Expanding Explainability: Towards Social Transparency in AI Systems

Upol Ehsan
Georgia Institute of Technology
Atlanta, GA, USA
ehsanu@gatech.edu

Q. Vera Liao
IBM Research AI
Yorktown Heights, NY, USA
vera.liao@ibm.com

Michael Muller
IBM Research AI
Yorktown Heights, NY, USA
michael_muller@us.ibm.com

Mark O. Riedl
Georgia Institute of Technology
Atlanta, GA, USA
riedl@cc.gatech.edu

Justin D. Weisz
IBM Research AI
Yorktown Heights, NY, USA
jweisz@us.ibm.com

ABSTRACT

As AI-powered systems increasingly mediate consequential decision-making, their explainability is critical for end-users to take informed and accountable actions. Explanations in human-human interactions are socially-situated. AI systems are often socio-organizationally embedded. However, Explainable AI (XAI) approaches have been predominantly algorithm-centered. We take a developmental step towards socially-situated XAI by introducing and exploring Social Transparency (ST), a sociotechnically informed perspective that incorporates the socio-organizational context into explaining AI-mediated decision-making. To explore ST conceptually, we conducted interviews with 29 AI users and practitioners grounded in a speculative design scenario. We suggested constitutive design elements of ST and developed a conceptual framework to unpack ST's effect and implications at the technical, decision-making, and organizational level. The framework showcases how ST can potentially calibrate trust in AI, improve decision-making, facilitate organizational collective actions, and cultivate holistic explainability. Our work contributes to the discourse of Human-Centered XAI by expanding the design space of XAI.

CCS CONCEPTS

• **Human-centered computing** → *Scenario-based design*; **Empirical studies in HCI**; **HCI theory, concepts and models**; *Collaborative and social computing theory, concepts and paradigms*; • **Computing methodologies** → *Artificial intelligence.*

KEYWORDS

Explainable AI, social transparency, human-AI interaction, explanations, Artificial Intelligence, sociotechnical, socio-organizational context

ACM Reference Format:
Upol Ehsan, Q. Vera Liao, Michael Muller, Mark O. Riedl, and Justin D. Weisz. 2021. Expanding Explainability: Towards Social Transparency in AI Systems. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3411764.3445188

1 INTRODUCTION

Explanations matter. In human-human interactions, they provide necessary delineations of reasoning and justification for one's thoughts and actions, and a primary vehicle to transfer knowledge from one person to another [65]. Explanations play a central role in sense-making, decision-making, coordination, and many other aspects of our personal and social lives [41]. They are becoming increasingly important in human-AI interactions as well. As AI systems are rapidly being employed in high stakes decision-making scenarios in industries such as healthcare [63], finance [76], college admissions [79], hiring [19], and criminal justice [37], the need for explainability becomes paramount. Explainability is not only sought by users and other stakeholders to understand and develop appropriate trust of AI systems, but also to support discovery of new knowledge and make informed decisions [58]. To respond to this emerging need for explainability, there has been commendable progress in the field of Explainable AI (XAI), especially around algorithmic approaches to generate representations of how a machine learning (ML) model operates or makes decisions.

Despite the recent growth spurt in the field of XAI, studies examining how people actually interact with AI explanations have found popular XAI techniques to be ineffective [6, 80, 111], potentially risky [50, 95], and underused in real-world contexts [58]. The field has been critiqued for its techno-centric view, where "inmates [are running] the asylum" [70], based on the impression that XAI researchers often develop explanations based on their own intuition rather than the situated needs of their intended audience. Currently, the dominant algorithm-centered XAI approaches make up for only a small fragment of the landscape of explanations as studied in the Social Sciences [65, 70, 71, 101] and exhibit significant gaps from how explanations are sought and produced by people. Certain techno-centric pitfalls that are deeply embedded in AI and Computer Science, such as Solutionism (always seeking technical solutions) and Formalism (seeking abstract, mathematical solutions) [32, 87], are likely to further widen these gaps.

CHI '21, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05…$15.00
https://doi.org/10.1145/3411764.3445188

One way to address the gaps would be to critically reflect on the status quo. Here, the lenses of Agre's Critical Technical Practice (CTP) [4, 5] can help. CTP encourages us to question the core epistemic and methodological assumptions in XAI, critically reflect on them to overcome impasses, and generate new questions and hypotheses. By bringing the unconscious aspects of experience to our conscious awareness, critical reflection makes them actionable [24, 25, 88]. Put differently, a CTP-inspired reflective perspective on XAI [26] will encourage us to ask: by continuing the dominant algorithm-centered paradigm in XAI, what perspectives are we missing? How might we incorporate the marginalized perspectives to embody alternative technology? In this case, a dominant XAI approach can be construed as algorithm-centered that privileges technical transparency and circumscribes the epistemic space of explainable AI around model transparency. An algorithm-centered approach can be effective if explanations and AI systems existed in a vacuum. However, it is not the case that explanations and AI systems are devoid of situated context.

On one hand, explanations (as a construct) are socially situated [64, 65, 70, 105]. Explanation is first and foremost a shared meaning-making process that occurs between an explainer and an explainee. This process is dynamic to the goals and changing beliefs of both parties [20, 38, 39, 45]. For our purposes in this paper, we adopt the broad definition that an explanation is an answer to a *why*-question [20, 57, 70].

On the other hand, implicit in AI systems are *human-AI assemblages*. Most consequential AI systems are deeply embedded in socio-organizational tapestries in which groups of humans interact with it, going beyond a 1-1 human-AI interaction paradigm. Given this understanding, we might ask: if both AI systems and explanations are socially-situated, then why are we not requiring incorporation of the social aspects when we conceptualize explainability in AI systems? How can one form a holistic understanding of an AI system and make informed decisions if one only focuses on the technical half of a sociotechnical system?

We illustrate the shortcomings of a solely technical view of explainability in the following scenario, which is inspired by incidents described by informants in our study.

> *You work for a leading cloud software company, responsible for determining product pricing in various markets. Your institution built a new AI-powered tool that provides pricing recommendations based on a wide variety of factors. This tool has been extensively evaluated to assist you on pricing decisions. One day, you are tasked with creating a bid to be the cloud provider for a major financial institution. The AI-powered tool gives you a recommended price. You might think, why should I trust the AI's recommendation? You examine a variety of technical explanations the system provides: visualizations of the model's decision-making process and descriptions of how the algorithm reached this specific recommendation. Confident at the soundness of the model's recommendation, you create the bid and submit it to the client. You are disheartened to learn that the client rejected your bid and instead accepted the bid from a competitor.*

Given a highly-accurate machine learning model, along with a full complement of technical explanations, why should the seller's pricing decision not have been successful? It is because the answer to the *why*-question is not limited to the machine explaining itself. It is also in the situational and socio-organizational context, which one can learn from how price recommendations were handled by other sellers. What other factors went into those decisions? Were there regulatory or client-specific (e.g., internal budgetary constraints) issues that were beyond the scope of the model? Did something drastic happen in the operating environment (e.g., a global pandemic) that necessitated a different strategy? In other words, situational context matters and it is with this context the "why" questions could be answered effectively and completely.

At a first glance, it may seem that socio-organizational context has nothing to do with explaining an AI system. Therein lies the issue — where we draw the boundary of our epistemic canvas for XAI matters. If the boundary is traced along the bounds of an algorithm, we risk excluding the human and social factors that significantly impact the way people make sense of a system. Sense-making is not just about opening the closed box of AI, but also about who is around the box, and the sociotechnical factors that govern the use of the AI system and the decision. Thus the "ability" in explainability does not lie exclusively in the guts of the AI system [26]. For the XAI field as a whole, if we restrict our epistemic lenses to solely focus on algorithms, we run the risk of perpetuating the aforementioned gaps, marginalizing the human and sociotechnical factors in XAI design. The lack of incorporation of the socio-organizational context is an epistemic blind spot in XAI. By identifying and critically reflecting on this epistemic blind spot, we can begin to recognize the poverty of algorithm-centered approaches.

In this paper, we address this blind spot and expand the conceptual lens of XAI by reframing explainability beyond algorithmic transparency, focusing our attention to the human and socio-organizational factors around explainability of AI systems. Building upon relevant concepts that promote transparency of social information in human-human interactions, we introduce and explore Social Transparency (ST) in AI systems. Using a scenario-based design, we create a speculative instance of AI-mediated decision-support system and use it to conduct a formative study with 29 AI users and practitioners. Our study explores whether and how proposed constitutive design elements address the epistemic blind spot of XAI – incorporating socio-organizational contexts into explainability. We also investigate whether and how ST can facilitate AI-mediated decision-making and other user goals. This paper is not a full treatise of how to achieve socially-situated XAI; rather a first step toward that goal by operationalizing the concept in a set of design elements and considering its implications for human-AI interaction. In summary, our contributions are fourfold:

- We highlight an epistemic blind spot in XAI – a lack of incorporation of socio-organizational contexts that impact the explainability of AI-mediated decisions – by using a CTP-inspired reflective approach to XAI.
- We explore the concept of Social Transparency (ST) in AI systems and develop a scenario-based speculative design that embodies ST, including four categories of design features

that reflect *What*, *Why*, *Who*, and *When* information of past user interactions with AI systems.

- We conduct a formative study and empirically derive a conceptual framework, highlighting three levels of context around AI-mediated decisions that are made visible by ST and their potential effects: technological (AI), decision, and organizational contexts.
- We share design insights and potential challenges, risks, and tensions of introducing ST into AI systems.

2 RELATED WORK

We begin with a in-depth review of related work in XAI field, further highlighting the danger of the epistemic blind spot. We then discuss a shift in broader AI related work towards sociotechnical perspectives. Lastly, we review work that pushed towards transparency of socio-organizational contexts in human-human interactions.

2.1 Explainable AI (XAI)

Although there is no established consensus on the complete set of factors that makes an AI system explainable, XAI work commonly shares the goal of making an AI system's functioning or decisions *easy to understand* by people [7, 14, 27, 31, 34, 62, 70, 82]. Recent work also emphasizes that explainability is an audience-dependant instead of a model-inherent property [7, 8, 26, 70, 73]. Explainability is often viewed more broadly than model transparency or intelligibility [31, 62, 82]. For example, a growing research area of XAI focuses on techniques to generate *post-hoc* explanations [27]. Instead of directly elucidating how a model works internally, post-hoc explanations typically justify an opaque' model's decision by rationalizing the input and output or providing similar examples. Lipton discussed the importance of post-hoc explanations to provide useful information for decision makers, and its similarity with how humans explain [62]. At a high level, Gilpin et al. [31] argued that the transparency of model behaviors alone is not enough to satisfy the goal of "*gain[ing] user trust or produc[ing] insights about the cause of the decisions*," but rather, explainability requires other capabilities such as providing responses to user questions and the ability to be audited.

Since an explanation is only explanatory if it can be consumed by the recipient, many recognize the importance of taking user-centered approaches to XAI [70, 89, 100], and the indispensable role that the HCI community should play in advancing the field. While XAI has experienced a recent surge in activities, the HCI community has a long history of developing and studying explainable systems, such as explainable recommender systems, context-aware systems, and intelligent agents, as outlined by Abdul et al. [1]. Moreover, XAI's disconnect with the philosophical and psychological grounds of human explanations has been duly noted [71], as best represented by Miller's call for leveraging insights from the Social Sciences [70]. Wang et al. reviewed decision-making theories and identified many gaps in XAI output to support the complete cognitive processes of human reasoning [101]. From these lines of work, we highlight a few critical issues that are most relevant to our work.

