Charting the Sociotechnical Gap in Explainable AI: A Framework to Address the Gap in XAI

UPOL EHSAN, Georgia Institute of Technology, USA
KOUSTUV SAHA, Microsoft Research, Canada
MUNMUN DE CHOUDHURY, Georgia Institute of Technology, USA
MARK O. RIEDL, Georgia Institute of Technology, USA

Explainable AI (XAI) systems are sociotechnical in nature; thus, they are subject to the sociotechnical gap—divide between the technical affordances and the social needs. However, charting this gap is challenging. In the context of XAI, we argue that charting the gap improves our problem understanding, which can reflexively provide actionable insights to improve explainability. Utilizing two case studies in distinct domains, we empirically derive a framework that facilitates systematic charting of the sociotechnical gap by connecting AI guidelines in the context of XAI and elucidating how to use them to address the gap. We apply the framework to a third case in a new domain, showcasing its affordances. Finally, we discuss conceptual implications of the framework, share practical considerations in its operationalization, and offer guidance on transferring it to new contexts. By making conceptual and practical contributions to understanding the sociotechnical gap in XAI, the framework expands the XAI design space.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; **User studies**; *Empirical studies in collaborative and social computing*; • **Computing methodologies** → **Artificial intelligence**.

Additional Key Words and Phrases: Explainable AI, Sociotechnical Gap, Human-AI interaction, Framework, Human-centered Explainable AI, FATE, Responsible AI, AI Ethics, organizational dynamics, AI governance, user study, participatory design

1 INTRODUCTION

Explainable Artificial Intelligence (XAI) systems are *sociotechnical* in nature because their technical (AI) components are embedded in social environments [3, 43, 87]. As a research area, XAI aims to provide human-understandable justifications for a system's behavior [4, 48, 65]. Given their sociotechnical nature and the increasing trend of being deployed in high-stakes domains like healthcare [29, 76, 78, 96], finance [102, 117], criminal justice [68, 80, 126], we need to account for both social and technical factors to mitigate biases and promote accountability in XAI [142].

All sociotechnical systems, XAI or otherwise, are subject to what Ackerman calls the *sociotechnical gap*—"the divide between what we know we must support socially and what we can support

Authors' addresses: Upol Ehsan, Georgia Institute of Technology, Atlanta, GA, USA, ehsanu@gatech.edu; Koustuv Saha, Microsoft Research, Montréal, Québec, Canada, koustuvsaha@microsoft.com; Munmun De Choudhury, Georgia Institute of Technology, Atlanta, GA, USA, munmund@gatech.edu; Mark O. Riedl, Georgia Institute of Technology, Atlanta, GA, USA, riedl@cc.gatech.edu.

technically" [3]. Ackerman posed the gap as a central challenge for sociotechnical systems, highlighting that without minding the gap, we cannot effectively design useful systems [3]. One of the most intriguing and challenging aspect of the sociotechnical gap is the difficulty to fully bridge permanently [3, 41, 116]. This is because the "social activity is fluid and nuanced" (which means user needs are dynamic) while the "technical systems are rigid and brittle", "not socially flexible", and "do not allow sufficient nuance", especially "in their support of the social world" [3].

With these challenges in mind, addressing the gap might appear futile at first: why should we bother addressing a gap that cannot be filled? However, therein lies a crucial point made by Ackerman—the goal of introducing the gap is *not* to "solve" this fundamental problem but to deeply understand it [3, 41]. Without understanding, we cannot address the gap. Without charting or mapping the gap—*how* the gap looks and *what* is on either side of it—we cannot understand enough to address it. The following analogy can be helpful: imagine two mountains with a canyon in the middle. If we do not know the geography of the gap formed by the canyon (e.g., where the gap is the largest or smallest), then it will be hard to find the best places to build a bridge. We need mechanisms to chart or map out the gap. This is what this paper does in the context of XAI—provide mechanisms to chart the technical and social dimensions of XAI systems so that we can better understand and address their sociotechnical gaps. Charting the gap can promote greater visibility of the design space and facilitate informed design decisions.

Charting the sociotechnical gap in XAI is challenging given AI is "a new and difficult design material" [42, 42, 61], especially given the non-deterministic, stochastic, and opaque-boxed nature of current systems [161, 162]. This implies that there is more volatility in mapping out the technical affordances of (X)AI-powered systems (vs. non-AI systems). Thus, there is a translational challenge in charting this gap in the context of XAI systems (as opposed to non-AI systems).

Inspired by Ackerman's goal, this paper proposes a shift in focus from a "gap filling" philosophy to a "gap understanding" one when addressing the sociotechnical gap in XAI. We argue that *engaging in the process of charting the gap can improve our problem understanding, which can reflexively provide actionable insights to improve explainability*. We operationalize our argument both conceptually and practically. At a conceptual level, we situate the process of charting the sociotechnical gap in XAI. By shifting the focus towards the utility of problem understanding, it can potentially highlight intellectual blinds spots in XAI, which can expand the design space. At a practical level, it can address the under-explored area of *how* we might go about mapping and using it to address explainability challenges. *We address the challenges around charting the sociotechnical gap in XAI by using two real-world case studies in distinct domains to empirically derive an analytical framework to systematically chart the gap*. This framework not only integrates existing guidelines in AI and HCI but also transfers the applications in the context of XAI, highlighting the interplay between the social and technical elements. We also explore how to connect the technical and social wings of the framework by utilizing emerging XAI concepts that take a sociotechnically-informed perspective on explainability. We apply our developed framework to a third case study in a new domain and showcase two main outcomes: (1) better understanding (charting) of the gap (how the gap looks) by being able to identify the most promising and challenging regions of the gap; (2) actionable insights to address the gap (what to do with the gap) that reflexively improved informed user actionability and system explainability. In summary, our contributions are fourfold:

- We situate the charting of the sociotechnical gap in the context of Explainable AI, expanding the design space
- Using two real-world case studies in different domains, we empirically derive a framework to chart the gap and apply it to a third case study in a new domain
- Our framework connects diverse threads of guidelines in AI and HCI, translates them to the context of XAI, elucidates how to use them to address the gap by connecting the social and technical aspects of the infrastructure

- We discuss the conceptual implications of the framework, share practical considerations in its operationalization, and offer guidance on transferring it to new use case domains.

Our work aims to help stakeholders chart the contours of sociotechnical gaps in XAI systems. The design philosophy of our framework is generative and iterative, not normative [33, 82, 107]. By "refocusing of design in terms of process rather than solutions" [137], it can help stakeholders avoid certain techno-centric pitfalls like Solutionism (always seeking technical solutions) [114] and Formalism (over-abstracting and seeking mathematical solutions) [63]. This paper is not an exhaustive treatise of the sociotechnical gap in XAI; rather, it takes a foundational step towards addressing it by operationalizing the charting of the gap.

2 BACKGROUND AND RELATED WORK

2.1 Sociotechnical Gap in HCI, CSCW, and AI systems

Walker et al. defined sociotechnical systems as the ones that include the interrelatedness of "social" and "technical" aspects. They argued that "an inevitable consequence of mixing 'socio' with 'technical' is that the socio does not necessarily behave like the technical, people are not machines; paradoxically, with growing complexity, even the 'technical' can start to exhibit non-linear behavior [153]. Ackerman, in his celebrated work [3], argued that there is an inherent gap between the social requirements of Computer-Supported Cooperative Work (CSCW) and its technical mechanisms, and this socio-technical gap concerns the divide between what we know we must support socially and what we can support technically. Over the years, evidence of the socio-technical gap has surfaced in various aspects and examples of Human-Computer Interaction (HCI) and CSCW technologies, such as computing affordances and data privacy [17, 55, 141].

AI-based computing systems are socially situated, and research recognizes the detrimental effect of a techno-centric view on AI [127, 139, 149]. If both social and technical factors are not considered in a balanced fashion, AI systems are hard to be integrated into individual and organizational workflows [105, 129, 157]. Further, ignoring social factors could lead to potential misuse, mistrust [25, 163, 164], and ethical risks and unintended consequences [111, 132, 147]. Prior work has explicitly highlighted the abstractions and traps with building AI-based computing systems in sociotechnical contexts [10, 63, 114, 137]. Notably, Selbst et al. discussed ways in which technical designers can mitigate the traps and draw abstraction boundaries to include social actors rather than purely technical ones [137]. Andrus et al. explored the many facets associated with developing AI for public interests, and how the *sociotechnics* of such AI systems are closely inter-related with three contemporary areas of Fair ML, AI Safety, and Human-in-the-Loop Autonomy [10].

Similarly, Alter published a conceptual paper to bridge the gap between thinking of systems as tools and thinking of systems as sociotechnical systems with human participants [8]. Munson et al. drew upon Ackerman's sociotechnical gap in CSCW concepts to review contemporary sociotechnical challenges of using social media to support health [116]. Malatji et al. examined the socio-technical gaps within organizational information and cybersecurity practices [106].

Ackerman posed the sociotechnical gap as a central challenge in CSCW [3]. While there are commendable efforts to grapple with the sociotechnical challenges in XAI [38, 43, 44, 87, 154], addressing the gap remains understudied. To our knowledge, this is the first paper that explicitly focuses on charting and addressing the sociotechnical gap within the context of explainable AI systems. Drawing upon the rich literature in CSCW, the case studies in this paper focus on AI-supported work that is not only *cooperative* but also bears *collaborative* aspects (e.g., cross-functional teams). In attempting to chart the sociotechnical gap in XAI, this paper thus builds on and extends the CSCW community's long-standing tradition of designing and studying explainable computing systems [52, 69, 83, 93, 101, 122, 135].

2.2 Explainable AI and its Sociotechnical Aspects

Broadly, an AI system can be considered to be explainable (or XAI), if it aims to make its decisions *easy to understand and interpret* by people [12, 27, 48, 62, 66, 95, 108, 123]. Explainability is often viewed more broadly than model transparency [62, 95, 123]. Recent XAI work showed post-hoc explanations [48, 95] and algorithmic auditability [62] can build trust in the AI systems. Taken together, explainability is a human-factor, not just a model-inherent property [12, 13, 45, 108, 112]. Therefore, the importance of adopting user-centered approaches to XAI has been advocated in recent research [47, 108, 139, 149]. Wang et al. reviewed decision-making theories and identified many gaps in XAI output to support the complete cognitive processes of human reasoning [154].

The HCI and CSCW communities have a long-standing interest in making computing systems explainable [9, 50, 60, 79, 94, 100, 155]. Abdul et al. studied HCI research on explainable systems spanning across expert systems, recommenders, context-aware technologies, and ML systems. In a series of works, Lim and colleagues have showcased the importance, opportunities, and challenges of AI intelligibility in context-aware systems [90–92, 94]. Miller noted the need to push XAI towards human-centered instead of algorithmic-centered approaches by calling out the gaps between XAI algorithmic output and properties of explanations sought by individuals [108]. Further, Doshi-Velez and Kim noted the need to involve intended stakeholders in the intended usage context [40].

Adopting human-centered approaches, recent works have studied explainable interfaces to help model developers diagnose and improve ML models [74, 79]. Kaur et al. examined how data scientists frequently misuse and over-trust interpretability tools [79]. Explainability features of AI systems have been explored in various domains, including health [154, 159], and recruitment/hiring [30]. The value of multi-stakeholder involvement has also been noted in content moderation [84] and active learning [61]. Complementarily, researchers have started investigating and evaluating the effectiveness of XAI techniques [7, 23, 26, 39, 46, 160]. However, XAI's disconnect with the philosophical and psychological grounds of human explanations has been duly noted [110], as best represented by Miller's call for leveraging insights from the Social Sciences [108]. Interestingly, while XAI is often claimed to be a critical step toward accountable AI, empirical studies have found little evidence that explanations improve a user's perceived accountability or control over AI systems [121, 142]. Given explainability is a human factor and XAI systems are sociotechnical, our work expands the design space of the sociotechnical needs of XAI systems.

2.3 Operationalizing User Needs in AI

As AI systems get more integrated into societies, they require the best practices to be operationalized and implemented. Along these lines, Amershi et al. proposed 18 design guidelines applicable for human-AI interactions [9]. There are several taxonomies of prototypical roles of XAI consumers [12, 71, 148], to provide valuable insights to tackle the challenges of selecting and translating between AI algorithms. Further, researchers have studied user-centered approaches to design XAI UX [51, 87, 158]. Liao et al. studied design challenges around XAI and proposed a question-driven design process to ground the user needs, choices of XAI techniques, design, and evaluation of XAI UX [87]. Recently, Ehsan et al. expanded the notion of explainability in AI systems by introducing and exploring the concept of *Social Transparency*, "a sociotechnically informed perspective that incorporates the socio-organizational context into explaining AI-mediated decision-making." [43]. The authors examined how socio-organizational information carried by the 4W—*who* did *what, when*, and *why*—design features can augment AI explainability [43]. Towards addressing the societal issues of algorithmic governance, Lee et al. presented a collective participatory framework that enables people to build an algorithmic policy for their communities [86].

