Sensible AI: Re-imagining Interpretability and Explainability using Sensemaking Theory

Harmanpreet Kaur
harmank@umich.edu
University of Michigan
Ann Arbor, MI, USA

Eytan Adar
eadar@umich.edu
University of Michigan
Ann Arbor, MI, USA

Eric Gilbert
eegg@umich.edu
University of Michigan
Ann Arbor, MI, USA

Cliff Lampe
cacl@umich.edu
University of Michigan
Ann Arbor, MI, USA

ABSTRACT

Understanding how ML models work is a prerequisite for responsibly designing, deploying, and using ML-based systems. With interpretability approaches, ML can now offer explanations for its outputs to aid human understanding. Though these approaches rely on guidelines for how humans explain things to each other, they ultimately solve for improving the artifact—an explanation. In this paper, we propose an alternate framework for interpretability grounded in Weick's sensemaking theory, which focuses on *who* the explanation is intended for. Recent work has advocated for the importance of understanding stakeholders' needs—we build on this by providing concrete properties (e.g., identity, social context, environmental cues, etc.) that shape human understanding. We use an application of sensemaking in organizations as a template for discussing design guidelines for *sensible AI*, AI that factors in the nuances of human cognition when trying to explain itself.

CCS CONCEPTS

• **Human-centered computing** → *HCI theory, concepts and models*; • **Computing methodologies** → *Artificial intelligence*; *Machine learning*.

KEYWORDS

interpretability, explainability, sensemaking, organizations

ACM Reference Format:
Harmanpreet Kaur, Eytan Adar, Eric Gilbert, and Cliff Lampe. 2022. Sensible AI: Re-imagining Interpretability and Explainability using Sensemaking Theory. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3531146.3533135

1 INTRODUCTION

With ML-based systems being deployed in the wild, it's imperative that all stakeholders of these systems have some understanding of how the underlying ML model works. From the experts who develop algorithms to practitioners who design and deploy ML-based systems, and end-users who ultimately interact with these systems—stakeholders require varying levels of understanding of ML to ensure that these systems are used responsibly. Approaches like interpretability and explainability have been proposed as a way to bridge the gap between ML models and human understanding. These include models that are inherently interpretable (e.g., decision trees [89], simple point systems [50, 124] or generalized additive models [18, 37]) and post-hoc explanations for the predictions made by complex models (e.g., LIME [92], SHAP [69]). Tools that implement interpretability and explainability approaches have also been made available for public use. In light of this, recent work in HCI has evaluated the efficacy of these tools in helping people understand ML models. These findings suggest that ML practitioners [52] and end-users [10, 54] are not always able to make accurate judgments about the model, even with the help of explanations. In fact, having access to these tools often leads to over-trust in the ML models. Ultimately, noting that interpretability and explainability are meant for the stakeholders, recent work has proposed design guidelines for explanations based on research in the social sciences about how people explain things to each other [77, 79]. Taking a human-centered or a model-centered approach, this prior work seeks to answer: *what are the characteristics of an explanation that can help people understand ML models?*

Let us consider a real-world setting. Imagine you are a doctor in a healthcare organization that has decided to use an ML-based decision-support software to help with medical diagnosis. The system takes as input information about patients' symptoms, demographics, family history, etc., and returns a predicted diagnosis. Naturally, you want to be able to overview why the software predicted a certain diagnosis before you suggest treatment based on its prediction. Further, you want to be able to explain to the patient why you (did not) trust and follow the predicted diagnosis. To aid with this, the software provider gives you access to an explanation system (e.g., LIME [92], SHAP [69]) which shows: (1) a local explanation (e.g., a bar chart) of the input features that were most important for the diagnosis made for a specific patient, (2) a global explanation for the features that are usually important to the model

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06…$15.00
https://doi.org/10.1145/3531146.3533135