First, there is a dearth of user studies and a lack of understanding on how people actually perceive and consume AI explanations [23, 100]. Only until recently have researchers began to conduct controlled lab studies to rigorously evaluate popular XAI techniques [12, 13, 15, 22, 27, 53, 80], as well as studies to understand real-world user needs for AI explainability [43, 50, 58]. Accumulating evidence shows that XAI techniques are not as effective as assumed. There have been rather mixed results on whether current XAI techniques could appropriately enhance user trust [15, 80, 107] or the intended task performance, whether for decision making [12, 58, 111], model evaluation [6, 13, 22], or model development [50]. For example, Alqarrawi et al. evaluated the effectiveness of saliency maps [6] – a popular explanation technique for image classification models – and found they provided very limited help for evaluating the model. Kauer et al. studied how data scientists use popular model interpretability tools and found them to be frequently misused [50]. Liao et al. interviewed practitioners designing AI systems and reported their struggle with popular XAI techniques due to a lack of actionability for end users. Recent studies also reported detrimental effects of explanations for AI system users including inducing over-trust or over-estimation of model capabilities [50, 90, 95], and increasing cognitive workload [2, 29]. Moreover, while XAI is often claimed to be a critical step towards accountable AI, empirical studies have found little evidence that explanations improve a user's perceived accountability or control over AI systems [81, 90].

Second, in human reasoning and learning, explanation is both a *product* and a *process*. In particular, it is a *social process* [70] as part of a conversation or social interaction. Current technical XAI work typically takes a product-oriented view by generating a representation of a model's internals [65]. However, explanations are also sought first and foremost as a knowledge transfer process from an explainer to an explainee. A process-oriented view has at least two implications for XAI. First, the primary goal of explanation should be to enable the explainee to gain knowledge or make sense of a situation or event, which may not be limited to a model's internals. Second, as a transfer of knowledge, explanations should be presented relative to the explainee's beliefs or knowledge gaps [70]. This emphasis on tailoring explanation according to explainee's knowledge gaps has been a focus of prior HCI work on explainable systems [58–61]. Recent work has also begun to explore interactive explanations that could address users' follow-up questions as a way to fill individual knowledge gaps [91, 104]. However, sometimes these knowledge gaps lie outside of the system, which may require providing information that is not related to its internal mechanics [1].

Finally, we argue that AI systems are socially situated, but sociotechnical perspectives are mostly absent in current XAI work. One recent study by Hong et al. [43] investigated how practitioners view and use XAI tools in organizations using ML models. Their findings suggest that the process of interpreting or making sense of an AI system frequently involves cooperation and mental model comparison between people in different roles, aimed at building trust not only between people and the AI system, but also between people within the organization [43]. Our work builds on these observations, as well as prior work on sociotechnical approaches to AI systems which we review below.

2.2 Sociotechnical approaches to AI

Our work is broadly motivated by work on sociotechnical approaches to AI. Academia and society at large have begun to recognize the detrimental effect of a techno-centric view on AI [85, 89, 100]. Since AI systems are socially situated, their development should carefully consider social, organizational, and cultural factors that may govern their usage. Otherwise one may risk deploying an AI system un-integrated into individual and organizational workflows [66, 106], potentially resulting in misuse, mistrust [108, 109], or having profound ethical risks and unintended consequences, especially for marginalized groups [72, 86, 99].

Researchers have proposed ways to make AI systems more human-centered and sensitive to socio-organizational contexts. Bridging rich veins of work in AI, HCI, and critical theory, such as Critical Technical Practices [5] and Reflective Design [88], Ehsan and Riedl delineate the foundations of a Reflective Human-centered XAI (HCXAI). *Reflective HCXAI* is a sociotechnically informed perspective on XAI that is critically reflective of dominant assumptions and practices of the field [26], and sensitive to the values of diverse stakeholders, especially marginalized groups, in its proposal of alternative technology. Zhu et al. proposed Value Sensitive Algorithm Design [112] by engaging stakeholders in the early stages of algorithm creation, to avoid biases in design choices or compromising stakeholder values. Several researchers have leveraged design fictions and speculative scenarios to elicit user values and cultural perspectives for AI system design [16, 17, 75]. Šabanovic developed a framework of Mutual-Shaping and Co-production [85] by involving users in the early stages of robot design and engaging in reflexive practices. Jones et al [47] 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.

More fundamentally, using a Science and Technology Studies (STS) lens [97], scholars have begun critically reflecting on the underlying assumptions made by AI algorithmic solutions. Mohamed et al. [72] examined the roles of power embedded in AI algorithms, and suggested applying decolonial approaches to enable AI technologies to center on vulnerable groups that may bear negative consequences of technical innovation. Green and Viljoen [32] diagnosed the dominant mode of AI algorithmic reasoning as "algorithmic formalism" – an adherence to prescribed form and rules – which could lead to harmful outcomes such as reproducing existing social conditions and a technologically-deterministic view of social changes. The authors pointed out that addressing these potential harms requires attending to the internal limits of algorithms and the social concerns that fall beyond the bounds of algorithmic formalism. In the context of fair ML, Selbst et al.[87] questioned the implications of algorithmic abstraction that are essential to ML. Abstracting away the broader social context can cause AI technical interventions to fall into a number of traps: Framing, Portability, Formalism, Ripple Effect, and Solutionism. The authors suggested to mitigate these problems by extending abstraction boundaries to include social factors rather than purely technical ones. In a similar vein, field work on algorithmic fairness often found that meaningful interventions toward usable and ethical algorithmic systems are non-technical, and that user community derive most value from localized, as opposed to "scalable" solutions [49, 54].

Our work is aligned with and builds on these views obtained through the sociotechnical lens. These perspectives inform our thinking as we expand the boundaries of XAI to include socio-organizational factors, and challenge a formalist perspective that peoples' meaning-making processes could be resolved through algorithmic formalisms. Our work takes an operational step towards sociotechnical XAI systems by expanding the design space with ST.

2.3 Social transparency and related concepts

Our work is also informed by prior work that studied social transparency and related concepts in human-human interactions. The concept of making others' activities transparent plays a central role in HCI and Computer-Supported Cooperative Work (CSCW) literature [92, 96]. Erickson and Kellogg proposed the concept of and design principles for Social Translucence, in which "social cues" of others' presence and activities are made visible in digital systems, so that people can apply familiar social rules to facilitate effective online communication and collaboration. Gutwin et al.'s seminal work on group awareness [35] for groupware supporting distributed teams provides an operational design framework. It sets out elements of knowledge that constitute group awareness, including knowledge regarding *Who*, *What*, and *Where* to support awareness related to the present, and *How*, *When*, *Who*, *Where*, and *What* for awareness related to the past. Theses theories have since inspired a bulk of work that created new design features and design spaces for social and collaborative technologies (e.g. [28, 30, 36, 51, 67]).

Building upon social translucence and awareness, Stuart et al. [94] conceptualized Social Transparency (ST) in networked information exchange. In particular, it extends the visibility of one's direct partner and the effect on their dyadic interactions, to also encompass one's role as an *observer* of others' interactions made visible in the network. Their framework describes three social dimensions made visible to people by ST: identity transparency, content transparency, and interaction transparency. This framework then considers a list of *social inferences* people could make based on these visible dimensions (e.g. perceived similarity and accountability based on identity transparency; activity awareness based on content transparency; norms and social networks based on interaction transparency), and their second order effects for the groups or community. Social transparency theory has been used to design and analyze various social media features and their impact on social learning [18, 78], social facilitation [44], and reputation management [18].

The above work focused on how ST – making others' activities visible – affects collaboration and cooperative behaviors with other people. Our work also draws on two other important aspects that ST could potentially support for decision-making. One is on knowledge sharing and acquisition. As reviewed by Ackerman et al, [3], CSCW systems supporting organizational knowledge management fall into two categories: a repository model that externalizes peoples' knowledge as sharable artifacts or objects; and an expertise-sharing model that supports locating the appropriate person to have in-situ access to knowledge. The CSCW community's shift from the former to the latter category represents a shift of emphasis from *explicit* to *tacit* knowledge. Transparency of others' communications could facilitate expertise location through the acquisition of organizational meta-knowledge (e.g., *who knows what* and *who knows whom*), as a type of "ambient awareness" coined by Leonardi in the context of enterprise social media [55, 56]. This position is also related to the development of *Transactive Memory Systems (TMS)* [10, 74, 77, 110] that relies on meta-knowledge to optimize the storage and retrieval of knowledge across different individuals. A sufficiently fluent TMS can evolve to a form of team cognition of or "collective mind" [46, 103] that can lead to better collective performance [9, 42].

Social transparency could also guide or validate peoples' judgment and decision as cognitive heuristics. Cognitive heuristics are a key concept in decision-making [48], which refers to "rules of thumb" people follow to quickly form judgments or find solutions to complex problems. By making visible what other people selected, interacted with, approved or disapproved, ST could invoke many social and group-based heuristics such as bandwagon or endorsement heuristics (following what many others' do), authority or reputation heuristics (following authority), similarity heuristics (following people in similar situations), and social presence heuristics (favoring a social entity over a machine) [68, 69, 98]. How these ST-rendered heuristics affect peoples' decisions and actions has been studied in a wide range of technologies such as reputation systems [83] and social media. In particular, they play a critical role in how people evaluate the trustworthiness, credibility, and agency of technologies [68, 69, 98], as well as the sources or organizations behind the technologies [44, 52]. While these heuristic-based judgments are indispensable for people to navigate complex technological and social environments, they also lead to biases and errors if inappropriately applied [48], calling for careful study of inferences people make based on ST features and their potential effect.

Our concept of social transparency in AI systems is informed by the aforementioned perspectives, but with several key distinctions: at the center of our work is a desire to support the explainability of AI systems, particularly in AI-mediated decision-making. We are not merely interested in making *others'* activities visible, but more importantly, how *others' interactions with AI* impact the explainability of the system. Within the view of a human-AI assemblage, in which both AI and people have decision-making agency, it is possible to borrow ideas and interpretative lenses from work studying ST in human-human interactions. To study the effects of ST in AI systems, our first-order focus is on users' sense-making of an AI system and their decision-making process, though it may inevitably impact their organizational behaviors as well.

3 SOCIAL TRANSPARENCY IN AI SYSTEMS: A SCENARIO BASED DESIGN EXPLORATION

After identifying an epistemic blind spot of XAI, we propose adding Social Transparency (ST) into AI systems–incorporating social-organizational contexts to facilitate explainability of AI's recommendations. This definition is intentionally left broad, as we follow a broad definition of explainability–ability to answer the why-question. We borrow the term ST from Stuart et al. [94], and similarly emphasize both making visible of other people in the human-AI assemblage, and other people's interactions with the "source", in our case, the AI system. Different from Stuart et al., which proposed the ST concept retrospectively at a time when ST enabling features were pervasive in CSCW systems, we had to consider, prospectively, what kind of features to add to an AI system to make ST possible.