Prior work has also advocated or proposed means to operationalize AI systems to be more human-centered and socio-organizationally situated [37]. Ehsan and Riedl adopted Critical Technical

Practice [6] and Reflective Design [138] to propose the foundations of a reflective Human-Centered XAI (HCXAI). Madaio et al. took a participatory approach to co-designing checklists to understand organizational challenges are AI ethics [103]. and Zhu et al. proposed Value Sensitive Algorithm Design (VSD) [170] by engaging stakeholders in the early stages of algorithm creation to avoid biases in design choices or compromising stakeholder values [170]. Taking an "AI lifecycle" perspective, Dhanorkar et al. underscore how the explainability needs of different types of explanation audiences evolve over time [38]. Prior work has also leveraged design fiction and speculative scenarios to understand user values and cultural perspectives for AI system design [31, 32, 115]. Šabanović proposed a framework of Mutual-Shaping and Co-production [127]. Jones et al. proposed a design process for intelligent sociotechnical systems with equal attention to analysis of social concepts in the deployment context and representing such concepts in computational forms. On the side of technical affordances, Sanneman and Shah designed a three-level framework for the development and evaluation of an explainable AI system's behavior [133].

In other works, Gebru et al. proposed *datasheets for datasets*, aimed at facilitating better communication between dataset creators and dataset consumers, and encourage the prioritization of transparency and accountability [59]. Sokol and Flach proposed explainability fact sheets to enable researchers and practitioners to grasp the capabilities and limitations of a particular explainable AI method [143]. On the side of models, Mitchell et al. proposed model cards to encourage transparent model reporting around creation, curation, and evaluation [109]. On trust and AI ethics, Bellamy et al., Hind et al. have explored operationalizing it in organization settings [20, 72, 81].

Our work builds upon and contributes to the above body of work to propose a framework that can help identify, operationalize, and address the gaps in various social and technical components in explainable AI systems and AI explainability. For this, we are not only informed by the above perspectives but also integrate them to chart the sociotechnical gap in explainable AI systems and bridge it at an infrastructural level.

3 CASE STUDIES ON REAL-WORLD APPLICATIONS OF XAI

We share two case studies from different domains: sales (Section 3.1) and mental health (Section 3.2). These case studies represent real-world projects spanning two years of engagement by one or more of the authors. For each case study, a background section provides necessary context, the problem, descriptions of the interplay between the social and technical components (which lead to scoping the gap), and steps we took to address the mapped out sociotechnical gap. While outlining the interplay, we focus on relevant aspects which ultimately help construct a framework (Section 4) that helps us map out the sociotechnical gaps in new use XAI use cases. Providing an exhaustive list of technical attributes and social needs is neither productive, nor necessary, nor potentially possible (the list might be uncountably large).

Instead of offering a top-down narration of providing the framework first, it may be helpful to the reader to take a bottom-up "build-up" approach to build up the contextually relevant parts first and allows us to link back to the concepts later. This approach is also representative of the real-life events of how we arrived at the framework. We ground our data and motivation for the building blocks of the framework through the first case study and use those lenses to describe the second. The structuring can be potentially helpful to the reader for transferring the process to other cases. We structure the interplay in a way that serves as building blocks for charting the sociotechnical gap and the eventual development of the framework. We not only share the two case studies to empirically derive the framework but also apply the framework through another real-world case study in a different domain (cybersecurity) (Section 5). We offer an end-to-end, conception-to-implementation, narrative of how charting the gap can help understand and address it. We also discuss how the framework can be transferred to other domains in Section 6

3.1 Case Study 1: Kuro, the AI-powered sales software

3.1.1 Background. SalesCorp (pseudonym), a Fortune 100 multi-national technology company, invested US $5 million to build an AI-powered pricing recommendation system to help its technology salesforce. Each sale typically ranges from US $500K to $2M; therefore, the stakes are high. If a client is priced out of their budget, SalesCorp might lose the immediate sale and risk losing long-term revenue from the client. SalesCorp follows a business-to-business (B2B) model and sells hardware, software, and cloud infrastructures. It follows a dynamic pricing model, dependent on multiple factors like a client's historical budget, company size, length of the relationship, etc.

The AI system, Kuro (pseudonym), was introduced as a "virtual experienced peer who is always there" and claims to utilize over 50,000 data points to drive recommended prices. The system was developed in a 70-30 partnership, 70% developed in-house using proprietary data and 30% developed with third-party vendors that offer AI infrastructural support (e.g., proprietary models licensed for exclusive use). A promised value proposition of AI-powered Kuro was its robustness, an alleged improvement from its "brittle" predecessor Sumo.

3.1.2 The Problem. On the surface, Kuro's metrics were impressive: 90% accuracy, good model transparency (showing top 3 features), and confidence range (0-100%) shown next to each recommendation. However, there was a breakdown — only 10% of the sellers used Kuro; of those who used it, only 2% found it useful. Given the socio-organizational context of the deployment, the problem here was not a purely technical one; it had strong social components (as we will see below). Considering Kuro's problems, SalesCorp recruited an author of this paper to work on this issue.

Chronicling a year-long collaboration, below we motivate and ground our methods and the data that ultimately form the buildings blocks the *social* and *technical* wings of the framework. Note that we structure the presentation of the following events to help the reader find correspondence between the case studies below and the framework derived from it (in the next section). Even though we structure the two prongs sequentially in our write-up, they took place in a largely parallel fashion, iteratively building on top of each other to provide a clearer picture of the sociotechnical gap. To protect client confidentiality, proprietary details are redacted from quotes.

3.1.3 Grounding the Data, Process, & Motivations for the Building Blocks for the Framework. Recognizing the socio the sociotechnical nature of the infrastructure, we focused on two relevant areas of the sociotechnical gap: understanding *the technical affordances* and *the socio-organizational* needs. To understand these two aspects, we took the following approach.

Taking a participatory approach [136], we conducted 2 workshops with 28 participants each. Participants had diverse roles (5 XAI Engineers, 6 Product Managers, 5 Responsible AI Managers, 3 Data Scientists, 4 VPs of Technology, and 5 UX Researcers). We took a participatory approach to leverage the domain expertise of our stakeholders and have their voices actively represented in our interventions. During orientation, we introduced the concept of the *sociotechnical gap* to our stakeholders so that we can focus on the divide between what the current system supports technically and what we must support socially. Almost half of them were already aware of the concept of the sociotechnical gap. Moreover, using examples of related work [38, 43, 88, 89], we oriented them to a human-centered XAI lens, one that does not restrict XAI to merely model transparency and incorporates the sociotechnical factors in its conceptualization [45]. This helped the participants adopt a sociotechnical stance (vs. a techno-centric one) in problem-solving.

Below we share **key takeaways** from each workshop to ground our data and motivate the structure of the framework.

Workshop 1: constructing building blocks. The goal in this 3-hour workshop was to *brainstorm and reach consensus on the relevant building blocks to consider when thinking about technical and social sides of the problem* (later forming the foundations of the resulting framework in Sec 4).

The first half of the workshop consisted of a generative (brainstorming) exercise where four cross-functional teams used sticky notes to list out the relevant topics (building blocks). In this stage, we explicitly asked our participants to focus on the generation of ideas and not an evaluation of them. The teams generated 7 blocks under the technical wing and 12 under the social side.

In the second half, we concentrated on narrowing down the building blocks in a way that balances practical feasibility with conceptual rigor. All four teams reported that a total of 19 blocks were too many to be practically useful. Next, we paired up the teams — four teams became two; once each of the two teams reached a consensus, two teams became one large team. Given the cross-functional nature of each team, the discussion challenged assumptions in a positive way relayed by this data scientist, "It's pretty refreshing to have your views challenged. I never really considered why I need to know the social aspects. This process helps me see exactly why I was tunnel-visioned." An iterative affinity diagramming process generated three thematic blocks for the technical side and three for the social side (we did not put any requirement for balancing the number of blocks).

The final task was to figure out the names for these blocks. We asked them to share the top function of each block. On the **technical** side, they listed: **data** genealogy (where the data comes from, how was it collected, etc.), **model**/algorithmic affordances (what can or cannot the model do), and how the AI system generates **explanations**. On the **social** side, they outlined: **trust** (how much or how little users rely on the AI), **actionability** of the explanations (what users need to act on the AI-generated explanations), and organizational **values** (how organizational incentives and values interplay with individual expectations). Thus, the group settled on the final blocks—on the technical side, there were data, models, and explanations; on the social side, we had trust, actionability, and values. We use these building blocks as analytic lenses to structure the case studies and also as foundations for the framework (Section 4).

Workshop 2: what is inside each building block. In this 4-hour workshop, the goal was to *operationalize each block in a systematic manner.* What questions or guidelines may we use to practically explore each block that will increase our understanding of the sociotechnical gap? The first step was to collaboratively explore sets of questions or guidelines we can use to flesh out the idea behind each block. To achieve this goal, initially, we proposed custom-made questionnaires (that we developed) for each block. However, we received substantial push-back against introducing new materials without adequately leveraging existing workflows and processes. For instance, the company had existing processes where engineers and data science teams prepared paradata documentation like Datasheets for Datasets [59]. Citing adoption concerns for new questionnaires, one product manager critically asked "who will say yes to doing new work without exploring how far we can go with what we have?" Building on this feedback, we pivoted into leveraging existing processes and documentation. Similar to the first workshop, we split participants into four cross-functional teams with the common mandate of finding existing processes and guidelines relevant to different building blocks. Next, we paired up the teams to facilitate the iterative filtering till a consensus was reached (four teams became two and two became one). The teams were able to highlight important. For instance, teams agreed that we could use Model Cards [109] for the model block and Datasheets for Datasets for the Data block.

The second step was to decide how many questions or guidelines should go under each block and how we may cluster them. To balance practical viability and burden, we received feedback to *not* generate an exhaustive list of questions. As a data scientist put it: "give us *starter questions* [from different guidelines] that kickstarts the conversation… [and] allow us to filter questions from the remaining set" (emphasis added). For each block, using thematic analysis [22], we worked with participants to incorporate relevant "starter" questions from different guidelines. Next, we used affinity diagramming [19, 21] with stakeholders to cluster relevant questions for each topic under

each wing (e.g., which ones from Datasheets for Datasets are most relevant). Last, we incorporated the feedback of 6 XAI and HCI experts to refine the final list of starter questions.

This grounding forms the foundation of the resulting framework (Section 4). Equipped with this understanding, we now share details on how we explored the technical and socio-organizational affordances to solve the problem with Kuro.

3.1.4 Understanding the technical affordances. There are three building blocks to unpack the technical wing: *data*, *model*, and *explainability*.

For **data**, using a participatory approach, we appropriated the questionnaires from Datasheets for Datasets [59] and customized it for the company, capturing insights on the data genealogies.

For the **model**, we hit "off limits" spots because the vendors supplied some of the decision mechanisms. When possible, we conducted interviews with account managers from the vendors and adapted Model Cards [109] to our use case. While we could not exhaust each entry from the original template, the empty slots were just as useful as the ones with data. It was an explicit reminder of what we do not have, which ultimately helps us scope the technical affordances.

For **explainability**, we leaned on two main artifacts – first, we selected relevant questions from Liao et al.'s XAI Question Bank [87]. Combined with the Datasheets and Model Cards, the questions provided a good starting point to explore different facets such as explanation type (how, why, what-if, why not, etc.). Second, we adapted IBM's Explainability Fact Sheets [20] to get a sense of the explanation generation mechanism (modality, global vs. local, etc.) at functional and validation levels. We had to use a subset because the original Fact Sheets had a list 36-item worklist. As one developer put it, "it was too much trouble for too little value."

3.1.5 Understanding the socio-organizational needs. The social side happened in tandem with the technical side. We want to get clear feedback on why the engagement was at 10%. To unpack this problem, we used a combination of interviews, Envisioning Card exercises [57], and workshops to understand three main things: trust in AI (what is harming? What is helping? What else is needed?), actionability of the explanations (what could help users take informed actions on the AI's decision), and values (tensions and alignments with organizational norms). When investigating **trust**, we learned that while the model transparency was helpful, it was not enough to calibrate the trust in AI. Less experienced sellers tended to see the multi-million-dollar Kuro as "godlike". They tended to over-trust the AI and felt they lacked the voice/agency to disagree with the AI. Instead of putting them in that position, they chose to completely ignore/not engage with the AI. More experienced sellers simply ignored the AI's recommendation because they felt they "knew better". Understanding the nuances of trust helped us scope the social needs better.

Thinking of **actionability**, Kuro's explanation information was deemed "inert". One said, "I've no idea of what to do with the numbers [Kuro] shows me. How is it confident? Confidence with respect to what?" Moreover, there was a mistake in the assumed interaction paradigm—it was 1 Human -1 AI, whereas the activity of selling is a group one (many-to-many). Many sellers worked around this siloed constraint by using analog means. They would walk over to other members of the team (each client was typically served by a team of 5 sellers of differing hierarchies) to consult.

When considering **values**, we found multiple avenues of tensions in organizational norms. Overriding Kuro's decision was extremely cumbersome—the audits and escalations involved created a disproportionate burden on the sellers, which is why they consistently ignored it. Moreover, entry-level sellers had no idea of what "acceptable" changes to Kuro's prices are. The sellers perceived the technical staff as "invaders", ones who want to automate the sellers out of their jobs. This created a tense relationship, hindering collaboration between the stakeholders.