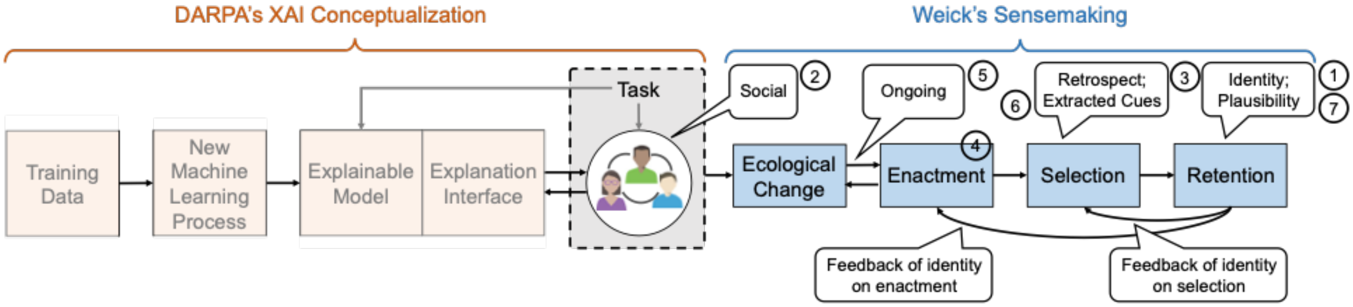


Figure 1: Left: DARPA's conceptualization of Explainable AI, adapted from [35]. Right: Weick's sensemaking properties (1–7) categorized using the high-level Enactment-Selection-Retention organizational model, adapted from [49]. Enactment includes properties about perceiving and acting on environmental changes; Selection, properties related to interpreting what the changes mean; and Retention, properties that describe storing and using prior experiences [56]. Our Sensible AI framework extends the existing definition of interpretability and explainability to include Weick's sensemaking properties.

when making a prediction, and (3) an overview of each feature's relationship with the output classes. The explanation system also includes interactive elements so you can ask "what if" questions based on different combinations of input features.

Is this enough to ensure that the ML-based decision-support software can be reliably used by the doctor? We claim that the answer to this question is no. This paper makes the argument that current interpretability and explainability solutions will always fall short of helping people reliably use ML-based systems for decision-making because of their focus on designing better explanations—in other words, improving an artifact. For example, while the explanation shows the symptoms that were important to the model's prediction (i.e., a local explanation), it does not tell the doctor to be cautious that the patient's other symptoms are fluctuating, that the patient belongs to a sub-group for which the model has limited training data, or that the nurses have noticed other relevant symptoms in the visiting family. From the patient's perspective, the explanation does not convey why, for example, their fear of having a particular disease (after an online symptom search or from family history) is unwarranted in this instance. These factors, that have little to do with the particular explanation, can alter the stakeholders' decision-making in significant ways. Here, *we propose a specific theoretical framework to shift from improving the artifact (e.g., an explanation or explanation system) to understanding how humans make sense of complex, and sometimes conflicting, information.* Recent work supports this shift from *what* an explanation should look like to *who* it is intended for. Properties of the *who* such as, prior experience with AI and ML [27], attitude towards AI (e.g., algorithmic aversion [17, 21]), the socio-organizational context [26], have been observed as being critical to understanding AI and ML outputs. We extend this work by providing a framework for *how* to incorporate human-centered principles to interpretability and explainability.

In this paper, we present Weick's sensemaking as a framework for envisioning the needs of people in the human-machine context. Weick describes sensemaking as, quite literally, "the making of sense," or "a developing set of ideas with explanatory possibilities" [118]. Although Weick's definition is similar to that of prior work in HCI and information retrieval, the two deviate in their goals; the latter defines sensemaking as finding representations that enable information foraging and question-answering [85, 96]. Weick's sensemaking is more procedural: "placement of items into frameworks, comprehending, redressing surprise, constructing meaning, interacting in pursuit of mutual understanding, and patterning" [118, p.6]. These processes are influenced by one's identity, environment, social, and organizational context—Weick expands these into the seven properties of sensemaking (Figure 1, Right). For example, for the doctor trying to diagnose a patient with the help of an ML-based system (with explanations), their understanding of the predicted diagnosis can be influenced by questions such as, have they recently diagnosed another patient with similar symptoms; is the patient's care team in agreement on a diagnosis; is the predicted diagnosis plausible; and, which symptoms are more visible and does the explanation present these as important to the prediction. The seven properties of sensemaking are a framework for identifying and understanding these contextual factors.