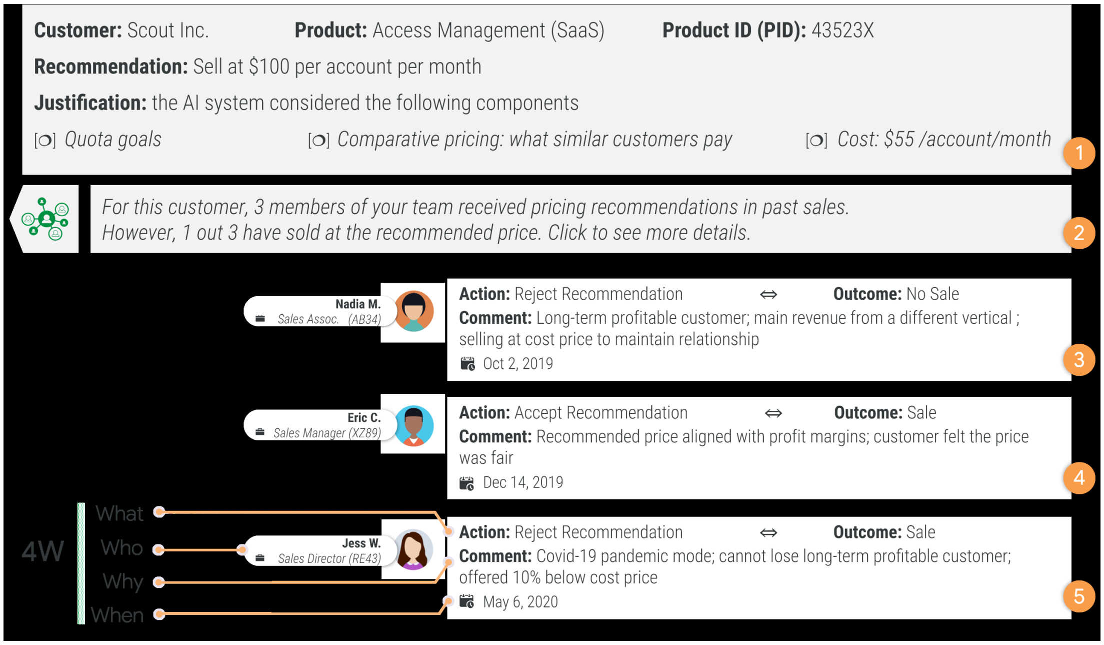
As a formative step, our goal was not to develop a finished treatise of ST in AI systems. Rather, we intended to create an exemplary design of an AI system with ST and use it to conduct formative studies to advance our conceptual development. We opted for a scenario-based design (SBD) method. SBD suspends the needs to define system operations by using narrative descriptions of how a user uses a system to accomplish a task [84]. SBD allows interpretive flexibility in a user journey by balancing between roughness and concreteness. SBD is an appropriate choice for our investigation because it is a method oriented for "envisioning future use possibilities" [84], focusing on people's needs, evocative, and has been adopted in prior XAI design work [106].

We started with a range of AI-mediated decision-making scenarios around cybersecurity, hiring (employment), healthcare, and sales, where a user encounters an AI recommendation and seeks answer to a *why*-questions about the recommendation, e.g. "why should I accept or trust the recommendation". We ran 4 workshops with a total of 21 people from 8 technology companies who are users or stakeholders of relevant AI systems. The scenarios started in a textual form, then we engaged participants in drawing exercises to create visual mock-ups of these scenarios (hereby referred to as visual scenarios), and brainstorming together what kind of information they wanted to see about *other users* of the AI system, and *other users' interactions with the AI system* if they were the user. When it came to types of design feature that could encode relevant socio-organizational context, people had many suggestions. For instance, suggestions about knowing what happened to other people getting recommendations from the AI systems, who got the recommendations, etc. quickly emerged in the discussions. The ideas converged to what our participants coined as the "4W"—*who* did *what* with the AI system, *when*, and *why* they did what they did— in order to have adequate socio-organizational context around the AI-mediated decisions.

We note an interesting observation that the 4W share similarity with the design elements for group awareness in groupware work [35], with the exception of "why", which is core to explainability. When thinking how to represent the "why", participants suggested an open ended textual representation to capture the nuances behind a decision. Eventually, we settled on a design of a "commenting" feature (why) together with traces of others' interactions with the AI system's recommendations (what), their identities (who) and time of interactions (when). In the rest of the paper, we refer to these constitutive design elements of ST as 4W. Figure 1 shows the final visual scenario with the 4W features used in the interview study.

We chose a sales scenario around an AI-mediated price recommendation tool, since it appeared to have a broader reach and accessibility even for workshop participants who did not work in a sales domain. In the study, we intended to interview sellers as targeted users of such an AI system, and also non-sellers to explore the transferability of the ST concept to other AI domains, as we will discuss in detail in the next section.

Figure 1: Visual scenario used in the interviews, labeled by blocks to be revealed in the interview in order: (1) Decision information and model explanation: Information of the current sales decision, the AI's recommended price and a "feature importance" explanation justifying the model's recommendation, inspired by real-world pricing tools; (2) ST summary: Beginning of ST giving a high-level summary of how many teammates in the past had received the recommendation and how many sold at the recommended price; (3-5): ST blocks with "4W" features containing the historical decision trajectory of three other users.



Customer: Scout Inc. **Product:** Access Management (SaaS) **Product ID (PID):** 43523X

Recommendation: Sell at $100 per account per month

Justification: the AI system considered the following components

Quota goals *Comparative pricing: what similar customers pay* *Cost: $55 /account/month*

For this customer, 3 members of your team received pricing recommendations in past sales. However, 1 out 3 have sold at the recommended price. Click to see more details.

Nadia M. *Sales Assoc. (AB34)*

Action: Reject Recommendation ⇔ **Outcome:** No Sale

Comment: Long-term profitable customer; main revenue from a different vertical ; selling at cost price to maintain relationship

Oct 2, 2019

Eric C. *Sales Manager (XZ89)*

Action: Accept Recommendation ⇔ **Outcome:** Sale

Comment: Recommended price aligned with profit margins; customer felt the price was fair

Dec 14, 2019

Jess W. *Sales Director (RE43)*

Action: Reject Recommendation ⇔ **Outcome:** Sale

Comment: Covid-19 pandemic mode; cannot lose long-term profitable customer; offered 10% below cost price

May 6, 2020

Design choices in the visual scenario: We ran 4 pilot studies to finalize the design of the visual scenario in Figure 1, and the procedure to engage participants with the design. We scoped the number of 4W blocks to three to strike a balance between a variety of ST information and avoiding overwhelming the participants, based on what we learned from the pilot studies. Each of the 4W are represented by one or more design features: accepting and rejecting the AI (action [what]), succeeding and failing to make the sale (outcome [what]), one's name, profile picture and organizational role ([who]), a comment on the reasons behind the action ([why]), and a time stamp ([when]). Contents in these components were inspired by the workshop discussions, and showcase a range of socio-organizational contexts relevant to the decision. The pilot runs revealed that presenting the entire visual scenario creates cognitive and visual clutter. Therefore, for the interview, we decided to reveal the five blocks shown in Figure 1 one by one, with the interviewer verbally presenting the narrative around each block.

4 STUDY METHODS

In this section we share the methodological details of the semi-structured interviews.

4.1 Recruitment

As mentioned, we intended to recruit both sellers and non-sellers, who are stakeholders of other AI-mediated decision-making domains. Stakeholders are not limited to end users. We also welcomed different perspectives from designers, data scientists, etc. With this in mind, we recruited participants from six different companies, including a large international technology company where we were able to recruit from multiple lines of products or sales divisions. The recruitment was initiated with an online advertisement posted in company-wide group-chat channels that we considered relevant, followed up by snowball sampling. The advertisement stated two recruiting criteria: First, they needed to have direct experience using or developing or designing an AI system. Second, the AI system should be interacted by multiple users, preferably with multi-user decision-making. We verified that these criteria were met through a series of correspondence (via online messaging) where each participant shared samples of the AI system they intended to discuss.

A total of 29 participants were recruited (17 self-identified as females while the rest as males). The recruitment of sellers turned out to be challenging, given their very limited availability. By using snowball sampling, we were able to recruit 8 sellers. For non-sellers, the snowball sampling resulted in participants clustered in two major domains – healthcare and cybersecurity. We conducted the study in the middle of Covid-19 (a global pandemic in 2020), which added non-trivial burden to the recruitment process and limited our interviews to a remote setting using video conferencing tools. Participants' ID, role, domains and domain experience is shared in Table 1. To facilitate traceability in the data presented hereafter, we differentiate sellers and non-sellers by appending the participant ID with *-S* for sellers and *-NS* for non-sellers (e.g., 1-S for a seller and 2-NS for a non-seller).

4.2 Interview procedure

The semi-structured interviews were conducted online with screen-sharing for the interviewer to present the visual scenario. All interviews were video recorded including the screen activities. The interview had 4 main parts. In the *first* part, after gaining informed consent, we asked participants to share about an AI system that they were currently engaged with, focusing on their or their users' needs for explainability. We also inquired about the socio-organizational context around the use case, both before and after the AI system was introduced.

The *second* part involved a deep dive into the speculative design with a walk-through of the visual scenario in Figure 1. This is where we explored how incorporation of ST can impact an AI-mediated decision-making scenario, as we revealed the different blocks of the visual scenario in a sequenced manner. Participants were asked to play the role of a salesperson trying to pitch a good price for an Access Management software to Scout Inc. (a client). In the first block revealed, the AI not only recommends a price, but also shows a technical explanation–a set of model features (e.g., cost price, quota goals, etc.) justifying the recommendation. Once the participant showed a good enough understanding on the Decision Information and Model Explanation portion (block 1 in Figure 1), we asked the participant to give a price they would offer and their confidence level (between 1-10, 10 being extremely confident) given what they saw on the screen. Next, we revealed the social transparency portions. First, it was the ST Summary (block 2 in Figure 1) followed by each of the 4W blocks (block 3-5 in Figure 1). We allowed participants to read through the content and guided them through any misunderstandings. They were encouraged to think-aloud during the whole process. Following this, we asked participants to share the top three reactions to the addition of the ST features, either positive or critical. After that, participants were asked to share their final price and confidence level. In addition, we asked them to rank the importance of the 4W (*who*, *what*, *when*, and *why*) for their decision-making process and justify their ranking.

The *third* part was about zooming out from the visual scenario and brainstorming plausible and impactful transfer scenarios of ST in domains our participants resided. At this point, we also gave them a conceptual definition and some vocabulary around ST so that they could brainstorm with us effectively. The goal of this part was to explore the design and conceptual space of ST in domains beyond the sales scenario. For sellers, this meant transferring to their own sales work context, which helped refining our own understanding of the sales scenario. Once participants shared their thoughts on the transferability of ST, they ranked the 4W in the transfer use cases. We wanted to see if there are variations in the rankings as the context switches—an aspect we discuss in the Findings section.

The *fourth* and final part involved discussions around potential unwanted or negative consequences of ST as well as reflective conversations on how incorporation of ST can impact explainability of AI systems.