3.1.6 Addressing the gap. Compared to when we started, these processes painted a clearer picture of the two sides of the sociotechnical gap and highlighted opportunities to use design as a bridge.

While we acknowledged that the gap would be ever-present, the goal was never to "fill" it, but to understand the boundaries and act on it through what Ackerman calls first-order approximations—"tractable solutions that partially solve specific problems with known trade-offs" [3].

A core takeaway that connects trust, actionability, and value tensions was sellers' desire to know how others acted on the AI's recommendation. They wanted "peripheral vision" of past user trajectories. For example, a junior seller would want to see what a senior person did with the AI system and why. This additional context, the seller felt, would help them judge the AI better, making the explanations actionable.

We tested if adding socio-organizational context would indeed help. In an ad-hoc manner, we offered a piece-meal approach. We asked the sellers to pin a Google Sheet (online collaborative spreadsheets) in their slack channel and put the following information: for any recommendation from Kuro that was borderline, they would notate the incident ID, what they did (accept/reject), and write a Tweet-length justification (sellers appreciated the "Tweetification" of the "why"). Within one week, the Google Sheet had over 400 entries. Sellers even appropriated the original format by adding "+1" (like) voting to useful comments. We saw the engagement rise to 42% (from 10%). This was particularly surprising because the sellers went out of their workflow to engage with the Google Sheets. Interviews showed that they found enough value in this "peripheral vision" to make it worth overcoming the context-switching burden.

With a proof of concept in place, we could propose engineering changes. Through participatory design activities, we designed an intervention where next to Kuro's AI recommendation and technical transparency, the seller would be able to see *who else did what, when*, and *why*. While developing the intervention, we followed the framework for Social Transparency [43] to foster calibration of trust in AI and consistency in collective actions.

After implementation, we carefully monitored engagement, which had improved to 87%, a remarkable jump from 10%. Moreover, sellers shared that they understood Kuro better because the AI's decision was now in context of human actions. The additional context cultivated a holistic form of explainability, which was not possible purely through technical transparency. Through follow-up interviews, surveys, and workshops, our conception of the boundaries of the gap got refined, which brought previously unconscious aspects into conscious awareness, thereby making it actionable. Our actions reflexively made Kuro's explanations actionable, improving their explainability.

3.2 Case Study 2: An AI-Powered Clinical Mental Health Tool

3.2.1 Background. The global burden of mental illnesses accounts for 32% of years lived with disability, making these conditions a major contributor to the global burden of disease [152]. While timely treatments are useful, mental heath treatment heavily relies on *what* the patient tells their clinician during in-person consultations. However, clinicians have often reported experiencing tension between a patient's self-report and external reality [15]. Fisher and Appelbaum [54] provided case reports where clinicians had begun to incorporate patients' electronic communications, such as social media, as a new form of collateral information. Prior work in social computing and HCI has shown that, from such data, AI and machine learning can help to infer whether an individual is vulnerable to mental health conditions [28, 67, 130, 131], such as major depression [35] and postpartum depression [36]. An emergent body of work in digital mental health has also argued how the computational power harnessed by AI systems could be leveraged to reveal the complex psychopathology of psychiatric disorders and thus better inform therapeutic applications and collaboration across different clinicians [85, 165].

3.2.2 The Problem. Although many patients already bring social media data to their appointments with clinicians and patients [16], currently, there is no technology to support these interactions, especially given that different members of the clinician team would have different needs to fulfill

the same types of data. To support realizing the potential of social media-derived AI insights at the point of care, in this case study, clinical researchers and practitioners from a large, urban, nationally-recognized health system in the Northeast of the U.S. collaborated with one of the authors of the current paper [165, 166]. Their *central goal* was to investigate if AI, when applied to these data voluntarily contributed by patients, could provide different faceted views (e.g., through a dashboard) that would support the varying needs of different treatment team members.

A prototype of the tool, developed via a co-design approach through a year-long collaboration and by harnessing voluntarily shared social media (Facebook) data of consented mental health patients, was tested with various types of mental health clinicians at this large health system through 1-1 interviews (*N*=13) and focus groups (*N*=13). This case study revealed promise for the prototype but also highlighted issues such as liability concerns around using a tool that was a "black-box" to the clinicians. They imagined scenarios where despite access to patient data that indicates an exacerbation of symptoms, clinicians might not be in a position to take action. Like with the previous case study, we outline the *social* and *technical* parts of the *infrastructure*.

3.2.3 Understanding the technical affordances. The interviews and focus groups revealed a variety of technical features that the clinicians deemed important from the perspective of supporting collaborative treatment to their patients. Following the style adopted for the previous case study, we discuss these technical affordances along the same four dimensions of *data*, *model*, and *explainability*.

Focusing on the **data**, the study highlighted important issues around mental health symptoms and clinical risk factors. The prototype was deemed promising to the clinician participants; however, they desired that the tool needs to facilitate collectively identifying and visualizing the range of different markers of the patient's mental health state that might be differentially valuable to different clinicians, as a patient navigates their interactions with psychiatrists, therapists, and social workers.

For the **model**, participants noted credibility issues regarding the AI-derived mental health insights the prototype provided. There were two dominant credibility issues that could diminish trust: if social media data correlates to actual life (construct validity) and if an AI applied on top of this data can distinguish different contexts and intents behind specific posts (clinical utility).

Finally, turning to **explainability**, the participants said that the tool can be useful to compare and negotiate what the social media data shows and what the patient verbally reports, especially if the two are mutually conflicting. It can also help one clinician understand how and why another member of the team appropriated certain social media-derived AI insights but may have chosen to ignore others. These technical needs necessitate incorporating explainability features in the tool.

3.2.4 Understanding the socio-organizational needs. In tandem with the technical needs, the interviews and focus groups also discovered a range of social and organizational needs for the AI-based tool to be successful at the point of care. We discuss these below.

The clinician participants raised a number of concerns surrounding **trust** in use of the tool. In the absence of an implicit level of trust in the functioning of the algorithms, clinician participants felt that they would consider the social media based AI insights with caution. Importantly, participants were concerned that trust in the tool's abilities and clinical usefulness might be undermined by the observer effect [34, 128], wherein patients may stop posting or begin to self-censor themselves on social media, knowing clinicians' awareness of and access to this information.

Speaking of **actionability**, different clinicians may also act upon the presented information in the tool in different ways: a therapist could use this data as a complement to reports from family members since the supposed "objectivity" of the AI insights could resolve and reconcile different collateral sources to construct a more accurate picture of the patient's status. In addition, some participants speculated if the tool could play some preventive role where it allowed comparing

a particular pattern of symptoms (say, suicidal thoughts) prior to hospitalization with current patterns, this could be useful for engaging preventive crisis intervention services.

Lastly, a number of conflicts in **values**, whether between the patient and the clinician, or different members of the clinician team were found. With regard to clinician-clinician value tensions, the case study unraveled underlying issues of burden. Participants were worried about the potential burden on their work hours and expectations in the use of the tool, perhaps more for some types of clinicians than others. Others felt that it may not always be feasible to review patients' social media information prior to consultations because patient loads can be exceedingly high. Overburdened clinicians (e.g., social workers) often adopt a work style and value systems distinct from clinicians who provide medication assistance to a handful of patients.

3.2.5 Addressing the gap. We now outline some possible interventions to mind the sociotechnical gap in the context of this case study. Like the first case study, the goal of these interventions is to understand the boundaries of the technical and the social needs [3], rather than seeking to "fill" it.

First, many clinical teams schedule synchronous meetings periodically to engage in peer-review of treatment protocols and discuss discrepancies in the standard of care. If we integrate AI-driven insight from patients' social media data, we would need a way to summarize, synthesize, and visualize them for later use. To facilitate collaboration, we can envision future iterations that integrate these AI-driven insights with meeting notes into the Electronic Health Record (EHR) for later review. Together, these adaptations to the design of the tool can facilitate bringing not only transparency to these discussions but also improving the actionability of the AI.

Second, we learned that the processes underpinning *how* different clinicians acted on the AI insights in the prototype were missing. For instance, if a trend shows exacerbated symptoms and a therapist sees this information was not used by the psychiatrist, this information gap can hamper the therapist's treatment plan, create confusion, and diminish trust in the tool. The clinician participants therefore strongly felt that *transparency to support these social interactions* will be needed in the future. For instance, we can design filtered views that showcase the decision-making pathways relevant to the specific patient based on their specific symptoms or risk factors.

4 A FRAMEWORK TO CHART THE SOCIOTECHNICAL GAP IN XAI

Here we share our design lens in XAI, the framework's setup, derivation, and its building blocks.

4.1 Our Design Lens in XAI: Human-Centered Explainable AI

Given our goal is to chart the sociotechnical gap in XAI, we adopt a design lens in XAI that is sociotechnically-informed—that of Human-centered XAI (HCXAI) [38, 45, 89]. HCXAI expands the concept of explainability beyond the bounds of the algorithm [43] and positions it as a relational and audience-dependent construct instead of a model-inherent one [12, 13, 108, 112]. Given AI systems (Human-AI assemblages [38, 43]) exist in sociotechnical settings [88, 146], it takes more than just algorithmic transparency to make them explainable [49, 108]. Thus, explaining what is happening "inside the black box" often requires us to also understand things "outside the black box" [38, 87], requiring us to consider the entire AI lifecycle (vs. just the algorithm). For instance, why a facial recognition system disproportionately misclassified women of color [24] can be explained by looking at demographic compositions in the training data. Emerging work in HCXAI [69, 120, 135] showcases how a broader XAI perspective can potentially address criticisms of popular algorithm-centered XAI techniques, which can be ineffective [7, 119, 169] and potentially risky [79, 145].

4.2 Overall Setup to Derive the Framework

We shared the data grounding along with the process and motivation behind the building blocks in Section 3. The framework begins by explicitly acknowledging the sociotechnical gap through two
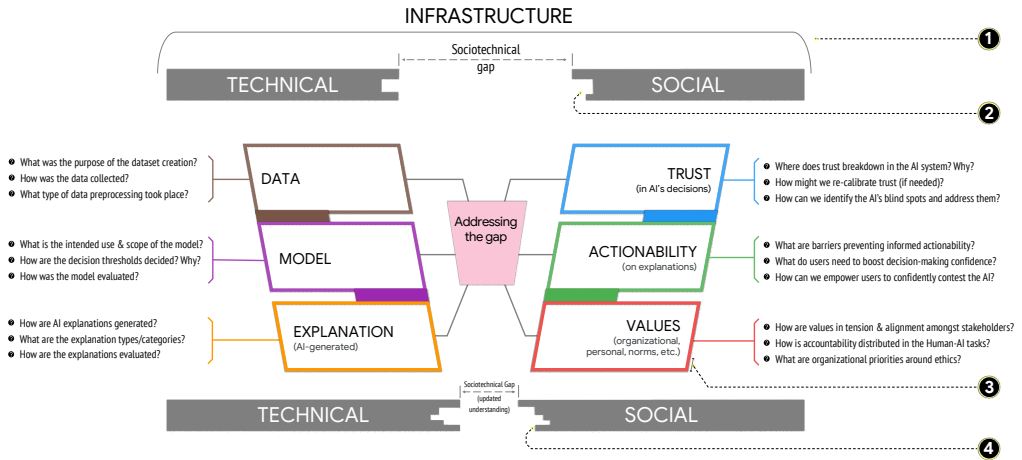
Fig. 1. Depicting the framework and its effect on the sociotechnical gap (in an XAI context) *before* and *after* the process. **(1)** *At the top*, we begin with the infrastructure consisting of both social and technical wings (grey boxes) with the sociotechnical gap in the middle. **(2)** Before charting the gap, we have a "rudimentary" idea of *how* its boundaries look like, depicted using the jagged edges. *In the middle*, we engage in the process of charting the gap with a "look inside" each wing by operationalizing the building blocks. The sequence of the blocks do not signify or impose any sequential limitations. As we will see during framework application in Sec. 5, the sequence of traversing the blocks is dependent on the problem and insights we gain from each block. **(3)** The slant in each wing is meant to convey that by engaging in the charting process, we gain actionable insights that address the gap (visually, bringing the two wings closer together). **(4)** Finally, *at the bottom*, having gone through the charting process, we emerge with a "higher resolution" understanding of the gap visually represented by the more fine-grained jagged edges of technical and social wings (different than the top grey boxes). The mapped out view reveals points where the gap is closer (highlighting promising areas to start addressing) and farther (challenging areas) from each other. *Overall*, the top represents a rudimentary idea of the gap before the mapping process, the middle signifies a "look inside" each half during the process, and the bottom depicts a refined and well-charted gap. Example "starter questions" are provided alongside each block that can jump-start its operationalization.

wings: the **technical** and the **social**. Figure 1 depicts the idea—each wing has three interconnected building blocks. On the *technical wing* are Data, Model, and AI-generated Explanations. On the *social side*, the building blocks are Trust in AI, Actionability of Explanations, and Values. To operationalize the goals for each block, we share plausible stakeholders (Table 1) at each step and offer "starter packets" with methodological recommendations and insights from existing guidelines. The starter questions provided next to each block in Fig. 1 should not be construed as being comprehensive because there is no one-size-fits-all formula in real-world AI deployments—the context, domain, and target audience govern how the process manifests. Example questions are meant to inspire and jump-start the generative process of charting the gap. The questions for each block were collaboratively scoped from existing guidelines (highlighted above) through our case studies. As Ackerman pointed out, the dynamic nature of user needs (social side) is primarily why the sociotechnical gap is hard to fully bridge [3, 41, 116]. Thus, problems often arise from the social side such as due to value tensions with AI's recommendation, users have difficulty acting on it.

