019).

[29] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376219

[30] David K Lewis. 1986. Causal explanation. (1986).

[31] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.

[32] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1–15.

[33] Q Vera Liao, Milena Pribić, Jaesik Han, Sarah Miller, and Daby Sow. 2021. Question-driven design process for explainable ai user experiences. *arXiv preprint arXiv:2104.03483* (2021).

[34] Q Vera Liao and Kush R Varshney. 2021. Human-centered explainable ai (xai): From algorithms to user experiences. *arXiv preprint arXiv:2110.10790* (2021).

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

[36] Sina Mohseni, Niloofar Zarei, and Eric D Ragan. 2018. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *arXiv* (2018), arXiv–1811. https://doi.org/10.1145/3387166 arXiv:1811.11839

[37] George Nott. 2017. Explainable artificial intelligence: Cracking open the black box of AI. *Computer world* 4 (2017).

[38] Andrés Páez. 2019. The pragmatic turn in explainable artificial intelligence (XAI). *Minds and Machines* 29, 3 (2019), 441–459.

[39] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2018. Manipulating and measuring model interpretability. *arXiv preprint arXiv:1802.07810* (2018).

[40] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. 2022. Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI. *arXiv preprint arXiv:2204.01075* (2022).

[41] Jakob Schoeffer and Niklas Kuehl. 2021. Appropriate fairness perceptions? On the effectiveness of explanations in enabling people to assess the fairness of automated decision systems. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*. 153–157.

[42] Phoebe Sengers, Kirsten Boehner, Shay David, and Joseph'Jofish' Kaye. 2005. Reflective design. In *Proceedings of the 4th decennial conference on Critical computing: between sense and sensibility*. 49–58.

[43] Ronal Singh, Tim Miller, Henrietta Lyons, Liz Sonenberg, Eduardo Velloso, Frank Vetere, Piers Howe, and Paul Dourish. 2023. Directive explanations for actionable explainability in machine learning applications. *ACM Transactions on Interactive Intelligent Systems* (2023).

[44] Simone Stumpf, Adrian Bussone, and Dympna O'sullivan. 2016. Explanations considered harmful? user interactions with machine learning systems. In *ACM SIGCHI Workshop on Human-Centered Machine Learning*.

[45] Jiao Sun, Q Vera Liao, Michael Muller, Mayank Agarwal, Stephanie Houde, Kartik Talamadupula, and Justin D Weisz. 2022. Investigating Explainability of Generative AI for Code through Scenario-based Design. In *27th International Conference on Intelligent User Interfaces*. 212–228.

[46] Kaige Xie, Sarah Wiegreffe, and Mark Riedl. 2022. Calibrating trust of multi-hop question answering systems with decompositional probes. *arXiv preprint arXiv:2204.07693* (2022).

[47] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining whether, why, and how human-AI interaction is uniquely difficult to design. In *Proc. CHI*. 1–13.

[48] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. ACM, Association for Computing Machinery, New York, NY, USA, 295–305. https://doi.org/10.1145/3351095.3372852