In summary, in addition to open-ended discussions, our interview collected the following data points from each participant: original and updated price decisions and associated confidence levels, rankings of 4W for both the sales scenario and one's own domain. While our study was not designed to quantitatively evaluate the effect of ST, we will report summary statistics of these data points in the Findings section, which helped guiding our qualitative analysis.

4.3 Qualitative Analysis of the interviews

The interviews lasted 58 minutes on average. We analyzed the transcription of roughly 29 hours of interview data using a combination of thematic analysis[11] and grounded theory [93]. Using an open coding scheme, two authors independently went through the videos and transcription to produce in-vivo codes (directly from the data itself). Then we separately performed a thematic analysis, clustering the codes from in-vivo coding to themes. We iteratively discussed and agreed upon the codes and themes, constantly comparing and contrasting the topics each of us found, refining and reducing the variations in each round till consensus was reached. We grouped the codes and themes at the topic level using a combination of mind-mapping and affinity diagramming. Our results section below is organized thematically, with the top-level topics as subsections. When discussing each topic, we highlight codes that add to that topic in **bold**.

5 FINDINGS

We begin by sharing how participants' own experience with AI systems demonstrates that technical transparency alone does not meet their explainability needs. They need context beyond the limits of the algorithm. Next, based on how participants reacted to the incorporation of ST in the design scenario, we unpack what context could be made visible by ST and break down the implications at three levels: **technological (AI)**, **decision-making**, and **organizational**, as summarized in Table 2. We further discuss specific aspects of socio-organizational context that the 4W design features carry and their effects, summarized in Table 3. Based on input from non-seller participants, we also share insights about the potential transferability of ST beyond the sales domain. We end this section with participants' discussions on the challenges, risks, and tensions of introducing ST into AI systems.

As mentioned, all participants, both sellers and non-sellers, experienced the sales scenario and reflected on its transferability to their own domains. Our analysis revealed substantial alignment between the two groups, possibly due to the accessible nature of our intentional choice of a sales domain and the content in the scenario. With the exception of Section 5.6, which focuses on non-sellers' reflection on transferability of ST, we report the results combining the two groups, but mark their IDs differently (*-S* or *-NS*) as shwon in Table 1.

5.1 Technical Transparency is not enough

As we began each interview with participants' own experience with AI systems, a core theme that lies at the heart of our findings is the realization that solely relying on technical or algorithmic transparency is not enough to empower complex decision-making. There is a shared understanding that AI algorithms cannot take into account all the contextual factors that matter for a decision: "not everything that you need to actually make the right decision for the client and the company is found in the data" (P25-NS). Participants pointed to the fact that even with an accurate and algorithmically sound recommendation, "there are things [they] never expect a

Table 1: Participant details

Participant ID	Role	Domain	Years of Experience
1-S	Seller	Sales	> 10
2-NS	Designer	Cybersecurity	> 5
3-NS	Designer	Finance and Travel	> 5
4-NS	Consultant	Gov and Non-profit	> 3
5-S	Seller	Sales	> 5
6-NS	Designer	Health- Oncology	> 5
7-NS	Data Scientist	Cybersecurity	> 8
8-S	Seller	Sales	> 3
9-NS	Designer	Health- Radiology	> 5
10-NS	Data Scientist	Cybersecurity	> 10
11-NS	Designer	Health	> 3
12-NS	Designer	Cybersecurity	> 5
13-S	Seller	Data Analytics	> 10
14-NS	Data Scientist	NLP	> 5
15-NS	Designer	Health- Radiology	> 5
16-NS	MD/ Data Scientist	Health- Oncology	> 10
17-NS	Manager	HR	> 5
18-S	Seller	Sales	> 3
19-S	Seller	Sales	> 3
20-S	Seller	Sales	> 10
21-S	Seller	Sales	> 3
22-NS	SOC analyst	Cybersecurity	> 3
23-S	Seller	Sales	> 5
24-NS	SOC analyst	Cybersecurity	> 5
25-NS	SOC analyst	Cybersecurity	> 3
26-NS	SOC analyst	Cybersecurity	> 5
27-NS	SOC Data Scientist	Cybersecurity	> 5
28-NS	SOC Architect	Cybersecurity	> 10
29-NS	SOC analyst	Cybersecurity	> 5

machine to know [such as] clients' allegiances or internal projects impacting budget behavior" (P1-S). Often, the context of social dynamics that an algorithm is unable to capture is the key: "real life is more than numbers, especially when you think of relationships" (P12-NS). Discussing challenges in interpreting and using AI recommendations in Security Operation Centers (SOC), P29-NS highlighted the need for awareness of others' activities in the organizational context:

> Sometimes, even with perfect AI, the most secure thing is to do nothing because you don't know what the machine doesn't know. There is no centralized process to tell us the context of what's going on elsewhere, what others are doing. One move has ripple effects, you know. So instead of using [the AI's recommendation], they end up basically doing the most secure thing– don't touch anything. That's where the context helps from your colleagues. That's how actually work really gets done. (P29-NS, a SOC director)

Moreover, even when provided, technical transparency is not always understandable for end users. While describing how he uses an AI-assisted pricing tool, this seller pointed to how the machine explained itself by sharing a "confidence interval" along with a description of how the AI works, which was meaningless to him:

> I hate how it just gives me a confidence level and gibberish that the engineers will understand. There is zero context. The only reason I am able to use this tool is [through] guidance from other sellers who gave me the background information on the lead I needed to generate a quote worth their time. (P23-S, senior salesperson using a pricing tool to generate a quote)

In complex organizational settings, answers to the why-question, i.e. knowledge needed to understand and take informed action for an AI mediated decision, might lie outside the bounds of the machine. As highlighted above, participants repeatedly desired for "context" to "fill in the gaps" (P27-NS). The ST information in our design scenario is intended to provide such context. After going through the ST portion, 26 out of the 29 participants lowered their sales prices, resulting in a mean final price of $73.8 (SD=$15.8), compared to a mean initial price of $110.7 (SD=$57.2) based only on the AI's recommended price of $100. 24 out of the 29 participants also increased the confidence ratings for their decisions, resulting in a mean final confidence score of 8.3 out of 10 (SD=0.9), compared to a mean initial confidence score of 6.4 (SD=1.7). These patterns suggest that ST information helped participants to set their price more cautiously and feel more confident about their decisions, by "help[ing] [them] to understand the situation more holistically" (P19-S). This participant succinctly summarized this perspective at the end of her interview:

> You can't just get everything in the data that trains the model. The world doesn't run that way. So why rely just on the machine to make sense of things that are beyond it? To get a holistic sense of the "why" you should or should not trust the AI, you need context. So the context from Social Transparency adds the missing piece to the puzzle of AI explainability". (P2-NS, a designer of cybersecurity AI systems)

Now we analyze participants' reaction and reflection from seeing ST features, and unpack the "context" made visible by ST and its effects at three levels: **technological (AI)**, **decision-making**, and **organizational**. For the subsections below each dedicated to a level of context, we begin by summarizing the effects of ST, with codes from the data in bold. These results are summarized in Table 2.

5.2 Technological (AI) context made visible

ST makes visible the socially-situated **technological context**: the trajectory of AI's past decision outputs as well as people's interactions with these technological outputs. Such contextual information could help people **calibrate trust in AI**, not only through **tracking AI performance**, but also by **infusing human elements** in AI that could invoke social-based perception and heuristics.

Records of others' past interactions with the AI system paints a concrete picture of the AI performance, which technical XAI solutions such as performance metrics or model internals would not be able to communicate. Participants felt that the technological context they understood through ST helped them better gauge the AI's limitations or "actual performance of the AI" (P10-NS). In fact, after going through the sales scenario, many reported on re-calibrating their trust in the AI, which is key to preventing both over-reliance of AI and "AI aversion" [21]:

> Knowing the past context helps me understand that the AI wasn't perfect. It's almost like a reality check. The comments helped because real life is more than numbers. I am more confident in myself that I am making the right decision but less trust[ing] the AI. (P12-NS, an XAI designer)

ST could also affect people's perception of and trust in the AI system by infusing the much needed human elements of decision-making in the machine. Participants from each of the domains (sales, cybersecurity, healthcare) highlighted that "there is a human aspect to [their] practice" (P6-NS), something that "can never be replaced by a machine" (P6-NS). Adding these human elements allows one to apply familiar social rules. Many participants commented on a "transitive trust" (P4-NS) from trusting their peers – "people are trained to believe [their] peers and trust them" (P25-NS) – to trusting the AI system, if others were using the AI systems or accepting the AI's recommendations. For instance, in the sales domain of the scenario, a transitive trust is often fostered by an organizational hierarchy or job seniority "as precedence and permission for doing the right thing" (P12-NS). Radiologists often want to "know who else used the same logic and for what reason" (P15-NS) when working with AI-powered diagnostic tools. In cybersecurity, "knowing that a senior analyst took a certain route with the recommendation [can be] the difference maker" (P28-NS). Some participants also commented on a positively perceived "humanizing effect" of AI by adding ST, that "[users] would potentially adopt... showing them like [AI is] supporting you not replacing you" (P6-NS).

The above discussions show that ST could support forming appropriate trust and evaluation of AI through two essential routes, as established in prior work on trust and credibility judgment of technologies [68, 69, 98]: a central route that is based on a better understanding of the AI system, and a peripheral route by applying social or group-based heuristics such as social endorsement, authority, identity, or social presence heuristics. While the central route tends to be cognitively demanding, the peripheral route is fast and easy, and could be especially impactful to help new users to enhance their trust and adoption of an AI system.

5.3 Decision-making context made visible

ST also makes visible the **decision context** – the local context of past decisions – for which many participants described as "in-situ access" to "crew knowledge" [1]. We will first elaborate on the notion of **crew knowledge**, then discuss how a combination of decision trajectory, historical context and elements of crew knowledge could 1) lead to **actionable insights**, which could **improve decision-making**, **boost decision confidence** and **support follow-up actions**; 2) provide social validation that facilitates **decision-making resilience** and **contestability of AI**.

The notion of *crew knowledge* emerged during our discussions with many participants regardless of their domains. When asked to elaborate on the concept, participants defined it as "informal knowledge acquired over time through hands-on experience", knowledge that is not typically "gained through formal means, but knowledge that's essential to do the job" (P8-S). Crew knowledge is learned "via informal means, mainly through colleague interactions" (P23-S). It can encode "idiosyncrasies like client specific quirks" (P27-NS). Participants referred to their team as their "crew", with a sense of identity and belonging to a community membership. We can think of crew knowledge as informal or tacit knowledge that is acquired over time and locally-situated in a tight-knit community of practice–an aggregated set of "know-hows" of sorts. While ST features may not explicitly encode a complete set of crew knowledge, they provide in-situ access to the vital context of past decisions that carry elements of crew knowledge.