Below we describe the operationalization of each building block, starting with the technical wing and followed by the social wing. As we highlighted in our case studies (in Section 3), even though we structure the two wings sequentially in our write-up, they are interconnected and iteratively build on top of each other to provide a clearer picture of the sociotechnical gap. The

Table 1. Summary of the framework to chart the sociotechnical gap in XAI.

Building Block	Goal	Operationalization Examples	Example Stakeholders
Technical Wing			
Data	To understand data genealogy, affordances, and scope what the data can tell us.	Datasheets for Datasets [59], AI FactSheets 360 [11], Dataset Nutrition Label [75]	Data Scientists, Data Analysts, Database Engineers
Model	To scope the model's affordances and limitations, which are important for trustworthiness.	Model Cards [109]	Data Scientists, AI Operations Engineers, Product Managers
Explanations (AI-generated)	To scope the affordances and limitations of AI-generated explanations especially around generation techniques, modality, and category.	XAI Question Bank [87]	Data Scientists, XAI Engineers, End-users
Social Wing			
Trust (in AI)	To understand the current level of trust and to scope calibrating trust.	Surveys [14, 73, 140], Interviews guided by the starter question, and Value Sensitive Design (VSD) exercises [58]	End-users including sales personnel and account managers
Actionability (of Explanations)	To understand what users need to act on the AI's explanation in an informed manner.	Social Transparency Framework [43]	End-users of the system (e.g., analysts or mental health specialists)
Values	To understand organizational values and their interplay with individual expectations.	VSD exercises, participatory design workshops	Multi-stakeholder involvement, including leadership, HR, developers, and end-users.

order in which the blocks appear (in the diagram or our writing) has no bearing on the order in which they can be operationalized (we can go back and forth, switch wings, etc.). The sequence of block traversal depends on the problem and findings at each stage; for instance, the framework application case study (in Section 5) will showcase how the process can start on the Actionability block (the social wing) and then move to the Data block (the technical wing). After describing the building blocks, we provide guidance on how to move around them using thematic labels (tags) that can keep track of the iterative process. Next, in terms of addressing the gap, we share strategies in the form of *palliatives* and *first-order approximations* [3] to address the gap. Finally, we share what end-products one might expect from engaging in this process. Beyond the mapping and charting affordances of the framework, a core contribution lies in integrating existing AI and XAI guidelines in the context of the sociotechnical gap.

4.3 The Technical wing

This wing of the framework aims to scope out what the system can afford and also cannot afford. We partition this space into *three* interconnected blocks: **data, model,** and (AI-generated) **explanations**. The motivation for this partition is two folds: first, the partition is theoretically situated in the spectrum of the technical infrastructure for XAI systems utilized by existing guidelines (e.g., Model Cards [109], AI Fairness 360 [20], XAI Question Bank [87], etc.). Second, it is empirically grounded in our case studies where we found that such a partition struck a balance between analytical demarcation and practical operationalization that resonated with stakeholders. This alignment promoted transferability from related guidelines, which facilitated the operationalization of each building block (detailed below).

Data: The goal of this building block is to understand data genealogy, affordances, and scope what the data can or cannot tell us (to the best extent available or possible). Data infrastructure is the food for the model, which entails it has fundamental implications on the entire lifecycle. Thus, we need a reliable and situated nutrition label for the data. To *operationalize* the goal, we can lean co-opt existing guidelines such as Datasheets for Datasets [59] and AI FactSheets 360 [11], and Dataset Nutrition Label [75]. Utilizing existing guidelines and adapting when needed, we are broadly interested in the origin story of the dataset. Fig. 1 shows some of the starter questions we can use to understand the purpose of the creation, duration of the data collection, types of preprocessing, how recent the dataset is, sampling information, etc. The starter questions often inspire follow-up questions that are more context-specific. It is important to balance breadth and depth when getting the data genealogy. The *stakeholders* of this process are variable across organizations but typically tend to be data scientists, data analysts, database engineers, etc.

Model: This block aims to scope the model's affordances and limitations, which are important for trustworthiness. Models are entities that transform the data into something of value (predictions, classifications, etc.). To *operationalize* the goal, we can co-opt guidelines like Model Cards [109]. Fig. 1 provides some starter questions around the model's intended purpose, architecture, performance evaluation, and datasets for evaluation. The starter questions are designed to seed deep generative discussions around specific use cases. The *stakeholders* are variable but typically tend to be data scientists, AI operations engineers (ones who validate and test models), product managers, etc.

Explanation: Here, we want to scope the affordances and limitations of AI-generated explanations, especially around generation techniques, modality, and category. We used the broad definition that an explanation is an answer to a why-question [97, 98, 108, 156]. To *operationalize* the goal, we can adapt some existing guidelines and tailor them for the use-case at hand. For instance, we can use aspects of Liao et al.'s XAI Question Bank [87] that provides a taxonomy of categories of explanations (global, local, counterfactual, etc.), their respective generation techniques, and connects them to different question types (how, why, what-if, why not, etc.) questions that facilitate user research. Fig. 1 shows a portion of the starter questions that we can use to scope the AI-generated Explanations space. The *stakeholders* typically tend to be data scientists, XAI engineers (who design the explanations), and end-users (who act on the explanations).

4.4 The Social Wing

This wing of the framework is meant to map out the social requirements of the stakeholders. We partition this wing into *three* interconnected blocks: **trust in AI** (towards the technology), **actionability** (at the individual level), and **values** (at the organizational level). This partition has two motivations: first, it is empirically situated in our case studies. Second, it is also aligned with the three levels of contexts (technical, individual, organizational) made visible by Social Transparency in AI – the perspective that incorporates the socio-organizational context into the conception of AI explainability [43]. This alignment is important because Social Transparency will play a part in connecting the two wings (detailed later) and addressing the sociotechnical gap.

As Ackerman pointed out, while the technical affordances (what the system can do) is often static, the social requirements are dynamic. The dynamic nature of user needs is primarily why the sociotechnical gap is hard to fully bridge. Often, problems arise on the social side—it could be issued where the AI system is over-trusted (technologically), users have difficulty acting on the AI explanations (individually), there are value tensions and conflicts between stakeholders, etc.

Trust (in AI): The goal here is to understand the current level of trust in the AI and scope how we might appropriately calibrate trust. Here, we are looking for situated baseline understandings of user trust in the AI and find points of breakdowns to act on. To *operationalize* the goal, we can use

existing surveys around trust in AI like [14, 73, 140], interviews guided by the starter question, and Value Sensitive Design (VSD) [56] exercises to get a sense of what aspects of the AI's performance matters to users, aspects where trust is missing, and how different facets impact the calibration of trust (when to trust the AI's decision vs. not), etc. Fig. 1 shows some of the starter questions that can guide the investigation. The *stakeholders* of this process are primarily end-users who need to act on the AI's decision. In the Kuro case, this would be sellers or account managers. While it's important to engage end-users, collaborating with data scientists and developers (secondary stakeholders) is important because they are often in charge of making changes.

Actionability (of Explanations): This block's goal is to understand what users need to act on the AI's explanation in an informed manner. Actionability is how users act on AI-generated explanations and it plays a role at the decision-making level [43]. Sometimes, even with well-functioning XAI systems, users do not feel empowered to decisive actions. Actionability resides at the junction between trust on technology and organizational values, where conflicted value systems can reduce the actionability of AI explanations. Recall that in the Kuro case, the audits required to override the AI increased the burden. To *operationalize* the goals, we can utilize insights offered by the Social Transparency framework that delineates how social validation (from peers), increased "peripheral vision" of past decisions, and support of follow-up actions can boost actionability and AI contestability (Fig. 1 shows some starter questions). To understand how these constructs should manifest in the system design, we can use similar methodological tools in the *Trust* block—interviews and VSD exercises. The *stakeholders* often include end-users of the system (e.g., analysts in an AI-powered cybersecurity system or mental health specialists in our use case).

Values: The goal here is to understand organizational values and their interplay with individual expectations. Recall that in the Kuro case, the misalignments in what the organization valued and the sellers wanted ultimately led to a breakdown in infrastructure. To *operationalize* the goal, we can utilize methodological tools such as VSD exercises, participatory design workshops, etc. to scope out points of conflict and value alignment between the organization and the individual, understand how accountability is distributed in the Human-AI collaboration, job performance expectations, understand positionality towards AI Ethics, etc. (Fig. 1 has some starter questions). This is a multi-stakeholder process that includes (but is not limited to) managers (in leadership), HR professionals, end-users, and developers.

While all the building blocks should be visited during the charting process of the gap, there is no prescribed order in which to visit each block. Each block is influenced by others, not just on the same wing but also across the gap. As shared earlier, the sequence of traversal is problem-dependent and insight-driven. Insights from addressing the Actionability block might prompt a visit to the Data block, updating our understanding, and warranting a revisit back to Actionability. We can iterate through the blocks, updating our understanding, till we achieve a stable saturation [134] of our understanding where further inquiries do not generate novel insights. To keep track of our progress and outcomes, we offer three thematic labels (tags)— *baseline understanding*, *updated understanding*, and *design recommendations*. *Baseline understanding* (BU) occurs the first time we visit a block and address it using the guidelines from our framework to get a sense of the status quo. *Updated understanding* (UU) can occur when revisiting the block with newfound insights from prior processes (e.g., BU of a different block). *Design recommendations* (DR) inform the practical development of tools, techniques, and policies (like first-order approximations). DRs can emerge from both baseline and updated understandings.

4.5 Addressing the Gap: Palliatives & First-order Approximations

Recall that there is no silver bullet to fully bridge the sociotechnical gap [116]. Moreover, success does not mean perfection in bridging the gap. To address the gap, we can use two approaches what

Ackerman calls *palliatives* and *first-order approximations* [3]. These approaches are not mutually exclusive and can be used in tandem.

Palliatives are interventions at the political, ideological, and education levels which impact how people perceive the affordances of the system [3, 116]. They are meant to ameliorate the gap without making system-level changes. They can include stakeholder analyses like the value-sensitive design exercises from our case studies. The findings can highlight blind spots in the system, which can then empower people to make informed decisions on usage [116]. As we saw in Kuro's case study, educational initiatives [99, 104]) can help bridge the gap where technology fails. As our case studies highlight, a competent system alone is not enough for people to find it useful. Combining analysis to understand where different organizational stakeholders are in the Technology Adoption Life Cycle [18, 113], we can use Technology Acceptance Models [150, 151] to understand perceived usefulness of the system. These steps can ameliorate certain disconnects between the social and technical elements and inform future technological interventions like first-order approximations.

First-order approximations are tractable solutions that partially solve specific problems with known trade-offs [3, 116]. Success with first-order solutions can involve connecting the social with the technical that *effectively* addresses some of the *most pressing* sociotechnical issues with *known* trade-offs. As our case studies showed, there are *multiple strategies* to develop first-order approximations. In Kuro's case, we illustrated how incorporating Social Transparency (ST) [43] acted as a first-order approximation to provide much needed "peripheral vision" and promote engagement. By providing visibility of who else did what, when, and why (4W), ST promotes explainability by making context visible at three levels (technological, individual, and organization), which align with the three blocks of the social and technical wings. With conceptual roots in XAI, ST can be a productive first-order approximation to address the gap for XAI systems. Using ST, increased visibility of the sociotechnical factors can provide actionable insights which can potentially address the gap. To operationalize ST, we can utilize the constitutive design elements – the 4W (who did what, when, and why) – to encapsulate relevant socio-organizational context.

4.6 Framework Outcomes

There are two main outcomes or end-products of our framework: (1) improved understanding of the gap (how the gap looks) and (2) actionable insights to address the gap (what to do with the gap). The thematic labels above are connected to the outcomes—BU (baseline understanding) and UU (updated understanding) connect to the first outcome while DR (design recommendations) connects to the second one. Our framework provides guidance that facilitates tractable navigation of complex sociotechnical landscapes and produces actionable insights to address the gap. We should have an idea of the most promising places to build a bridge but also where the gap is the largest. Thus, stakeholders are empowered to make informed decisions on what to address in the gap vs. not– this ability to, which a core tenet of Ackerman's call to action in addressing the gap [3].

5 APPLYING THE FRAMEWORK

We now share a third case study where we applied this framework in an applied real-world setting. To showcase the versatility of the framework, we use a new domain for the application: *cybersecurity*. The following delineation covers 1.5 years of fieldwork. After the application details, we also share a study investigating the effectiveness of the framework (did the stakeholders find it useful?).