What does this knowledge of sensemaking offer to interpretability and explainability researchers and tool designers? A sensemaking perspective tells us how things beyond the individual (i.e., the environmental, social, and organizational contexts) shape individual cognition. It gives us a path forward. Prior work in organizational studies has used sensemaking to identify ways in which teams and organizations can be made more reliable. These high-reliability organizations (HROs) can serve as a template for designing *Sensible AI*, AI that accounts for the nuances of human cognition when explaining itself to people. We extend the principles that make HROs reliable (e.g., a preoccupation with failure, a sensitivity to low-level operations, a reluctance to simplify anomalous situations) as guidelines for designing Sensible AI. Within our healthcare example, Sensible AI might take the form of a system that highlights the most significant ways in which a change in input features would change the predicted diagnosis; shows cases with similar input features but different diagnosis; presents input features that were considered less important by the model; asks all members of the patient care team to review the diagnosis individually first, allowing for a diversity of opinions and discussion opportunities; and asks for further explanation for cases in which the predicted diagnosis was disregarded, to inform future test cases. Our hope is that researchers and designers can translate our Sensible AI design guidelines as technical and social checks and balances in their tools, to better support human cognition as described by sensemaking.

2 INTERPRETABILITY AND EXPLAINABILITY

2.1 What are interpretability and explainability?

Interpretability is defined from a model's perspective as the "ability to explain or to present in understandable terms to a human" [24, p.2]. It serves as a proxy for other desiderata for ML-based systems such as reliability, robustness, transferability, informativeness, etc. These properties in turn promote trustworthiness, accountability, and fair and ethical decision-making [24, 65]. At a high-level, interpretability approaches can be categorized into glassbox models (e.g., [18, 37, 50, 58, 89, 124]) or post-hoc explanations for blackbox models (e.g., [6, 69, 92, 99, 102]). Instantiating these approaches into user-facing tools, static explanations output by mathematical representations of interpretability now includes interactive visuals output by explainable AI. Although similarly defined, this idea of *explainability* is more human-centered and is "associated with the notion of an explanation as an interface between humans and a decision maker that is, at the same time, both an accurate proxy of the decision maker and comprehensible to human" [9, p.85] (Figure 1, Left). Scholars have incorporated prior work from philosophy (e.g., [34, 39, 64, 84, 86, 111]), the social sciences (e.g., [40, 60, 66–68, 72, 77, 79, 81, 103]), and HCI (e.g., [12, 25, 32, 82, 83, 121, 126]) with the motivation that by translating ideas from how people explain things to each other, we can design better solutions for how ML models can be explained to people. As a result, increasingly, interpretability and explainability tools include characteristics such as interactivity [7, 41], counterfactual "what-if" outputs [78, 115], and modular and sequential explanations [75].

Several comprehensive reviews (e.g., [3, 9, 63, 116, 125]) synthesize and describe design considerations for the field. Based on a review of 289 core papers and 12412 citing papers, Adbul et al. highlight the trends as (1) a move from early AI work (e.g., in Expert Systems [19, 109]) to FAccT-centric ways of providing explanations; and (2) addressing macroscopic societal accountability in addition to helping individual users understand ML outputs [3]. Arrieta et al. taxonomize 409 papers to clarify terminology (e.g., interpretability, understandability, comprehensibility, etc.); describe interpretability approaches for shallow and deep learning models; and highlight the challenges for responsible AI [9].

2.2 Understanding the "who" in interpretability and explainability

Scholars in ML, HCI, and social science communities have advocated for the importance of understanding *who* the explanations are intended for. Their work identifies principles about stakeholders that are relevant in the human-machine context. *Cognitive factors* (e.g., mental models, type of reasoning) have been shown to be important. For example, accurate mental models and deliberative reasoning can help avoid ML practitioners' misuse of, and over-reliance on, interpretability outputs [52]. This also applies to end-users without ML expertise [16], otherwise explanations increase the likelihood that an end-user will accept an AI's output, regardless of its correctness [10]. For end-users, completeness (rather than soundness) of explanations helps people form accurate mental models [57]. Accuracy and example-based explanations can similarly shape people's mental models and expectations, albeit in different ways [54].

Prior experience and background in ML is also important. Variance in these can result in preset expectations, which can lead to over- or under-use of explanations [27]. *Job- and task-dependent information needs* also shape how (much) people internalize explanations. Explanation interfaces that are interactive and collaborative can improve overall accuracy [108]. Additionally, explanations from glassbox models with fewer number of features are easier for end-users to understand [88]. For ML practitioners, specific types of visuals of explanations (e.g., local vs. global, sequential vs. collective) differ in how much they help them understand and debug models, and explain them to customers [41, 75]. Finally, *social, organizational, and socio-organizational context* is important. For example, [42, 45, 70, 113, 126] all highlight the challenges of operating within an organization that either develops or employs an AI-based system. Stakeholders within and outside the organization can have conflicting needs from the system—technical interpretability and explainability approaches are unable to account for these.