[1] The original term used by most participants was "tribal knowledge", which is a term often used in business and management science to refer to unwritten knowledge within a company. We note that, from an Indigenous perspective particularly in North America, the words "tribe" and "tribal" connote both an official status as a recognized Nation, and also a profound sense of identity, often rooted in cultural heritage, a specific ancestral place, and a lived experience of the on-going presence of tribal elders and ancestors (past, present, and future). In our case, participants used the word "tribal" in a non-Indigenous meaning. Being sensitive to potential mis-use of the word, we engaged in critical conversations with potentially affected community members to understand their perspectives. The conversations revealed that it is best to avoid using that word. We went back to the participants who used the word "tribal" and asked if "crew" captures the essence of what they meant by "tribe". All of them agreed that the words were interchangeable. As such, we only present the data using the term "crew knowledge".

Table 2: Results on the three levels of context made visible by ST and their effects. "–" in the last column indicates first-order to second-order effect(s)

Levels	Context made visible	Effects of the visibility
Technological (AI)	Trajectory of AI's past decision outputs and people's interactions with these outputs	Tracking AI performance – Calibrate AI trust Infusing human elements – Calibrate AI trust
Decision-making	Local context of past decisions and in-situ access to decision-related (crew) knowledge	Actionable insights – Improve decisions; Boost decision confidence; Support follow-up actions Social validation – Decision-making resilience; AI contestability
Organizational	Organizational meta-knowledge and practices	Understanding organizational norms and values – Improve decisions; Set job expectation Fostering auditability and accountability Expertise location – Develop TMS

The central position of crew knowledge in participants' responses demonstrates that ST can act as a vehicle for knowledge sharing and social learning in "one consolidated platform" (P21-S). Participants repeatedly mentioned two types of insights they gained from ST to be particularly actionable for AI-mediated decisions. The first is additional variables important for the decision-making task that are not captured in the AI's feature space. For example: "I have a lot more variables that I'm aware of to consider, like, the whole pandemic thing..."(P12-NS). These additional variables are often tacit knowledge, idiosyncratic to the decision, or constantly changing, making them impossible to be formalized in an algorithm. ST could support in-situ access to these variables.

Second, ST supports analogical reasoning with similar decisions and their actual outcomes. Participants exhibited a tendency to reason about the similarity and differences between the contexts of the current decision and past decisions made visible by ST. For example: "what did other oncologists do for a patient like that? So, what treatments were chosen for patients like this person?" (P6-NS) or "I see the reasoning why they didn't pay the recommended price the other time... but those were different circumstances and look now, they're were growing customer and we need to push them up closer to the more profitable price" (P1-S).

Gaining actionable insights could ultimately boost decision confidence, as most participants commented on increasing their confidence in the final price. We also observed an interesting bifurcation on how they conceptualize confidence in the AI versus confidence in oneself after being empowered with knowledge about the decision context. This quote encapsulated that perspective well:

> The system will go by the numbers but I have my "instincts" thanks to my [crew] knowledge. With these comments, you can say I also have my team's "instincts" to help me. So I am less confident on the AI but more in myself due to the 360 view I have of things– I have more information than the machine. (P22-NS)

Moreover, participants commented that learning from the decision context could also support follow-up actions such as interacting with clients or "justifying" (P1-S) the decision to supervisors, as illustrated in the quote below:

> And I actually learned a lot. I learned from their comments... I feel like this is an education for the next sale. Even [if it is] another customer, I will be more confident...and know what to do with [the AI's recommendation] because I know how to evaluate it. (P12-NS)

Learning about past decisions from others, especially higher echelons of the organizational hierarchy, also provided social validation. Social validation can reduce the feelings of individual vulnerability in the decision-making process. While going through the sales scenario, participants would often comment how "the director (Jess) offering discounts gives [them] the permission to do the same" (P12-NS). Being able to have a "direct line of sight into the trajectory of how and why decisions were done in the past" (P24-NS) can make one feel empowered, especially if one has to contest the AI. For most participants, their use of AI systems was mandated by their employers. Many a time, the technology got in the way, becoming a "nuisance" (P1-S) they needed to "fight" (P5-S). Contesting the machine often requires time-consuming reporting and manual review, which creates a feeling that one "can't just say no to the AI" (P25-NS). This participant elaborated on the vulnerability and how social validation could empower one to act:

> People are afraid—they don't want to screw up. You look like a dumb*** if you end up in the war room and say you goofed up because you blindly followed the machine. Even if you have at least one other person doing something similar with the AI, you are safe. Just that knowledge is enough to act less scared. [If] your neck is on the line, someone else's is also on the line. It distributes the risk. (P26-NS)

5.4 Organizational context made visible

Lastly, ST gives visibility to the broader **organizational context**, including the meta-knowledge about the organization such as who knows what and organizational practices. Different from decision context, which makes visible knowledge localized to the decision, organizational context reflects macro-information about the organization. This differentiation shares similarity with the concepts of content versus interaction transparency in Stuart et al.'s ST in social network [94], which emphasizes that transparency of others' interactions enables awareness of "normative behaviors as well as the social structure within a community". We observed that such awareness could then: 1) inform an **understanding of organizational norms and values** that help **improve decision-making** and **calibrate people's overall job expectations**; 2) foster **accountability and auditability**, and 3) facilitate **expertise location**, and if done right, over time the **formation of a Transactive Memory System (TMS)** [110]. In short, organizational context made visible by ST could foster effective collective actions in the organization and strengthen the human-AI assemblage.

Visibility of others' actions in an organization (what's done) could translate into an understanding of organizational norms (what's acceptable) and values (what's important), which might be otherwise neglected since "norms are often not enshrined in a rule book" (P25-NS). From the comments in the scenario, participants were informed of organizational norms: "the fact that a director offered the discount below cost price means that this is something that's acceptable. I might be able to do" (P12-NS) and values: "seeing Jess [the director in our scenario] give such a steep discount and noting how she did it to retain a customer, tells [us] that relationship matters to this company" (P28-NS). This type of insight is crucial for making informed decisions and setting overall job expectation, especially for new employees to "learn about the culture of the company" (P16-NS). The following participant succinctly summarized this point:

> The comments...get me a sense of what should be done, what's expected of me, and what I can also get away with. It tells me what this company values. This helps me understand why certain things are done the way they are, especially if they go against what the AI wanted me to do. This actually explains why I need to do something. (P25-NS).

The enactment of ST in an AI system shared across an organization enables accountability. Participants felt that if they knew "who did what and why, [then] it provides a nice way to promote accountable actions" (P26-NS). Participants noted that currently there is a level of opaqueness in workers' decision-making processes, making it difficult to uphold accountability, be it during bank audits, sales audits, or standardization on health interventions. ST, according to them, can provide "peripheral vision" (P29-NS) that can boost accountability by not only making past decisions traceable, but also socially-situated to better evaluate and attribute responsibilities for, as highlighted by this quote:

> I think these comments would be extremely important for audits and postmortems after an attack. The traceability is huge. (P26-NS, a senior SOC analyst)

That being said, there is a potential double-edged-sword nature to traceability and accountability, where people might feel they are being watched or surveilled. The same participant (P26-NS) articulated this concern:

> You know, there is a dark side to this. If you are part of organizations that love to surveil people, then you are out of luck. That is why organizational culture is so important... [In our company], we focus on the problem not the person. But you can't really say this applies [everywhere]. (P26-NS)

ST also provides awareness of organizational meta-knowledge [56], such as who does or knows what, and who knows whom. Many participants reacted to the scenario with reaching out to relevant people made visible through ST: such as "who was driving that sales" (P3-NS), or "reach out to Jeff just because it's the most recent and find out what's going on" (P5-S). It shows that ST could potentially solve a pain point for larger, distributed organizations by supporting expertise location.

Beyond expertise sharing, some participants commented that knowing whom to reach out to could facilitate the creation of an "institutional memory" (P28-NS), the passing of "legacy knowledge" (P2-NS), and the ability to "leverage broader resources to lean on" (P8-S). These comments resonate with the core concept of transactive memory systems (TMS) [10, 74], which explains how a group or organization collectively manages the distribution and retrieval of knowledge across different individuals, often through informal networks rather than formal structures [77]. TMS could facilitate employee training and benefit new members:

> You can't survive without institutional memory. . . [but] it's never written anywhere and is always in the grapevines. Even if some of it could be captured like this [with ST], then that's a game changer... Training newcomers is hard especially when it comes to getting that "instinct" on the proper way to react to the [security] alerts. Imagine how different training would be if everything was there in one place!" (P28-NS)

A TMS could also facilitate a peer-to-peer support system that gives employees a sense of community:

> What I really love is the support system you can potentially create over time using ST. This actually reminds of the knowledge repo[sitory] my colleagues and I have set up where we add our nuggets of client specific wisdom which helps others operate better. As you know, we are a virtual team so having this collective support is crucial. (P27-NS, a SOC data scientist)

Through tight interactions of the community and repeatedly seeing others' decision processes, a TMS can, over time, enable to formation of a collective mind [103, 110]–members of a group form a shared cognitive or decision schema and construct their own actions accordingly. Collective mind is associated with enhanced organizational performance and creativity. Interestingly, one participat speculated on how ST can be construed as "mindware":

> This almost reminds me of a mindware in a team, sort of like a group mind. Currently, we tie our [security] incident reports to a slack channel and that acts as a storage of our collective memories... We even have tagged comments, so when you showed me your thing, it reminded me of that. (P25-NS)

5.5 Design for ST: the 4W

With the effects of ST at the three levels in mind, now we discuss how participants reacted to specific design features that are intended to reflect ST. As discussed in Section 3, our co-design

Table 3: Summary of the design features, supported effect, and rank of the "4W" features

Category	Design Features	Supported Effect	Overall Rank
What	Action taken on AI Decision outcome Summary statement	Tracking AI performance Machine contestability (Social validation)	1st
Why	Comments with rationale justifying the decision	Tracking AI performance Actionable insights Understanding organizational norms and values Social validation	2nd
Who	Name Organizational role/ job title Profile picture	Social validation Transitive trust (Infusing human elements) Expertise location	3rd
When	Timing of the decision	Temporal relevance (actionable insights)	4th

exercises informed the choices of constitutive elements of ST: *Who* did *What*, *When*, and *Why*, referred as the 4W. The reader might recall that participants were asked to rank and justify the relative importance of 4W twice during the interview. The first ranking was done in the sales scenario. The second was done in discussing the transferability to participants' own domains. By explicitly inquiring about their thoughts and preferences around the 4W, it helped us understand the effects of each of these design features in facilitating ST and the remaining challenges.

Despite domain-dependent variations, we found overall patterns of preference that are informative. In the sales scenario, first, participants wanted to know "what happened?" (mean rank= 1.90). If the outcome was interesting, then they wanted to delve deeper into the "why" (mean rank= 1.97), followed by "who" (mean rank= 3.03) did it and "when" (mean rank= 3.10). The relative order of the 4W remained stable when discussing transferability to participant's individual domains. Table 3 summarizes the 4W design features and the types of effect they support based on the codes emerged in the interviews. These codes correspond to the effects of the three levels of context shown in Table 2.