5.1 Background: Atlas, An AI-powered Cybersecurity System

SecureCorp (pseudonym), a Fortune 500 muti-national cybersecurity technology company, invested $10 million to build Atlas (pseudonym), an AI-powered cybersecurity system to help analysts manage firewall configurations, especially "bloat" that happens when people forget to close open ports. Over time, as an organization evolves, people forget to close open ports, causing bloat

and serious security vulnerability. This has high stakes—the wrong port left open could lead to a breach, incurring huge losses for the company (in the order of millions of dollars) because the clients of SecureCorp are primarily banks. On paper, Atlas is impressive—it is purported to have commendable accuracy (92%); however, the user engagement is only at 2%. Given the significant investment in Atlas and its internal deployment, SecureCorp expected a much higher engagement. Since less than 2% of the workforce was engaged with Atlas, the question was: why?

5.2 Problem Identification & Alignment Using the Framework

Our framework facilitates a systematic *start* to problem-solving beyond helping us address it. Once we identify the problem, we need to align the problem with the framework: classify the appropriate wing (technical or social) and find the nearest matching building block. Shadowing 10 cybersecurity analysts (hereon, referred to as analysts) for 2 weeks, we identified the problem—*users were not comfortable* **acting on the AI's explanations**. Checking with our framework, it fell under the **actionability** block in the *social wing*.

As we visit the conceptual blocks in the rest of Atlas' case study, we use the labels of *baseline understanding* (BU), *updated understanding* (UU), and *design recommendations* (DR) to track the progress of our insights.

5.3 Baseline Understanding (BU)

As noted in Section 4.6, a baseline understanding provides formative information about the problem. It typically occurs during the first visit to a block, where the visits are like fact-finding missions.

Actionability (Social). To obtain a *baseline understanding*, we conducted interviews with the analysts, guided by the starter questions in our framework. We found that analysts hesitated to act on the explanations because they felt vulnerable and siloed in a vacuum from other colleagues. The nature of work had elements of collaboration that Atlas did not serve. The story of Julie (pseudonym) was infamous: Julie got fired for following the AI's mistaken recommendation to close some ports(false-positive). The AI, however, suffered no consequences.

Trust (Social). The vulnerability stemmed from lack of **trust in the AI** (top of the social wing in Figure 1). Guided by the framework, we conducted a series of Value-Sensitive Design (VSD) exercises [56, 58] to get a *baseline understanding* of what is important to them regarding trust. We found two things that informed the next steps: first, accountability was displaced unfairly in the Human-AI relationship. Discussing Julie's story, another analyst Simon (pseudonym) shared the key concern, "the AI is immune…it has nothing to lose. If the AI is wrong and I trust it, my neck is on the line. So why should I trust something that has no skin in the game?" Second, the analysts could not trust the AI because they lacked an understanding of the technical infrastructure. As Simon said, "[they] really didn't know what's going on."

These insights prompted a transition to the *technical wing* to get *baseline understandings* of its components. We used the guidelines provided for each block (discussed in Section 4) to operationally scope the technical affordances and limitations.

Data (Technical). For the **data** block, we interviewed system administrators and data scientists using our framework around data origins, the purpose of collection, policies, etc. When the development team learned about the analysts' lack of *trust* and *actionability*, they were motivated to carry out customized exercises following the Datasheets for Dataset framework [59]. We learned actionable information; for instance, that the dataset has proprietary information originally collected for compliance reasons, and that there was very little documentation on preprocessing steps, which encouraged the development team to address this deficiency.

Model (Technical). For the **model**, the data science team shared a substantial limitation— the model is proprietary and purchased from a vendor (third-party). Thus, SecureCorp's team had no way to "open" the "black-box". We also learned that this was quite normal—most organizations lack the in-house infrastructure to build systems end-to-end. Often, companies use an ensemble of technologies to provide a service. However, the charting process highlighted blind spots that profoundly impacted *trust* and *actionability*. Upon realizing that they cannot answer our starter questions, the development team had concrete steps to address—their unknowns became known.

Explanation (Technical). Next, we addressed the **explanation** from Atlas. Using the starter questions, we conducted multiple focus group discussions with the XAI engineers and analysts (end-users). We learned that Atlas generates post hoc explanations (explanation generation after model's training) [48, 95, 108, 124, 167] that are *local* (explaining a prediction vs. explaining a model) [64, 125, 168]. The explanations are fine-tuned to the company's proprietary data. The generation mechanism is model-agnostic to accommodate the lack of access to the vendor's model.

   While these insights from the social and technical blocks are informative, we cannot act on them if they are not harmonized with the organizational values, which took us back to the *social wing*.

Values (Social). To get a baseline understanding of the interplay between organizational and individual values, we used the framework guidelines to find the points of alignment and conflict through multiple interviews and focus group discussions with leadership, HR representatives, product managers, analysts, and auditors. We found *conflicting values around accountability*; for instance, the organization was set up such that the liability fell on the human's shoulder, something that hurt actionability (the core problem). Here, we successfully used the insights from previous steps to garner empathy and understanding, mitigating the conflicts between the value systems.

5.4   Updated Understanding (UU)

The baseline understanding propelled revisits to certain blocks to update our understanding, which often carried design recommendations. For clarity purposes, we share the updated understanding in this subsection. In the next subsection, we report the design recommendations in one place. We begin with revisits to *data* and *model*.

Data (Technical). Conversations with auditors (during the BU of *values*) generated two notable insights applied to the data block. First, we learned that, contrary to our previous understanding, the dataset can be updated yearly (as opposed to every 3 years). Second, the auditors highlighted compliance risks around the lack of pre-processing documentation (problem highlighted during BU of Data). We relayed these findings to the development teams through focus group discussions, which resulted in concrete design recommendations (shared in the next subsection).

Model (Technical). When the leadership learned (during the BU of *values*) about the lack of access to the model's inner workings, they re-negotiated liability terms with vendors that reduced the burden on analysts, updating our understanding of the model block and generating design recommendations (outlined later).

   Using these updated insights, we revisited the *trust* and *actionability* blocks in the *social wing*.

Trust (Social). Recall that the main reason to get a BU of the technical wing was the *mistrust* in AI due to *lack of understanding* of the technical blocks. We relayed the BU and UU of data, model, and explanation to the analysts. The increased visibility of what was possible and not possible alone recalibrated their **trust** in AI. For instance, learning about the new data update policies calibrated their expectations on the correctness of AI decisions on the latest security threats. The re-negotiated model liability with the vendor provided psychological relief.

Actionability (Social). The recalibration of trust updated **actionability** given the close relationship between them. Recall that analysts felt vulnerable and siloed. We wanted to address this problem. Guided by the framework, we utilized the lenses of Social Transparency (ST) to increase the "peripheral vision" of decision-making by providing the context of past decisions. Interviews revealed that analysts strongly resonated with the idea of reducing individual vulnerability by distributing accountability through increased visibility of each other's interactions. We used the guidelines from ST and engaged in participatory design sessions to operationalize *how* the 4W (who did what, when, and why) of ST could manifest at a design level to promote informed actionability. Aligned with the ST guidelines, we learned how the increased socio-organizational context can promote peer-to-peer learning, which adds social validation to individual decision-making. These led to design recommendations (detailed in the next subsection).

After these sessions, we reached a saturation [134] of our understanding where further investigations did not generate novel information. We proceeded to make design recommendations.

5.5 Design Recommendations (DR)

The updated understanding produced design recommendations at technological and policy levels.

On the ***technological*** side, there are recommendations on **explanation** and **data**. On the **explanation** side, analysts wanted the 4W design insights (derived from the UU of Actionability) to be incorporated alongside the AI's explanation. Next to each AI-generated explanation, they wanted the socio-organizational context—who else did what with similar AI decisions in the past and why (the 4W). There were concrete design manifestations of the 4W: for *who*, analysts wanted years of experience to be displayed to situate the authority behind their recommendation. They voted against revealing names to protect privacy. For *what*, they wanted not just what the past decision was (accept/reject) but also what the outcome was (threat nullified/not). For *when*, beyond the date, they wanted more temporally relevant information such as last dataset update. For *why*, they collaboratively agreed on a standardized template for consistency purposes. On the **Data** side, given the new understanding of update policies and pre-processing details, this information should be tagged along with other details alongside the *when* in the 4W.

In terms of ***policies***, during the BU of **value**, we productively engaged with leadership. Empowered by our findings, they launched an educational campaign (a palliative measure [3, 116]) led by the analysts (the end-users) that reformed policies and training procedures. On the **model** end, we revised employee training programs to cover the re-negotiated liability contract with the vendor so that new analysts do not suffer Julie's fate. These policy changes align with one of Ackerman's proposed methods—educational interventions—to address the sociotechnical gap [3].

5.6 Framework Outcomes

Harmonizing user requirements, organizational incentives, and technical affordances, we worked with the development team to implement the design recommendations (4W of Social Transparency and policy changes). Within one month of deployment, *the engagement was 93%, a drastic change from the initial 2%*. We continued to track progress over the next four months through interviews, surveys, and workshops. The results were promising—the improved engagement (94% on average) addressed the actionability problem. Similar to the Kuro case study, the increased "peripheral vision" afforded by the 4W of ST (who did, what, when, and why) [43] helped analysts gain a deeper understanding of the socio-organizational context, which appropriately calibrated their trust on the AI and self-confidence to act on the AI's recommendation. The increased visibility of the technical affordances enabled informed opinions about the limitations of Atlas. Most importantly, analysts reported an improvement of Atlas' explainability because it was no longer devoid of the socio-organizational context that was necessary for informed actionability.

5.7 Investigating Framework Effectiveness

While the engagement gains are important for SecureCorp, as a research team, we wanted to investigate the following: how effective was the framework? Did the stakeholders find the framework useful? To answer these questions, we conducted a follow-up study with the following procedure.

5.7.1 Methods. We hosted a two-day workshop attended by 23 participants (referred to as P1-P23 below) with diverse backgrounds (XAI engineers, Product Managers, UX Researchers, HR managers, Responsible AI managers, etc.). The workshop lasted 5 hours each day. We analyzed 10 hours of recorded footage through the lenses of thematic analysis [22]. Using an open coding scheme, we produced in-vivo codes (directly from the data) and clustered them themes. We iterated till we reached stability in our codes and respective themes, grouping them at a topic level using a combination of mind-mapping and affinity diagramming.

5.7.2 Findings. Below we present our findings on the effectiveness of the framework, related challenges, and how we may address the challenges.

The framework empowered participants to appreciate the multi-stakeholder and sociotechnical nature of the problem, expanding their understanding of explainability. Participants highlighted how the framework helped them appreciate that "everything is situated, nothing happens in a silo, and we all need to depend on each other" (P3). Reflecting on a newfound sociotechnical perspective, participants with engineering backgrounds appreciated that "it's a multiplayer game, and [they] were playing it a 1-player game" (P2). They also expressed welcomed relief at "not being alone to handle all the complex issues of AI like fairness, bias, explainability" (P23). "The framework helped [them] see that it is unfair to expect engineers to carry the entire weight of AI explainability because the problem is not just technical, it's sociotechnical" (P14). Most importantly, a sociotechnical lens afforded participants to view "XAI as something more than model transparency" (P8). This participant captures the newfound realization:

> "It's sad that I had such a very narrow view of XAI where XAI was equal to algorithmic transparency. AI systems are made of people and algorithms. I now realize explainability *of the system* is much more than explainability *of the model*... This is why [engineers] like me need to know where the data comes from, what the model cannot do... otherwise, how can I help my end user to know *when* to rely on the AI's explanation vs. not?" (P12, XAI engineer, emphasis added)

Participants emerged with a deeper understanding of the sociotechnical gap in XAI. Before participants went through the framework, they "lacked a clear understanding of the gap and how the technical and social sides mattered" (P10). The explicit sociotechnical focus with "blocks dedicated to each side made the process modular and traceable" (P14). When asked how the framework augmented their understanding of the gap, this participant summarized it well:

> "The process provided a high-resolution image of the gap as opposed to a very blurry picture when we started. We know where the gap is the widest and the closest. This knowledge is crucial for interventions. Even as someone with a PhD and who knew Ackerman's work, it never really dawned on me that we could address XAI issues like this. " (P15, UX Researcher)

Blind spots in existing organizational practices were revealed. Participants expressed that the "framework allowed [them] to clearly highlight blind spots" (P19). For example, while obtaining the BU of the Values block we found that there were value tensions around how accountability was structured and how analysts felt vulnerable, which impacted their actionability, and hurt engagement (the root problem). Most importantly, "the framework helped [stakeholders] see that [they] needed the 4W—the social transparency—and that technical transparency wasn't cutting it" (P6). Participants felt that "had [they] not gone through the process, these blind spots may not have been identified and addressed" (P21).