5.5.1 What: In our scenario, the "what" is conveyed by two design features – whether (a) a previous person accepted or rejected the AI's recommendation and (b) whether the sale was successful or not [the outcome]. There was also a summary feature of What appeared at the top (block 2 in Figure 1). Citing that the outcome is the "consequence of [their] decision" (P16-NS), participants felt that it was a must-have element of ST. Participants referred to the "what" as the "snapshot" of all ST information (P5-S, P8-S, P15-NS, P28-NS) which gave them an overview of the AI's performance and others' actions, and guided them to decide "do I want to invest more time and dig through" (P15-NS). As the scenario unfolded in the beginning, seeing the summary What feature often invoked a reaction that one should be cautious from over-relying on the AI's recommendation, but seek further information to make an informed decision. On a more constructive note, especially when thinking through transfer scenarios in radiology and cybersecurity, participants highlighted the need to present the appropriate level of details so that it does cognitively burden the user– "the outcome should be a TL;DR. The 'why' is there if I am interested" (P29-NS).

5.5.2 Why: The "Why" information was communicated in free-form comments left by previous users in our scenario. Participants often referred to the "why" as the "context behind the action" and "to understand the human elements of decision-making" (P17-NS). They felt that the "why" could not only help them understand areas that the technology might be lacking, but also "explain the human and the organization" (P28-NS). "Understanding the rationale behind past decisions allows [one] to make similar decisions… [and] gives you an idea of what you should be doing" (P18-S). Prior rationales can also "give [humans] a justification to reject the machine" (P7-NS) by "know[ing] why someone did something similar" (P28-NS). In short, insights into the why can inform AI performance, provide actionable insights and social validation for the decision, as well as facilitate a better understanding of organizational norms and values. Social validation, in particular, can enable contestability of AI. On a constructive note, participants highlighted the need to process or organize the comments to make them consumable: "not all whys are created equal, [and that there is a] need to ensure things are standardized" (P25-NS). There were concerns that if comments are not quality controlled, they might not serve the purpose of shedding context appropriately. Citing "no one wants a lawsuit on their hands" (P26-NS), participants also suggested the need to be vigilant about compliance and legal requirements to ensure private details (e.g., proprietary information) is not revealed.

5.5.3 Who: The "Who" information in our scenario included multiple elements: a previous user's name, position and a displayed profile picture. Participants engaged with the implications of "who" at multiple levels. For many, "who" was the bare minimum that they needed for expertise location – "if [I] knew who to reach out to… [I] could find out the rest of the story" (P5-S). For others, knowing someone's organizational role or level of experience is more important, because "hierarchy matters" (P16-NS) and one's experience level influences the "degree of trust we can place on other people's judgement" (P2-NS). Thus the identity information could affect both social validation for one's own decision and transitive (dis) trust in AI. On a constructive note, some reflected on how "the collective 'who' matter[ed]" (P6-NS) and there needs to be consistency across personnel for them to make sense of the decision. Some participants raised concerns of the profile picture and the name displayed in our scenario. They felt that these features can lead to biases in weighing different ST information. Others welcomed the profile information because it "humanizes" the use of AI (P1-S). Here, the domain of the participant appeared to matter – most salespeople welcomed complete visibility; many of the stakeholders from healthcare and government service domains raised concerns. Such perception differences across domains highlight that we need to pay attention to the values in the community of practice as we design these features.

5.5.4 When: The "When" information is expressed by a timestamp. Participants felt that the timing can dictate "if the information is still relevant" (P11-NS), which informs the actionability of context they gain from ST. Knowing the *when* "puts things into perspective" (P16-NS) because it adds "context to the decision and strengthen[s] the why" (P18-S). Timing was particularly useful when participants deliberated on which prior decision they should give more weight to. One comment in the scenario highlighted how Covid-19 (a global pandemic in 2020) influenced the decision-maker's actions. At the time of the interviews, the world was still going through Covid-19. The "when" "aligned things with a timeline of events and how they transpired" (P21-S).

5.6 Transferability of ST to other domains

After participants engaged with the sales scenario, we debriefed them on the conceptual idea of adding ST to AI systems. We then asked them to think of transfer scenarios by envisioning how ST might manifest in their own domains or use cases. As Table 1 shows, except for 3 people (P4-NS, P11-NS, P17-NS), our participants came from three main domains: sales, cybersecurity, and healthcare (radiology and oncology). Here we give an overview of how participants viewed the potential needs and impact of ST in cybersecurity and healthcare domains.

5.6.1 Cybersecurity: Participants working in cybersecurity domain were unanimous in their frustration about the lack of awareness of how their peers make use of AI's recommendations. They saw a rich space where the incorporation of ST could improve their decision-making abilities and provide social validation to foster decision-making resilience. For example, many participants felt that ST would be extremely useful in ticketing systems, where the SOC analyst is tasked with a binary classification deciding if the threat should be escalated or not. Current AI systems have a high rate of false positive alerting a security threat when there is none. This can be stressful for new analysts, as "newcomers [who] always escalates things because they are afraid" (P22-NS). Others pointed out ST can provide insights into organizational practices in the context of compliance regulations. Participants also highlighted ST's potential to augment a standardized AI with local contexts in different parts of an organization. This would be particularly useful when the AI was trained on a dataset from the Global North but deployed in the Global South:

> A lot of the companies operate internationally, right? So one of the things we struggle with is working with international clients whose laws are different. On top of that, the system is trained in North American data. Cyber threats mean different things to different people—what's harmless to me can breach your system. So yeah, if we can do something like this to augment the AI, I think we can catch threats better in a personalized manner to the client. Also, justifying things would be easier because now you have data from both sides [humans and AI]. (P26-NS)

Most participants highlighted how visibility of the crew knowledge would be instrumental to pass on "client specific legacy knowledge" (P2-NS). In fact, many cybersecurity teams have existing tools to track past decisions "beyond the model details" (P26-NS); for instance, one team manually keep a historical timeline of false positive alerts. This knowledge helps them calibrate their decisions because "no clients like the boy who shouts wolf every single time" (P25-NS). However, none of these aspects were integrated. Some even expressed surprise on the similarity after going through our scenario. This participant commented on how integration of their tracking system could facilitate ST to improve decision-making:

> It's not like we don't have crew knowledge now, you know. But I never really thought about the whole explainability thing from both sides before you showed me this [pointing to the comments in the sales scenario]. Why just have it from the machine? People are black boxes too, you know. Coming at [explainability] from both ends is kind of holistic. I like it. (P29-NS, a SOC analyst)

Some participants, mainly data scientists, speculated on how one can use the "corpus of social signals" (P10-NS) to feed back into the machine as training data. They wanted to "incorporate the human elements into the machine" (P14-NS) or expert knowledge of "top" analysts back into the AI. They wished the ingestion to not only improve the AI performance, but also to generate socially-situated "holistic explanations", a point we come back to later in the Discussions section.

5.6.2 Healthcare (Radiology and Oncology): Our participants in the healthcare space mainly work in the imaging decision-support domain. Participants felt that ST has promising transfer potential because it would facilitate peer-review and cross-training opportunities. For radiologists and oncologists, participants highlighted that doctors need "explainability especially when their mental models do not match with the AI[s' recommendation]" (P16-NS). This is where peer feedback and review for similar AI recommendations can be instrumental. One person shared a story of how oncologists rely on "tumor boards" (a meeting made up of specialized doctors to discuss challenging cases). The goal is to decide on the best possible treatment plan for a patient by collaboratively thinking through similar tough cases. This participant equated the tumor board activity to those in the comments, highlighting how the 4W adds a "personal touch" to situate the information amongst "trustworthy peers" (P6-NS).

Participants also valued the context brought in by ST for multi-stakeholder problems, such as deciding on treatment plans for patients going through therapy. ST can help ensure the plans are personalized because doctors can not only see the AI's recommendation that's trained on a standard dataset, but can also "consult or reach out to other doctors [who have] treated similar patients and what were all the surrounding contexts that dictated the treatment plan" (P6-NS). According to them, one of the strengths of ST was that the technical and the socio-organizational layers of decision support were integrated in one place, presented side by side, as highlighted in the following quote:

> You need both [social and technical aspects] integrated. Without integration in one place, context switching just takes a lot of time and no one would use it. Just having these things in one place makes all the difference. It's funny how we actually IM each other to ask what people did with the AI's alerts. (P27-NS)

5.7 Challenges around ST

Providing ST in AI systems is not without its challenges, risks, and tensions. We discuss four themes that emerged from the interviews on the potential negative consequences of ST. Future work should strive to mitigate these problems.

First and foremost, there is a vital tension between transparency and **privacy**. Similar issues have been discussed in prior work on social transparency in CSCW [28]. Participants were concerned about making themselves visible to others in the organization, especially with job-critical information such as past performance and competing intellect. Some were also worried that individuals could be coerced into sharing such sensitive information. For example, P6-NS commented, based on her experience working with health professionals, that people may be unwilling to disclose detailed information about their work:

> I've definitely got on the phone with colleges where they're like, well, you know, not everybody at my practices [are willing to talk about it]... just everybody having access to the outcome probably is not great. Especially if they're not really in a position...and it just becomes like, a point of contention. (P6-NS)

Some were concerned about revealing personal information. For example, P2-NS reacted by asking: "do I really want this info about me? Who will see it? What can they do with it?" Several suggested to anonymize the *Who* by revealing only general profiles such as position or present the ST information at an aggregated level.

The second tension is around **biases** that ST could induce on decision-making. The most prominent concern is on group-thinking, by conforming to the group or the majority' choices. Other biases could also happen by following eminent individuals such as someone in a "senior position" (P17-NS) or "a friend" (P14-NS). As discussed in previous sections, ST could invoke social-based heuristics, which could support both decision-making and judgment of AI. However, biases and cognitive heuristics are inevitably coupled, and should be carefully managed. Users in some domains might be more subject to biases from ST than others. For example, P17-NS were hesitant about introducing ST features into the human resource domain, for example for AI assisted hiring: "issues of bias, cherry picking, and groupthink are much more consequential in HR situation" (P17-NS).

A third challenge is regarding **information overload and consumption** of ST. While the design scenario listed only 3 comments, participants were concerned about how to effectively consume the information if the number of entries increases, and how to locate the most relevant information in them. There was also a tension in integrating ST in one's decision-making workflow, which was especially prominent in time-sensitive contexts such as clinical decision-support. Some participants suggested avoiding a constant flow of ST and only provide ST where needed, e.g. "[ST] is not the information that always needs to be up in front of their face, but there should be a way to get back to it, especially when you're building confidence in kind of assistance"(P6-NS). Others suggested providing ST in a structured or processed format such as "summarization" (P17-NS) or "providing some statistics" (P5-S).

Lastly, a tension for the success of ST lies in the **incentive to contribute**. While there are clear benefits for consumers of ST, it is questionable whether there is enough motivation for people to take the extra effort to contribute, as illustrated by this quote:

> One thing that we found interesting is oncology... It's a really hard sell to get them to give feedback into a system because they're so time pressed for their workflow...they're giving you work for free. Like, systems should be doing this for them... but the system can break if they don't participate in that loop. (P6-NS)

This is a classic problem in CSCW systems [33], which may require both lowering the barriers and cost to contribute, and incentivizing contributions with visible and justifiable benefits.

6 DISCUSSION & IMPLICATIONS

Our results identify the potential effects of ST in AI systems, provide design insights to facilitate ST, and point to potential areas of challenges. In this section, we discuss three high-level implications of introducing ST into AI systems: how ST could enable holistic explainability, how ST could strengthen the Human-AI assemblage, and some technical considerations for realizing ST to move towards a socially-situated XAI paradigm.

6.1 Holistic Explainability through ST

After participants concluded the scenario walk-through, we debriefed them on the concept of ST and the idea of facilitating explainability of AI-mediated decisions with ST. Despite frequently using AI systems and facing explainability related issues in their daily workflows, many participants were initially surprised by the association between socio-organizational contexts and AI explainability. The surprise was met with a pleasant realization when they reflected on the scenario and how it could transfer to their real-world use cases. Perhaps their reaction is not surprising given that the epistemic canvas of XAI has largely been circumscribed around the bounds of the algorithm. The focus has primarily been on "the AI in X-AI instead of the X [eXplainable], which is a shame because it's the human who matters" (P25-NS). The following sentiment captures this point:

> I was taught to think [that] all that mattered [in XAI] was explanations from the model...This is actually the

> first time I thought of AI explainability from a social perspective, and I am an expert in this space! This goes to show you how much tunnel-visioned we have been. Once you showed me Social transparency…it was clear that organizational signals can definitely help us make sense of the overall system. It's like we had blinders or something that stopped us from seeing the larger picture. (P27-NS, a SOC data scientist)

A most common way participants expressed how ST impacted their "ways of answering the why-question" (P28-NS) was how incorporation of the context makes the explainability more "holistic" (P2-NS, P23-S, P6-NS, P27-NS, P29-NS). Acknowledging that "context is king for explanations… [and] there are many ways to answer why" (P27-NS), participants felt that the ST goes "beyond the AI" to provide "peripheral vision" of the organizational context. This, in turn, allows them to answer their why-questions in a holistic manner. For instance, in SOC situations "there is often no single correct answer. There are multiple correct answers" (P2-NS). Since AI systems "don't produce multitudes of of explanations", participants acknowledged that "incorporating the social layer into the mix" can expand the ways they view explainability (P28-NS). Moreover, the humanization of the process can also make the decision explainable to non-primary stakeholders in a way that technical transparency alone cannot achieve. For instance, participants felt that having the 4W can make it easier to justify the decisions to clients and regulators.

While in this work our focus is on how a holistic explainability through ST could better support decision-makers, we recognize that there are other types of stakeholders and explanation consumers, as well as other types of AI systems, that could benefit from ST. For example, collecting the 4W in the deployment context could help model developers to investigate how the system performs and why it fails, then incorporate the insights gained about the technological, decision and organizational contexts to improve the model. Auditors or regulatory bodies could also leverage 4W information to better assess the model's performance, biases, safety, etc. by understanding its situated impact. The contributors of ST information are not limited to decision-makers. For example, with automated AI systems where there isn't a human decision-maker involved, its explainability could be facilitated by making visible the social contexts of people who are impacted by the AI systems.

6.2 Making the Human-AI assemblage concrete

We highlight that a consequential AI system is often situated in complex socio-organizational contexts, where many people interacting with it. By bringing the human elements of decision-making to the fore-front, ST enables the humans to be explicitly represented, thereby making the Human-AI assemblage concrete. As one participant put it, the socially-situated context can ensure "the human is not forgotten in the mix of things" (P25-NS). In our Findings section, we discussed how organizational meta-knowledge can facilitate formation of Transactive Memory Systems (TMS) [42, 74, 102], allowing "who knows what" to be explicitly encoded for future retrieval. Over time, the "heedful interrelations" enabled by TMS and repeatedly seeing others' decision processes with the AI through ST could possibly enable a shared decision schema in the community, leading to the formation of a collective mind [103, 110]. This collective mind is one that includes AI as a critical player. With this conceptualization future work could explore the collective actions and evolution of human-AI assemblages.

By prioritizing the view of human-AI assemblage over the AI, adding ST to AI systems calls for critical consideration on what information of the humans and whose information is made visible to whom. Prior work on ST in CSCW systems has warned against developing technologies that make it easy to share information without careful consideration on its longer-term second-order effect on the organization and its members [94]. In Section 5.7 we identified four potential issues of ST as foreseen by our participants, including privacy, biases, information overload and motivation to contribute, all of which could have profound impact on the functioning of the human-AI assemblage. In general, future work implementing ST in AI systems should take socioechnical approaches to developing solutions that are sensitive to the values of stakeholders and "localized" to a human-AI assemblage. For example, regarding the privacy issue, participants were sensitive to how much visibility the rest of the organization has to their shared activities and knowledge. When asked how they might envision the boundaries, participants highlighted that one should "let the individual teams decide because every 'tribe' is different" (P28-NS).

6.3 Towards socially-situated XAI

Using a scenario-based design (SBD) method, we suspended the needs to define system operations and technical details. Some practical challenges may arise in how to present the 4W to explanation seekers. The first challenge is to handle the quantity of information, especially to fit into the workflow of the users. In addition to utilizing NLP techniques to make the content more consumable, for example by providing memorization or organizing it into facets, it could also be beneficial to give users filtering options, allow them to define "similarity" or choose past examples they want to see. Secondly, there needs to be mechanisms in place to validate the quality and applicability of ST information, since "not all whys are created equal" (P25-NS). This could be achieved by either applying quality control on the recorded ST information, or through careful design of interfaces to elicit high-quality 4W information from the contributors. Another caveat is that it is common for an AI system to receive model updates or adapt with usage, so its decisions may change over time. In that case, it is necessary to flag the differences of the AI in showing past ST information. Lastly, in certain domains or organizations, it is not advisable or possible to gather all 4W information, sometimes due to the tension with privacy, biases and motivation to contribute, so alternative solutions should be sought, for example by linking past decision trajectories with relevant guidelines or documentation to help users decode the *why* information when it is not directly available.

Several participants, especially those with a data science background, suggested an interesting area for technical innovation–if "the AI can ingest the social data" (P12-NS) to improve both its performance and its explanations. While recent work has started exploring teaching AI with human rationales [29], ST could enable acquiring such rationales in real-usage contexts. As suggested by what participants learned from the 4W, the decision and organizational contexts made visible by ST could help the AI to learn additional features and localized rules and constraints, then incorporate them into its future decisions. Moreover, a notable area of XAI work focuses on generating human-consumable and domain-specific machine explanations by learning from how humans explain [27, 40], which could be a fruitful area to explore when combined with the availability of 4W information. That being said, it may be desirable to explicitly separate the technical component (to show how the AI arrives at its decision) and the socio-organizational component (as further support for or caution against AI's decision) in the explanations, as participants had strong opinions to be able to "know how and where to place the trust" (P27-NS).

7 LIMITATIONS & FUTURE WORK

We view our work as the beginning of a broader cross-disciplinary discourse around what explainability entails in AI systems. With this paper, we have taken a formative step by exploring the concept of Social Transparency (ST) in AI systems, particularly focusing on how incorporation of socio-organizational context can impact explainability of the human-AI assemblage. Given this first step, the insights from our work should be viewed accordingly. We acknowledge the limitations that come with using a scenario-based design [84], including the dependency between the scenario and data. The insights should be interpreted as formative instead of evaluative. We acknowledge that we need to do more work in the future to expand the design space and consider other design elements for ST, further unpack the transferability of our insights, especially where this transfer might be inappropriate. We should also investigate how ST impacts user trust over longitudinal use of ST-infused XAI systems.

Our conception of ST is rooted in and inspired by Phil Agre's notion of Critical Technical Practice [4, 5] where we identify the dominant assumptions of XAI and critically question the status quo to generate alternative technology that brings previously-marginalized insights into the center. Agre stated that "at least for the foreseeable future, [a CTP-inspired concept] 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." [4]. As such, ST will, at least for the foreseeable future be a work-in-progress, one that is continuously pushing the boundaries of design and reflexively working on its own blind spots. We have "planted one foot" in the work of design by identifying a neglected insight–the lack of incorporation of socio-organizational context as a constitutive design element in XAI– and exploring the design of ST-infused XAI systems. 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.

8 CONCLUSION

Situating XAI through the lens of a Critical Technical Practice, this work is our attempt to challenge algorithm-centered approaches and the dominant narrative in the field of XAI. Explainability of AI systems inevitably sits at the intersection of technologies and people, both of which are socially-situated. Therefore, an epistemic blind spot in that neglects the "socio" half of sociotechnical systems would likely render technological solutions ineffective and potentially harmful. This is particularly problematic as AI technologies enter different socio-organizational contexts for consequential decision-making tasks. Our work is both conceptual and practical. Conceptually, we address the epistemic blind spot by introducing and exploring Social Transparency (ST)–the incorporation of socio-organizational context–to enable holistic explainability of AI-mediated decision-making. Practically, we progressively develop the concept and design space of ST through design and empirical research. Specifically, we developed a scenario-based design that embodies the concept of ST in an AI system with four constitutive elements–*who* did *what* with the AI system, *when*, and *why* they did what they did (4W). Using this scenario-based design, we explored the potential effect of ST and design implications through 29 interviews with AI stakeholders. The results refined our conceptual development of ST by discerning three levels of context made visible by ST and their effects: technological, decision, and organizational. Our work also contributes concrete design insights and point to potential challenges of incorporating socio-organizational context into AI systems, with which practitioners and researchers can further explore the design space of ST. By adding formative insights that catalyzes our journey towards a socially-situated XAI paradigm, this work contributes to the discourse of human-centered XAI by expanding the conceptual and design space of XAI.

ACKNOWLEDGMENTS

With our deepest gratitude, we acknowledge the time our participants generously invested in this project. We are grateful to members of the Human-Centered AI Lab at Georgia Tech whose continued input refined the conceptualizations presented here. We are indebted to Werner Geyer, Michael Hind, Stephanie Houde, David Piorkowski, John Richards, and Yunfeng Zhang from IBM Research AI for their generous feedback and time throughout the duration of this project. Special thanks to Intekhab Hossain and Samir Passi for conversations and feedback throughout the years that have constructively added to the notion of Social Transparency. This project was partially supported through an internship at IBM Research AI and by the National Science Foundation under Grant No. 1928586.

REFERENCES

[1] 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*. 1–18.

[2] Ashraf Abdul, Christian von der Weth, Mohan Kankanhalli, and Brian Y Lim. 2020. COGAM: Measuring and Moderating Cognitive Load in Machine Learning Model Explanations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[3] Mark S Ackerman, Juri Dachtera, Volkmar Pipek, and Volker Wulf. 2013. Sharing knowledge and expertise: The CSCW view of knowledge management. *Computer Supported Cooperative Work (CSCW)* 22, 4-6 (2013), 531–573.

[4] P Agre. 1997. Toward a critical technical practice: Lessons learned in trying to reform AI in Bowker. *G., Star, S., Turner, W., and Gasser, L., eds, Social Science, Technical Systems and Cooperative Work: Beyond the Great Divide, Erlbaum* (1997).

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

[6] Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze. 2020. Evaluating saliency map explanations for convolutional neural

networks: a user study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 275–285.

[7] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115.

[8] 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).

[9] John R Austin. 2003. Transactive memory in organizational groups: the effects of content, consensus, specialization, and accuracy on group performance. *Journal of applied psychology* 88, 5 (2003), 866.

[10] David P Brandon and Andrea B Hollingshead. 2004. Transactive memory systems in organizations: Matching tasks, expertise, and people. *Organization science* 15, 6 (2004), 633–644.