Table 2. Framework effectiveness workshop participant details

ID	Organizational Role	Experience	ID	Organizational Role	Experience
	Engineering and Data Science			*Product Management*	
P1	XAI Engineer	>5 ys.	P5	Product Manager	>15 ys.
P7	XAI Engineer	>10 ys.	P8	Product Manager	>8 ys.
P12	XAI Engineer	>4 ys.	P9	Product Manager	>11 ys.
P23	XAI Engineer	>3 ys.		*Organizational Governance*	
P2	Data Scientist	>10 ys.	P10	Responsible AI Manager/ AI Ethics	>3 ys.
P6	Data Scientist	>5 ys.	P11	Responsible AI Manager/ AI Ethics	>4 ys.
P14	Data Scientist	>6 ys.	P19	Responsible AI Manager/ AI Ethics	>3 ys.
	User Research		P20	Responsible AI Manager/ AI Ethics	>5 ys.
P3	UX Researcher	>5 ys.	P13	Compliance Officer	>10 ys.
P4	UX Researcher	>5 ys.	P22	Compliance Officer	>11 ys.
P15	UX Researcher	>10 ys.	P17	VP of Technology	>15 ys.
P16	UX Designer	>13 ys.	P21	SVP of Technology	>20 ys.
			P18	HR Manager	>16 ys.

The collaborative process led to improvements in inter-departmental dynamics. One of the "best effects of going through the blocks has been improvement in [participants'] relationships with other departments. In large organizations like SecureCorp, work is often siloed along verticals. Participants shared that the framework "provided peripheral vision of who else was there, how they can help, and most importantly, brought everyone under the same roof to solve the joint-problem" (P17). Even though "the start was a rough one and [non-engineering participants] didn't see eye to eye with the engineering team, [they] now have a better working relationship and mutual understanding" (P15). The process, participants expressed, "brought people from different bubbles together, which is an amazing achievement in a complex organization" (P22).

The process activated latent, pre-existing organizational information into actionable knowledge. Being able to re-use and re-purpose existing practices and guidelines was a major win for us in terms of getting buy-in and adoption. As shared in Section 3, we made a conscious effort to ensure we leverage as much of the existing infrastructure as possible; for instance, we leveraged familiar HCI techniques and AI guidelines like Model Cards that already had built-in adoption. Product Managers particularly appreciated this "sustainable way of thinking because it made the most out of what [the company] already had" (P9). These choices minimized extra burden on our participants, which they appreciated. It also activated latent information into actionable knowledge without adding significant burden to stakeholders, elaborated by this participant:

> "Thanks for not forcing us to do yet another new checklist. We're often jaded of researchers telling us to do new things without looking at how we can leverage what we've! I'm really happy that we got a lot of mileage from existing information and processes. I loved how this process helped us see old things in a new light and solve a really sticky problem." (P17, VP of Technology)

Beyond benefits of using the framework, participants highlighted two main challenges. The **first challenge** is around **"the issue of ownership and responsibility of each block"** (P1). Participants shared that they were expecting more guidance from the framework. Those from the Responsible AI and Policy teams, wondered: "Since it's a multi-stakeholder problem, it's not clear who owns which block. Who is ultimately accountable? Without clear guidance, we may end up playing hot potato or resort to 'not my problem'-type thinking" (20). *When asked how we may alleviate that hurdle*, we had generative and critically constructive suggestions. First, given problem contexts can vary, participants agreed it was "impossible to decide [ownership] before-hand, so it should not be top-down" (P22). Issues of responsibility should be discussed "when [stakeholders] have a clear idea of the problem" (P4). Second, since each block in the framework often has multiple stakeholders, ownership "need not be on only one team; it can be shared" (P12). The relative split of

accountability for the block can be decided by asking—"which group has the most to lose if things go wrong in this block?". Leadership "may need to proactively engage in cases of disagreements" (P20). Last, the conversations and should start early—"before embarking on a baseline understanding and revisited during updated understanding to see if things have changed" (P2).

Another challenge is around the **complications of obtaining collective buy-in during the early stages** of the framework. Here, teams struggled to get buy-in during the early stages and wanted the framework to better steer the process. Data scientists and developers "initially struggled to understand why non-developers had to be involved when discussing the model" (P7) *Participants offered constructive solutions to this problem of buy-in.* They suggested to add "an all-hands meeting with team leads during onboarding. To get buy-in, it's important that value propositions are clearly laid out—what's in it for each team and what we do lose if we don't put our heads together?" (P18). To give stakeholders a shared sense of ownership to facilitate the buy-in, participants advocated and appreciated our participatory methods. They "felt included; it was clear that [they] were problem-solving *with* [the researchers]" (P16). This Product Manager put it succinctly: "if you want people to join you, first you need to make them feel like their voices matter; next you make it crystal clear early on why it's a team sport and we all need to pull our weight"(P9)

6 DISCUSSION AND IMPLICATIONS

6.1 Transferring the Framework to New Domains

We constructed the framework with the ethos of customizability and versatility, striving to balance specificity with generality. This is why we developed it from distinct domains and applied it in a third new domain. The diversity of domains is an asset and speaks to the versatility of the framework's applicability. Despite their diversity, they share a vital common element: *all domains have the sociotechnical gap.* Thus, the framework can apply to domains beyond sales and cybersecurity like robotics, autonomous driving, etc. As long as the problem has an explainability component embedded in sociotechnical environments, the framework can be used to chart the sociotechnical gap. This is because the answers to the questions we ask and how we operationalize each block, and the starter questions are similar. Moreover, when transferring the framework to new domains, we should be mindful of whether the contexts are permitting. Inherent in transferability is context-sensitivity [53, 70, 144]. Real-world XAI settings have different contexts, domains, and application area—as such, the transferability of our framework to new domains is subject to context-sensitivity. Given its construction and the existence of the sociotechnical gap in all sociotechnical systems, our framework is likely to have broad transferability but not blanket generalizability.

We offer *three* "knobs" (entities) to consider as we calibrate the framework to new domains: *nature of collaboration, infrastructural complexity,* and *stakes (or consequences).* These three knobs are not the only ones that can potentially inform transfer and are not meant to be exhaustive. They are, however, informative and provide guidance during transfer.

For the *nature of collaboration,* the higher the collaborative or cooperative nature of work, the higher the impact from the social wing of the framework, and the higher the potential impact of Social Transparency in bridging the social and technical factors. The level of collaboration in our use case is relatively high—in all three cases, we had multiple users interacting with the AI system. In contrast, if the use case is geared towards 1-1 Human-AI interaction, then the social factors might not weigh as much. Beyond the number of users involved in the Human-AI interaction, we should also consider the nature of geographic co-location and organizational culture. A more distributed workforce (e.g., multi-nationals) might require higher levels of Social Transparency.

For the *infrastructural complexity,* we should consider complexities both at the technical and social (organizational) levels. Technical complexities can include complex data infrastructures, models (ensembles of deep neural networks), etc. On the social side, complex organizational structures

can impact how we understand the interplay of values, how actionability works, and what type of human factors govern trust in AI. The level of complexity can correlate with the number of "passes" or iterations we need to do over each block—the higher the level of complexity, the more time/effort we might need to invest in operationalizing each block (e.g., multiple Model Cards reporting for ensemble models), and the higher the chances of multiple iterations for updated understandings. This is why our framework provides thematic labels (baseline understanding, updated understanding, etc.) to manage progress and track challenges.

For the *stakes* (or consequences), we should be mindful of the eventual impact on human lives. These can range from distributive justice issues like the allocation of relief funds (algorithmically) to users risking job losses due to accepting a false positive AI decision. An AI system responsible for recommending music has a different level of stakes than one that recommends loan approvals. For example, in the Atlas (cybersecurity) case study, the stakes were high—a missed step could cost someone's job (case in point, Julie's story). The liability of the AI's error resided entirely on the human. This displacement of accountability created a sensitive situation than required careful stakeholder management. Not only did we have to revisit blocks (e.g., *actionability*) in the social wing to update our understanding, we also had to be careful in operationalizing the *values* block—when engaging with leadership, we had to strike a balance where no single party felt blamed for the situation. Thus, there are socio-political aspects to consider when considering stakes.

6.2 Practical Implications

Our work bears practical implications that can help researchers and practitioners (stakeholders on the technology development end such as data scientists, AI engineers, technical product managers, UX Researchers) operationalize our proposed framework.

6.2.1 Implications for Practitioners. First, thinking sociotechnically pays dividends— technology problems in organizational environments are sociotechnical. This entails that they have both social and technical components. Thus, technical interventions alone cannot solve sociotechnical problems. When looking for solutions, developers and data scientists are likely to make progress if they consider both social and technical dimensions. *Second,* the problem space requires multi-stakeholder engagements; thus, practitioners should actively seek collaboration with other teams. As we highlighted in the framework effectiveness study (Section 5.7), engineers and developers were relieved to not feel alone in handling sociotechnical complex issues in AI. The framework helped them realize that "it's a multi-player game" (P2) and that a purely technical perspective imposes an unfair burden on engineers to solve problems that are multi-stakeholder and sociotechnical. *Third,* there is value in going through the activity of charting the gap even though the sociotechnical gap may never be fully bridged. As we saw in Section 5, from highlighting organizational value tensions to unearthing non-obvious technical limitations, the framework-driven process provided actionable insights on not just understanding the sociotechnical gap in XAI but also how to address it. The process also had the positive side-effect of surfacing organizational barriers that may hinder cross-functional teamwork that, if addressed formally or informally, can improve future best practices.

6.2.2 Implications for Researchers. First, taking a participatory approach helps get stakeholder buy-in—if stakeholders feel that they are designing *with* the researchers, rather than the researchers designing *for* them, they tend to feel included and take engaged ownership of the process. We framed all problems as cooperative joint problems [1] where the power dynamics were calibrated as much as feasible. Being participatory can also complicate collective buy-ins (a challenge identified in our effectiveness study); e.g., who calls the shots for a given block. We share mitigation strategies in Section 5 such as clearly articulating value propositions before starting the process. *Second,* researchers should leverage existing information and documentation before introducing new

checklists and processes. As our framework effectiveness workshops showcased (in Section 5.7), such a sustainable approach improved (a) the chances of stakeholder buy-in and acceptance, (b) activated latent information into actionable knowledge without adding a significant burden to stakeholders, and (c) mitigated what participants called *checklist burnout* where people feel overwhelmed with long checklists and lose engagement. This is why our framework encourages using established checklist guidelines and familiar experimental methods. *Third*, where feasible, we suggest engaging in group settings that have lower barriers to participation because they tend to be generative. Relatively low-stakes group settings like "lunch-n-learns" where staff gather to listen to presentations are resourceful avenues to get initial interest in a project. Given the relevant stakeholder(s) might already be there, it reduces the barrier to getting people in a room. This is why, whenever possible, we conducted workshops. We often recruited through lunch-n-learn sessions and used virtual workshops to reduce the barrier to participation. We hope that these practical implications will help future researchers and practitioners operationalize the framework in real-world settings.

6.3 Conceptual & Theoretical Implications

By situating the sociotechnical gap in the context of XAI, our framework adds to its epistemology by *expanding the conceptual landscape* in three ways. *First,* it showcases the *utility of conceptually understanding the gap* by connecting it with practical outcomes. By providing a systematic process-driven structure to chart the gap, our framework showcases how the very act of understanding the gap can produce design interventions to address the gap. This connection motivates the practical utility in the pursuit of Ackerman's main goal in proposing the gap—the need to understand it [3]. It also speaks to the shift in focus we advocated in the introduction—one that goes from a "gap filling" philosophy to a "gap understanding" one when addressing the sociotechnical gap in XAI.

Second, our framework *connects diverse threads* of work in AI and HCI and *translates* them in the *context of XAI*, adding to its conceptual landscape. For example, on the technical wing, we integrate existing guidelines (like Datasheets for Datasets [59], Model Cards [109], XAI Question Bank [87], etc.) that previously existed in their own niches. On the social wing, we derive translational insights from XAI paradigms (like Social Transparency [43]) that incorporates socio-organizational elements. More importantly, the framework delineates the interplay between the building blocks of each wing as well as offers methodological guidelines (like Value Sensitive Design (VSD) [58] and Participatory Design (PD) [136]) on operationalizing the goals.

Third, by drawing boundaries that include the social factors rather than purely technical ones, our framework reframes the conceptualization of the relevant factors, potentially empowering stakeholders to *avoid certain techno-centric traps* such as Solutionism [114] and Formalism [63]. Selbst et al.[137] highlight falling into these traps (by excluding the social context) can cause AI interventions to exacerbate societal inequities and unfairness. By "refocusing of design in terms of process rather than solutions" [137], our framework can help researchers and practitioners avoid these pitfalls. It encourages a sociotechnically-informed "heterogeneous engineering approach" [137], one that critically reflects on *what* social factors to use *why*, and *when* to implement design interventions [63, 137]. For example, in the Atlas case study, the framework not only helped us devise design recommendations around the 4W (who, what, when, why) of Social Transparency [43], but it also highlighted the areas (actionability and trust) in which it is most impactful.

The conceptual implications are connected to our *design philosophy*. As we shared in the introduction, our framework is generative, not normative. Its process-driven nature is meant to empower stakeholders (e.g., researchers, designers, developers) to systematically chart the sociotechnical gap and generate insights to address the gap. We do not anticipate there exists a one-size-fits-all paradigm that could chart the gap universally for all contexts. There could be many paths that lead

to the same destination. As such, we do not impose restrictions on the exact path to mapping the gap. Instead, we focus on operationalizing the building blocks systematically where the generative insights guide the path-finding process. For example, a startup company might take a different sequence of actions in charting the gap compared to a large multinational organization. While they might have different types of gaps, the gap still exists, which is where our generative framework can help in both cases. Incorporating the lenses of Social Construction of Technology developed by Pinch and Bijker [118], we encourage *interpretive flexibility* of the process from different *relevant social groups*, each with its own goals and visions of the technology. This design philosophy not only permeates this framework but also has conceptual implications on future work based on it.

6.4 Limitations & Future Work

We have taken a formative step on how to chart the gap and address it by deriving an analytical framework using two case studies and applying it on a new one. Given this first step, the insights from our work should be scoped accordingly. We need to do future work on applying the framework across more domains and in different settings. In our case studies, both the development and end-users were part of the same organization. Future work could explore contexts where the development and end-user teams are from different organizations (e.g., service providers and customers).

For future iterations, we are inspired by Agre's design philosophy of Critical Technical Practice [5]. Aligned with Agre's notion, "at least for the foreseeable future, [we] will require a split identity – one foot planted in the craft work of design and the other foot planted in the reflexive work of critique." [6]. As such, our exploration of the sociotechnical gap in XAI, at least for the foreseeable future, is a work-in-progress. We have "planted one foot" in the work of design by offering a framework that helps map the gap and address it. Now, we seek to learn from and with the broader HCI and XAI communities as we "plant the other foot" in the self-reflective realm of critique.

7 CONCLUSION

Explainable AI (XAI) systems are inherently sociotechnical; thus, they are subject to the sociotechnical gap. This paper targeted the problem of charting the sociotechnical gap between the technical affordances and the social needs in XAI systems. We used two case studies in two different domains (sales and mental health) to empirically derive an analytic framework to facilitate charting the sociotechnical gap in XAI. This framework consisted of three building blocks in each of the technical and social wings. The technical wing included data, model, and explanations components, and the social wing included trust, actionability, and values components. For each of these building blocks, we provided a set of starter questions that would help address the sociotechnical gap in XAI systems. Further, we delineated how to address the gap through the lenses of palliatives and first-order approximations [3]. Next, we applied the proposed framework to a third case study in a new domain (cybersecurity) to showcase its affordances, and how it helped to not only understand the problem but also to generate actionable insights to adequately address the problem and improve the system's explainability. Our work provided design recommendations and potential challenges, risks, and tensions in mapping out and addressing the sociotechnical gap in XAI.

REFERENCES

[1] Leena Aarikka-Stenroos and Elina Jaakkola. 2012. Value co-creation in knowledge intensive business services: A dyadic perspective on the joint problem solving process. *Industrial marketing management* 41, 1 (2012), 15–26.

[2] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 582.

[3] Mark S Ackerman. 2000. The intellectual challenge of CSCW: the gap between social requirements and technical feasibility. *Human–Computer Interaction* (2000).

[4] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160.

[5] 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).

[6] Philip E Agre. 1997. *Computation and human experience.* Cambridge University Press.

[7] 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*.

[8] Steven Alter. 2010. Design spaces for sociotechnical systems. (2010).

[9] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.

[10] McKane Andrus, Sarah Dean, Thomas Krendl Gilbert, Nathan Lambert, and Tom Zick. 2020. AI development for the public interest: From abstraction traps to sociotechnical risks. In *2020 IEEE International Symposium on Technology and Society (ISTAS)*. IEEE.

[11] Matthew Arnold, Rachel KE Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilović, Ravi Nair, K Natesan Ramamurthy, Alexandra Olteanu, David Piorkowski, et al. 2019. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development* 63, 4/5 (2019), 6–1.

[12] 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).

[13] 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* (2019).

[14] Maryam Ashoori and Justin D Weisz. 2019. In AI we trust? Factors that influence trustworthiness of AI-infused decision-making processes. *arXiv preprint arXiv:1912.02675* (2019).

[15] American Psychiatric Association et al. 2015. *The American Psychiatric Association practice guidelines for the psychiatric evaluation of adults.* American Psychiatric Pub.

[16] Aaron Balick. 2014. Technology, social media, and psychotherapy: Getting with the programme. *Contemporary Psychotherapy* 6, 2 (2014).

[17] Natalya N Bazarova, Yoon Hyung Choi, Victoria Schwanda Sosik, Dan Cosley, and Janis Whitlock. 2015. Social sharing of emotions on Facebook: Channel differences, satisfaction, and replies. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*. 154–164.

[18] GM Beal and JM Bohlen. 1957. The diffusion process (Special Report Nᵒ 18, Agricultural Experiment Station). *Iowa State College* (1957).

[19] Martin Bella and Bruce Hanington. 2012. Universal methods of design. *Beverly, MA: Rockport Publishers* (2012), 204.

[20] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. 2018. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943* (2018).

[21] Hugh R Beyer and Karen Holtzblatt. 1996. Contextual techniques starter kit. *interactions* 3, 6 (1996), 44–50.

[22] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[23] Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*.

[24] 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.

[25] Daniel Buschek, Lukas Mecke, Florian Lehmann, and Hai Dang. 2021. Nine Potential Pitfalls when Designing Human-AI Co-Creative Systems. *arXiv preprint arXiv:2104.00358* (2021).

[26] 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*.

[27] 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.

[28] Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine* (2020).

[29] Zhengping Che, Sanjay Purushotham, Robinder Khemani, and Yan Liu. 2016. Interpretable deep models for ICU outcome prediction. In *AMIA Annual Symposium Proceedings*, Vol. 2016.

[30] 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*.

[31] EunJeong Cheon and Norman Makoto Su. 2016. Integrating roboticist values into a Value Sensitive Design framework for humanoid robots. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 375–382.

[32] EunJeong Cheon and Norman Makoto Su. 2018. Futuristic autobiographies: Weaving participant narratives to elicit values around robots. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*.

[33] Clifford Christians. 1989. A theory of normative technology. In *Technological Transformation*. Springer, 123–139.

[34] Massimiliano Sassoli de Bianchi. 2013. The observer effect. *Foundations of science* 18, 2 (2013), 213–243.

[35] Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*.

[36] Munmun De Choudhury, Scott Counts, Eric J. Horvitz, and Aaron Hoff. 2014. Characterizing and predicting postpartum depression from shared facebook data. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing - CSCW '14*. ACM Press, Baltimore, Maryland, USA, 626–638.

[37] Munmun De Choudhury, Min Kyung Lee, Haiyi Zhu, and David A Shamma. 2020. Introduction to this special issue on unifying human computer interaction and artificial intelligence. *Human–Computer Interaction* (2020).

[38] 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*.

[39] 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*.

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

[41] Paul Dourish and Genevieve Bell. 2011. *Divining a digital future: Mess and mythology in ubiquitous computing*. Mit Press.

[42] 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

[43] 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*.

[44] 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).

[45] Upol Ehsan and Mark O Riedl. 2020. Human-centered explainable ai: Towards a reflective sociotechnical approach. In *International Conference on Human-Computer Interaction*. Springer, 449–466.

[46] Upol Ehsan and Mark O Riedl. 2021. Explainability pitfalls: Beyond dark patterns in explainable AI. *arXiv preprint arXiv:2109.12480* (2021).

[47] Upol Ehsan and Mark O Riedl. 2022. Social Construction of XAI: Do We Need One Definition to Rule Them All? *arXiv preprint arXiv:2211.06499* (2022).

[48] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark Riedl. 2019. Automated Rationale Generation: A Technique for Explainable AI and its Effects on Human Perceptions. In *Proceedings of the International Conference on Intelligence User Interfaces*.

[49] 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*.

[50] Malin Eiband, Daniel Buschek, and Heinrich Hussmann. 2021. How to support users in understanding intelligent
      systems? Structuring the discussion. In *26th International Conference on Intelligent User Interfaces*. 120–132.
[51] Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. 2018.
      Bringing transparency design into practice. In *23rd international conference on intelligent user interfaces*. 211–223.
[52] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin
      Hamilton, and Christian Sandvig. 2015. "I always assumed that I wasn't really that close to [her]" Reasoning about
      Invisible Algorithms in News Feeds. In *Proceedings of CHI conference on human factors in computing systems*.
[53] Deborah Finfgeld-Connett. 2010. Generalizability and transferability of meta-synthesis research findings. *Journal of
      advanced nursing* 66, 2 (2010), 246–254.
[54] Carl E. Fisher and Paul S. Appelbaum. 2017. Beyond Googling. *Harvard Review of Psychiatry* (2017), 1. https:
      //doi.org/10.1097/hrp.0000000000000145
[55] Andrea Forte and Cliff Lampe. 2013. Defining, understanding, and supporting open collaboration: Lessons from the
      literature. *American behavioral scientist* 57, 5 (2013), 535–547.
[56] Batya Friedman. 1996. Value-sensitive design. *interactions* 3, 6 (1996), 16–23.
[57] Batya Friedman and David Hendry. 2012. The envisioning cards: a toolkit for catalyzing humanistic and technical
      imaginations. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1145–1148.
[58] Batya Friedman, Peter Kahn, and Alan Borning. 2002. Value sensitive design: Theory and methods. *University of
      Washington technical report* 2-12 (2002).
[59] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III,
      and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010* (2018).
[60] Katy Ilonka Gero, Zahra Ashktorab, Casey Dugan, Qian Pan, James Johnson, Werner Geyer, Maria Ruiz, Sarah Miller,
      David R Millen, Murray Campbell, et al. 2020. Mental Models of AI Agents in a Cooperative Game Setting. In
      Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 1–12.
[61] Bhavya Ghai, Q Vera Liao, Yunfeng Zhang, Rachel Bellamy, and Klaus Mueller. 2021. Explainable Active Learning
      (XAL) Toward AI Explanations as Interfaces for Machine Teachers. *Proceedings of the ACM on Human-Computer
      Interaction* 4, CSCW3 (2021).
[62] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining
      explanations: An approach to evaluating interpretability of machine learning. *arXiv preprint arXiv:1806.00069* (2018).
[63] Ben Green and Salomé Viljoen. 2020. Algorithmic realism: expanding the boundaries of algorithmic thought. In
      Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 19–31.
[64] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. 2018.
      Local rule-based explanations of black box decision systems. *arXiv preprint arXiv:1805.10820* (2018).
[65] 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.
[66] David Gunning. 2017. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd
      Web* 2 (2017), 2.
[67] Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2017. Detecting
      depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences* (2017).
[68] Karen Hao. 2019. AI is sending people to jail – and getting it wrong. *MIT Technology Review* (21 January 2019).
      Retrieved 26-August-2019 from https://www.technologyreview.com/s/612775/algorithms-criminal-justice-ai/
[69] 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.
[70] Gillian R Hayes. 2011. The relationship of action research to human-computer interaction. *ACM Transactions on
      Computer-Human Interaction (TOCHI)* 18, 3 (2011), 1–20.
[71] Michael Hind. 2019. Explaining explainable AI. *XRDS: Crossroads, The ACM Magazine for Students* 25, 3 (2019), 16–19.
[72] Michael Hind, Sameep Mehta, Aleksandra Mojsilovic, Ravi Nair, Karthikeyan Natesan Ramamurthy, Alexandra
      Olteanu, and Kush R Varshney. 2018. Increasing trust in AI services through supplier's declarations of conformity.
      arXiv preprint arXiv:1808.07261 18 (2018), 2813–2869.
[73] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that
      influence trust. *Human factors* 57, 3 (2015), 407–434.
[74] 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*.
[75] Sarah Holland, Ahmed Hosny, and Sarah Newman. 2020. The dataset nutrition label. *Data Protection and Privacy:
      Data Protection and Democracy (2020)* 1 (2020).
[76] Andreas Holzinger, Chris Biemann, Constantinos S Pattichis, and Douglas B Kell. 2017. What do we need to build
      explainable AI systems for the medical domain? *arXiv preprint arXiv:1712.09923* (2017).

[77] Andrew JI Jones, Alexander Artikis, and Jeremy Pitt. 2013. The design of intelligent socio-technical systems. *Artificial Intelligence Review* 39, 1 (2013), 5–20.

[78] Gajendra Jung Katuwal and Robert Chen. 2016. Machine learning model interpretability for precision medicine. *arXiv preprint arXiv:1610.09045* (2016).

[79] 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*.

[80] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. Human Decisions and Machine Predictions. *The Quarterly Journal of Economics* 133, 1 (2017), 237–293. https://doi.org/10.1093/qje/qjx032

[81] Bran Knowles and John T Richards. 2021. The Sanction of Authority: Promoting Public Trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 262–271.

[82] Sivam Krish. 2011. A practical generative design method. *Computer-Aided Design* 43, 1 (2011), 88–100.

[83] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on visual languages and human centric computing*. IEEE.