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

[12] 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*. 454–464.

[13] 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*. 258–262.

[14] 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.

[15] 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*. 1–12.

[16] 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.

[17] 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*. 388–397.

[18] Laura Dabbish, Colleen Stuart, Jason Tsay, and Jim Herbsleb. 2012. Social coding in GitHub: transparency and collaboration in an open software repository. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*. 1277–1286.

[19] Ben Dattner, Tomas Chamorro-Premuzic, Richard Buchband, and Lucinda Schettler. 2019. The Legal and Ethical Implications of Using AI in Hiring. *Harvard Business Review* (25 April 2019). Retrieved 26-August-2019 from https://hbr.org/2019/04/the-legal-and-ethical-implications-of-using-ai-in-hiring

[20] Daniel Clement Dennett. 1989. *The intentional stance*. MIT press.

[21] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.

[22] 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*. 275–285.

[23] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. *stat* 1050 (2017), 2.

[24] Paul Dourish. 2004. *Where the action is: the foundations of embodied interaction*. MIT press.

[25] Paul Dourish, Janet Finlay, Phoebe Sengers, and Peter Wright. 2004. Reflective HCI: Towards a critical technical practice. In *CHI'04 extended abstracts on Human factors in computing systems*. 1727–1728.

[26] Upol Ehsan and Mark O Riedl. 2020. Human-centered Explainable AI: Towards a Reflective Sociotechnical Approach. *arXiv preprint arXiv:2002.01092* (2020).

[27] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O Riedl. 2019. Automated rationale generation: a technique for explainable AI and its effects on human perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 263–274.

[28] Thomas Erickson and Wendy A Kellogg. 2003. Social translucence: using minimalist visualisations of social activity to support collective interaction. In *Designing information spaces: The social navigation approach*. Springer, 17–41.

[29] 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* CSCW (2021).

[30] Eric Gilbert. 2012. Designing social translucence over social networks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2731–2740.

[31] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 80–89.

[32] 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.

[33] Jonathan Grudin. 1988. Why CSCW applications fail: problems in the design and evaluationof organizational interfaces. In *Proceedings of the 1988 ACM conference on Computer-supported cooperative work*. 85–93.

[34] David Gunning. 2017. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web* 2 (2017), 2.

[35] Carl Gutwin and Saul Greenberg. 2002. A descriptive framework of workspace awareness for real-time groupware. *Computer supported cooperative work* 11, 3-4 (2002), 411–446.

[36] Carl Gutwin, Reagan Penner, and Kevin Schneider. 2004. Group awareness in distributed software development. In *Proceedings of the 2004 ACM conference on Computer supported cooperative work*. 72–81.

[37] 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/

[38] Fritz Heider. 1958. The psychology of interpersonal relations Wiley. *New York* (1958).

[39] Denis J Hilton. 1996. Mental models and causal explanation: Judgements of probable cause and explanatory relevance. *Thinking & Reasoning* 2, 4 (1996), 273–308.

[40] Michael Hind, Dennis Wei, Murray Campbell, Noel CF Codella, Amit Dhurandhar, Aleksandra Mojsilović, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2019. TED: Teaching AI to explain its decisions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 123–129.

[41] Robert R Hoffman and Gary Klein. 2017. Explaining explanation, part 1: theoretical foundations. *IEEE Intelligent Systems* 32, 3 (2017), 68–73.

[42] Andrea B Hollingshead and David P Brandon. 2003. Potential benefits of communication in transactive memory systems. *Human communication research* 29, 4 (2003), 607–615.

[43] Sungsoo Ray Hong, Jessica Hullman, and Enrico Bertini. 2020. Human Factors in Model Interpretability: Industry Practices, Challenges, and Needs. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–26.

[44] Shih-Wen Huang and Wai-Tat Fu. 2013. Don't hide in the crowd! Increasing social transparency between peer workers improves crowdsourcing outcomes. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 621–630.

[45] David Hume. 2000. *An enquiry concerning human understanding: A critical edition*. Vol. 3. Oxford University Press.

[46] Edwin Hutchins. 1991. The social organization of distributed cognition. (1991).

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

[48] Daniel Kahneman, Stewart Paul Slovic, Paul Slovic, and Amos Tversky. 1982. *Judgment under uncertainty: Heuristics and biases*. Cambridge university press.

[49] Michael Katell, Meg Young, Dharma Dailey, Bernease Herman, Vivian Guetler, Aaron Tam, Corinne Bintz, Daniella Raz, and PM Krafft. 2020. Toward situated interventions for algorithmic equity: lessons from the field. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 45–55.

[50] 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*. 1–14.

[51] Jennifer G Kim, Ha-Kyung Kong, Hwajung Hong, and Karrie Karahalios. 2020. Enriched Social Translucence in Medical Crowdfunding. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. 1465–1477.

[52] Roderick M Kramer. 1999. Trust and distrust in organizations: Emerging perspectives, enduring questions. *Annual review of psychology* 50, 1 (1999), 569–598.

[53] 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*. 1–13.

[54] 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.

[55] Paul M Leonardi. 2014. Social media, knowledge sharing, and innovation: Toward a theory of communication visibility. *Information systems research* 25, 4 (2014), 796–816.

[56] Paul M Leonardi. 2015. Ambient awareness and knowledge acquisition: using social media to learn 'who knows what'and 'who knows whom'. *Mis Quarterly* 39, 4 (2015), 747–762.

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

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

[59] 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.

[60] 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*. 2119–2128.

[61] Brian Y Lim, Qian Yang, Ashraf M Abdul, and Danding Wang. 2019. Why these Explanations? Selecting Intelligibility Types for Explanation Goals.. In *IUI Workshops*.

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

[63] 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* 155, 2 (02 2020), 148–158. https://doi.org/10.1001/jamasurg.2019.4917 arXiv:https://jamanetwork.com/journals/jamasurgery/articlepdf/2756311/jamasurgery_loftus_2019_rv_190007.pdf

[64] Tania Lombrozo. 2011. The instrumental value of explanations. *Philosophy Compass* 6, 8 (2011), 539–551.

[65] Tania Lombrozo. 2012. Explanation and abductive inference. (2012).

[66] Erin E Makarius, Debmalya Mukherjee, Joseph D Fox, and Alexa K Fox. 2020. Rising with the machines: A sociotechnical framework for bringing artificial intelligence into the organization. *Journal of Business Research* 120 (2020), 262–273.

[67] David W McDonald, Stephanie Gokhman, and Mark Zachry. 2012. Building for social translucence: a domain analysis and prototype system. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*. 637–646.

[68] Miriam J Metzger and Andrew J Flanagin. 2013. Credibility and trust of information in online environments: The use of cognitive heuristics. *Journal of pragmatics* 59 (2013), 210–220.

[69] Miriam J Metzger, Andrew J Flanagin, and Ryan B Medders. 2010. Social and heuristic approaches to credibility evaluation online. *Journal of communication* 60, 3 (2010), 413–439.

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

[71] Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining explanations in AI. In *Proceedings of the conference on fairness, accountability, and transparency*. 279–288.

[72] Shakir Mohamed, Marie-Therese Png, and William Isaac. 2020. Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. *Philosophy & Technology* (2020), 1–26.

[73] 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.

[74] Richard L Moreland and L Thompson. 2006. Transactive memory: Learning who knows what in work groups and organizations. *Small groups: Key readings* 327 (2006).

[75] Michael Muller and Q Vera Liao. [n.d.]. Exploring AI Ethics and Values through Participatory Design Fictions. ([n. d.]).

[76] 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

[77] Bonnie A Nardi, Steve Whittaker, and Heinrich Schwarz. 2002. NetWORKers and their activity in intensional networks. *Computer Supported Cooperative Work (CSCW)* 11, 1-2 (2002), 205–242.

[78] Duyen T Nguyen, Laura A Dabbish, and Sara Kiesler. 2015. The perverse effects of social transparency on online advice taking. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. 207–217.

[79] DJ Pangburn. 2019. Schools are using software to help pick who gets in. What could go wrong? *Fast Company* (17 May 2019). Retrieved 16-September-2020 from https://www.fastcompany.com/90342596/schools-are-quietly-turning-to-ai-to-help-pick-who-gets-in-what-could-go-wrong

[80] 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).

[81] 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*. 1–13.

[82] 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, 19–36.

[83] Paul Resnick, Ko Kuwabara, Richard Zeckhauser, and Eric Friedman. 2000. Reputation systems. *Commun. ACM* 43, 12 (2000), 45–48.

[84] Mary Beth Rosson and John M Carroll. 2009. Scenario based design. *Human-computer interaction. boca raton, FL* (2009), 145–162.

[85] Selma Šabanović. 2010. Robots in society, society in robots. *International Journal of Social Robotics* 2, 4 (2010), 439–450.

[86] 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.

[87] 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*. 59–68.

[88] 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.

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

[90] 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.

[91] Thilo Spinner, Udo Schlegel, Hanna Schäfer, and Mennatallah El-Assady. 2019. explAIner: A visual analytics framework for interactive and explainable machine learning. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 1064–1074.

[92] Susan Leigh Star and Anselm Strauss. 1999. Layers of silence, arenas of voice: The ecology of visible and invisible work. *Computer supported cooperative work (CSCW)* 8, 1-2 (1999), 9–30.

[93] Anselm Strauss and Juliet Corbin. 1994. Grounded theory methodology. *Handbook of qualitative research* 17, 1 (1994), 273–285.

[94] H Colleen Stuart, Laura Dabbish, Sara Kiesler, Peter Kinnaird, and Ruogu Kang. 2012. Social transparency in networked information exchange: a theoretical framework. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. 451–460.

[95] 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*.

[96] Lucy Suchman. 1995. Making work visible. *Commun. ACM* 38, 9 (1995), 56–64.

[97] Lucy A Suchman. 1987. *Plans and situated actions: The problem of human-machine communication.* Cambridge university press.

[98] S Shyam Sundar. 2008. *The MAIN model: A heuristic approach to understanding technology effects on credibility.* MacArthur Foundation Digital Media and Learning Initiative.

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

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

[101] 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*. 1–15.

[102] Daniel M Wegner, Ralph Erber, and Paula Raymond. 1991. Transactive memory in close relationships. *Journal of personality and social psychology* 61, 6 (1991), 923.

[103] Karl E Weick and Karlene H Roberts. 1993. Collective mind in organizations: Heedful interrelating on flight decks. *Administrative science quarterly* (1993), 357–381.

[104] Daniel S Weld and Gagan Bansal. 2019. The challenge of crafting intelligible intelligence. *Commun. ACM* 62, 6 (2019), 70–79.

[105] 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.

[106] 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.

[107] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L Arendt. 2020. How do visual explanations foster end users' appropriate trust in machine learning?. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 189–201.

[108] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. Unremarkable ai: Fitting intelligent decision support into critical, clinical decision-making processes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–11.

[109] 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 *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 4477–4488.

[110] Youngjin Yoo and Prasert Kanawattanachai. 2001. Developments of transactive memory systems and collective mind in virtual teams. *International Journal of Organizational Analysis* 9, 2 (2001), 187–208.

[111] 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.

[112] 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.