[84] Vivian Lai, Han Liu, and Chenhao Tan. 2020. " Why is' Chicago'deceptive?" Towards Building Model-Driven Tutorials for Humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*.

[85] Ellen E Lee, John Torous, Munmun De Choudhury, Colin A Depp, Sarah A Graham, Ho-Cheol Kim, Martin P Paulus, John H Krystal, and Dilip V Jeste. 2021. Artificial Intelligence for Mental Healthcare: Clinical Applications, Barriers, Facilitators, and Artificial Wisdom. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* (2021).

[86] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, et al. 2019. WeBuildAI: Participatory framework for algorithmic governance. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–35.

[87] 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*.

[88] Q. Vera Liao, Milena Pribić, Jaesik Han, Sarah Miller, and Daby Sow. 2021. Question-Driven Design Process for Explainable AI User Experiences. *arXiv:2104.03483 [cs]* (Sept. 2021). http://arxiv.org/abs/2104.03483 arXiv: 2104.03483.

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

[90] 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.

[91] Brian Y Lim and Anind K Dey. 2010. Toolkit to support intelligibility in context-aware applications. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*. 13–22.

[92] Brian Y Lim and Anind K Dey. 2011. Investigating intelligibility for uncertain context-aware applications. In *Proceedings of the 13th international conference on Ubiquitous computing*. 415–424.

[93] 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 (CHI '09)*. ACM, New York, NY, USA, 2119–2128. https://doi.org/10.1145/1518701.1519023

[94] 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*.

[95] Zachary C Lipton. 2018. The mythos of model interpretability. *Queue* 16, 3 (2018), 31–57.

[96] Tyler J. Loftus, Patrick J. Tighe, Amanda C. Filiberto, Philip A. Efron, Scott C. Brakenridge, Alicia M. Mohr, Parisa Rashidi, Jr Upchurch, Gilbert R., and Azra Bihorac. 2020. Artificial Intelligence and Surgical Decision-making. *JAMA Surgery* (2020).

[97] Tania Lombrozo. 2011. The instrumental value of explanations. *Philosophy Compass* 6, 8 (2011).

[98] Tania Lombrozo. 2012. Explanation and abductive inference. *Oxford handbook of thinking and reasoning* (2012).

[99] Duri Long and Brian Magerko. 2020. What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–16.

[100] Ewa Luger and Abigail Sellen. 2016. " Like Having a Really Bad PA" The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 5286–5297.

[101] Henrietta Lyons, Eduardo Velloso, and Tim Miller. 2021. Conceptualising contestability: Perspectives on contesting algorithmic decisions. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–25.

[102] Donald MacKenzie. 2018. Material Signals: A Historical Sociology of High-Frequency Trading. *Amer. J. Sociology* 123, 6 (2018), 1635–1683. https://doi.org/10.1086/697318

[103] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[104] Jonathan Magnusson. [n.d.]. Improving Dark Pattern Literacy of End Users. ([n. d.]).

[105]   Erin E Makarius, Debmalya Mukherjee, Joseph D Fox, and Alexa K Fox. 2020. Rising with the machines: A so-
        ciotechnical framework for bringing artificial intelligence into the organization. *Journal of Business Research* 120
        (2020).
[106]   Masike Malatji, Sune Von Solms, and Annlizé Marnewick. 2019. Socio-technical systems cybersecurity framework.
        Information & Computer Security (2019).
[107]   Jon McCormack, Alan Dorin, and Troy Innocent. 2004. Generative Design: A Paradigm for Design Research. (2004).
[108]   Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267
        (2019), 1–38.
[109]   Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer,
        Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on
        fairness, accountability, and transparency*. 220–229.
[110]   Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining explanations in AI. In *Proceedings of the
        conference on fairness, accountability, and transparency*. 279–288.
[111]   Shakir Mohamed, Marie-Therese Png, and William Isaac. 2020. Decolonial AI: Decolonial Theory as Sociotechnical
        Foresight in Artificial Intelligence. *Philosophy & Technology* (2020), 1–26.
[112]   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.
[113]   Geoffrey A Moore and Regis McKenna. 1999. Crossing the chasm. (1999).
[114]   Evgeny Morozov. 2013. *To save everything, click here: The folly of technological solutionism.* Public Affairs.
[115]   Michael Muller and Q Vera Liao. [n.d.]. Exploring AI Ethics and Values through Participatory Design Fictions. ([n. d.]).
[116]   Sean A Munson, Hasan Cavusoglu, Larry Frisch, and Sidney Fels. 2013. Sociotechnical challenges and progress in
        using social media for health. *Journal of medical Internet research* 15, 10 (2013), e226.
[117]   John Murawski. 2019. Mortgage Providers Look to AI to Process Home Loans Faster. *Wall Street Journal* (18 March
        2019). Retrieved 16-September-2020 from https://www.wsj.com/articles/mortgage-providers-look-to-ai-to-process-
        home-loans-faster-11552899212
[118]   Trevor J Pinch and Wiebe E Bijker. 1984. The social construction of facts and artefacts: Or how the sociology of
        science and the sociology of technology might benefit each other. *Social studies of science* 14, 3 (1984), 399–441.
[119]   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).
[120]   Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. 2022. Data Cards: Purposeful and Transparent Dataset
        Documentation for Responsible AI. *arXiv preprint arXiv:2204.01075* (2022).
[121]   Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as Mechanisms for Supporting Algorithmic
        Transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 103.
[122]   Emilee Rader and Rebecca Gray. 2015. Understanding User Beliefs About Algorithmic Curation in the Facebook
        News Feed. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*.
        Association for Computing Machinery, New York, NY, USA, 173–182. https://doi.org/10.1145/2702123.2702174
[123]   Gabriëlle Ras, Marcel van Gerven, and Pim Haselager. 2018. Explanation methods in deep learning: Users, values,
        concerns and challenges. In *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Springer.
[124]   Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should I trust you?: Explaining the predictions
        of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data
        mining*. ACM, 1135–1144.
[125]   Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations.
        In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.
[126]   Cynthia Rudin, Caroline Wang, and Beau Coker. 2020. The Age of Secrecy and Unfairness in Recidi-
        vism Prediction. *Harvard Data Science Review* 2, 1 (31 3 2020). https://doi.org/10.1162/99608f92.6ed64b30
        https://hdsr.mitpress.mit.edu/pub/7z10o269.
[127]   Selma Šabanović. 2010. Robots in society, society in robots. *International Journal of Social Robotics* 2, 4 (2010), 439–450.
[128]   Koustuv Saha, Pranshu Gupta, Gloria Mark, Emre Kıcıman, and Munmun De Choudhury. 2023. Observer Effect in
        Social Media Use. (2023).
[129]   Koustuv Saha, Jordyn Seybolt, Stephen M Mattingly, Talayeh Aledavood, Chaitanya Konjeti, Gonzalo J Martinez, Ted
        Grover, Gloria Mark, and Munmun De Choudhury. 2021. What life events are disclosed on social media, how, when,
        and by whom?. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–22.
[130]   Koustuv Saha, Benjamin Sugar, John Torous, Bruno Abrahao, Emre Kıcıman, and Munmun De Choudhury. 2019. A
        social media study on the effects of psychiatric medication use. In *Proceedings of the International AAAI Conference on
        Web and Social Media*, Vol. 13. 440–451.
[131]   Koustuv Saha, Asra Yousuf, Ryan L Boyd, James W Pennebaker, and Munmun De Choudhury. 2022. Social media
        discussions predict mental health consultations on college campuses. *Scientific reports* (2022).

[132] Javier Sánchez-Monedero, Lina Dencik, and Lilian Edwards. 2020. What does it mean to'solve'the problem of discrimination in hiring? social, technical and legal perspectives from the UK on automated hiring systems. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 458–468.

[133] Lindsay Sanneman and Julie A Shah. 2020. A Situation Awareness-Based Framework for Design and Evaluation of Explainable AI. In *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*.

[134] Benjamin Saunders, Julius Sim, Tom Kingstone, Shula Baker, Jackie Waterfield, Bernadette Bartlam, Heather Burroughs, and Clare Jinks. 2018. Saturation in qualitative research: exploring its conceptualization and operationalization. *Quality & quantity* 52, 4 (2018), 1893–1907.

[135] 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.

[136] Douglas Schuler and Aki Namioka. 1993. *Participatory design: Principles and practices*. CRC Press.

[137] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*.

[138] 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.

[139] Ben Shneiderman. 2020. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction* 36, 6 (2020), 495–504.

[140] Keng Siau and Weiyu Wang. 2018. Building trust in artificial intelligence, machine learning, and robotics. *Cutter business technology journal* 31, 2 (2018), 47–53.

[141] Supriya Singh, Anuja Cabraal, Catherine Demosthenous, Gunela Astbrink, and Michele Furlong. 2007. Password sharing: implications for security design based on social practice. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 895–904.

[142] Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Daniel S Weld, and Leah Findlater. 2020. No explainability without accountability: An empirical study of explanations and feedback in interactive ml. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[143] Kacper Sokol and Peter Flach. 2020. Explainability fact sheets: a framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*.

[144] Ernest T Stringer. 2007. Action research third edition. (2007).

[145] 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*.

[146] 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.

[147] Harini Suresh and John V Guttag. 2019. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002* (2019).

[148] Richard Tomsett, Dave Braines, Dan Harborne, Alun Preece, and Supriyo Chakraborty. 2018. Interpretable to whom? A role-based model for analyzing interpretable machine learning systems. *arXiv preprint arXiv:1806.07552* (2018).

[149] Jennifer Wortman Vaughan and Hanna Wallach. [n.d.]. 1 A Human-Centered Agenda for Intelligible Machine Learning. ([n. d.]).

[150] Viswanath Venkatesh and Fred D Davis. 2000. A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management science* 46, 2 (2000), 186–204.

[151] Viswanath Venkatesh, Michael G Morris, Gordon B Davis, and Fred D Davis. 2003. User acceptance of information technology: Toward a unified view. *MIS quarterly* (2003), 425–478.

[152] Daniel Vigo, Graham Thornicroft, and Rifat Atun. 2016. Estimating the true global burden of mental illness. *The Lancet Psychiatry* 3, 2 (2016), 171–178.

[153] Guy H Walker, Neville A Stanton, Paul M Salmon, and Daniel P Jenkins. 2008. A review of sociotechnical systems theory: a classic concept for new command and control paradigms. *Theoretical issues in ergonomics science* 9, 6 (2008).

[154] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems*.

[155] Qiaosi Wang, Koustuv Saha, Eric Gregori, David Joyner, and Ashok Goel. 2021. Towards Mutual Theory of Mind in Human-AI Interaction: How Language Reflects What Students Perceive About a Virtual Teaching Assistant. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[156] Daniel A Wilkenfeld and Tania Lombrozo. 2015. Inference to the best explanation (IBE) versus explaining for the best inference (EBI). *Science & Education* 24, 9-10 (2015), 1059–1077.

[157] Christine Wolf and Jeanette Blomberg. 2019. Evaluating the promise of human-algorithm collaborations in everyday work practices. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23.

[158] Christine T Wolf. 2019. Explainability scenarios: towards scenario-based XAI design. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 252–257.

[159] Yao Xie, Melody Chen, David Kao, Ge Gao, and Xiang'Anthony' Chen. 2020. CheXplain: Enabling Physicians to Explore and Understand Data-Driven, AI-Enabled Medical Imaging Analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*.

[160] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L Arendt. 2020. How do visual explanations foster end users' appropriate trust in machine learning?. In *Proc. IUI*.

[161] Qian Yang. 2019. *Profiling Artificial Intelligence as a Material for User Experience Design*. Ph.D. Dissertation.

[162] 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*.

[163] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. Unremarkable ai: Fitting intelligent decision support into critical, clinical decision-making processes. In *Proc. CHI*.

[164] Qian Yang, John Zimmerman, Aaron Steinfeld, Lisa Carey, and James F Antaki. 2016. Investigating the heart pump implant decision process: opportunities for decision support tools to help. In *Proc. CHI*.

[165] Dong Whi Yoo, Michael L Birnbaum, Anna R Van Meter, Asra F Ali, Elizabeth Arenare, Gregory D Abowd, and Munmun De Choudhury. 2020. Designing a clinician-facing tool for using insights from patients' social media activity: Iterative co-design approach. *JMIR Mental Health* (2020).

[166] Dong Whi Yoo, Sindhu Kiranmai Ernala, Bahador Saket, Domino Weir, Elizabeth Arenare, Asra F Ali, Anna R Van Meter, Michael L Birnbaum, Gregory D Abowd, and Munmun De Choudhury. 2021. Clinician perspectives on using computational mental health insights from patients' social media activities: design and qualitative evaluation of a prototype. *JMIR Mental Health* 8, 11 (2021), e25455.

[167] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. 2015. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579* (2015).

[168] Quanshi Zhang, Yu Yang, Haotian Ma, and Ying Nian Wu. 2019. Interpreting cnns via decision trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6261–6270.

[169] 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*. ACM.

[170] Haiyi Zhu, Bowen Yu, Aaron Halfaker, and Loren Terveen. 2018. Value-sensitive algorithm design: Method, case study, and lessons. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–23.