These studies from the ML, HCI, and social science communities have all highlighted relevant factors about the "who" in interpretability and explainability. Our proposed framework complements these evaluations: it unifies them based on sensemaking theory translated from organizational studies. We explain how individual, social, and organizational factors can affect the human-machine context, and provide a path forward that accounts for these *who*-centered factors.

3 SENSEMAKING

Sensemaking describes a framework for the factors that influence human understanding; "the sensemaking perspective is a frame of mind about frames of mind" [118, p.xii]. It is most prominent in discrepant or surprising events. People try to put stimuli into frameworks, particularly when predictions or expectations break down. That is, when people come across new or unexpected information, they like to add structure to this unknown. The process by which they do this, why they do it, and how it affects them and their understanding of the world are all central to sensemaking.

Sensemaking subsumes interpretability[1]. They share the same goal: understanding an outcome or experience. If an ML-based system could explain itself, we can verify if the reasoning is sound based on auxiliary criteria (e.g., safety, nondiscrimination), and determine whether the system meets other desiderata such as fairness, reliability, causality, and trust [24, 65]. Sensemaking includes all of this and more. Sensemaking not only considers the information being presented to the person doing the meaning-making, but also

[1] Although interpretability is defined as model-centric and explainability as human-centric, there is not yet consensus on how these terms are different from an implementation point of view. Since "interpretability" is commonly used in describing tools that output explanations, we use this term for the rest of the paper. We follow similar terminology choices with ML- (rather than AI-) based systems since interpretability is attributed to ML models.

Property	Human-Human Context	Human-Machine Context
Identity Construction	Sensemaking is a question about who I am as indicated by the discovery of how and what I think.	Given multiple explanations, people will internalize the one(s) that support their identity in positive ways.
Social	What I say and single out and conclude are determined by who socialized me and how I was socialized, and by the audience I anticipate will audit the conclusions I reach.	Differences in micro- and macro-social contexts affect the effectiveness of explanations.
Retrospective	To learn what I think, I look back over what I said earlier.	Providing explanations before people can reflect on the model and its predictions negatively affects sensemaking.
Enactive	I create the object to be seen and inspected when I say or do something.	The order in which explanations are seen affects how people understand a model and its predictions.
Ongoing	Understanding is spread across time and competes for attention with other ongoing projects, by which time my interests may already have changed.	The valence and magnitude of emotion caused by an interruption during the process of understanding explanations from interpretability tools change what is understood.
Focused on Extracted Cues	The 'what' that I single out and embellish is only a small portion of the original utterance, that becomes salient because of context and personal dispositions.	Highlighting different parts of explanations can lead to varying understanding of the underlying data and model.
Plausibility over Accuracy	I need to know enough about what I think to get on with my projects, but no more, which means sufficiency and plausibility take precedence over accuracy.	Given plausible explanations for a prediction, people are not inclined to search for the accurate one amongst these.

Table 1: An overview of the seven properties of sensemaking, their description in the human-human context, and our proposed claims for the human-machine context grounded in each property.

additional contextual nuances that affect whether and how this information is internalized. This includes factors such as, the enacted environment, the individual's identity, their social and organizational networks, and prior experiences with similar information.

In the subsections that follow, we describe Weick's seven properties of sensemaking in the human-human context and translate them for the human-machine context (see Table 1 for an overview). To concretize how these properties might affect stakeholders of ML-based systems, we present an example user vignette for each property. Prior work has applied similar methodology when translating theory [4, 76]. While the examples are crafted based on popular press articles and research papers, they are not intended as being representative of these cases. We use them to highlight a sensemaking property, but we do not claim that the property has a causal relationship with the example, i.e., there could be other reasons for why the ML-based systems functioned the way that they are described in these articles.

3.1 Grounded in Identity Construction

Identity is critical for AI/ML sensemaking because people only understand these systems in ways that they are congruent with their existing beliefs or update their beliefs while shedding a positive light on them. For interpretability, this suggests that, given multiple explanations, people will internalize the one(s) that support their identity in positive ways.

3.1.1 Identity Construction in the ***Human-Human*** *Context.* Sensemaking begins with the sensemaker. In this way, sensemaking is innately human-centered: "how can *I* know what *I* think until *I* see what *I* say?" [118, p.18]. It is grounded in the individual's need to have a clear sense of identity. People make sense of something to either support their existing beliefs or update them when applying their beliefs leads to a breakdown in their understanding. Weick notes five things of importance for identity and sensemaking [118, pp.23-24]: (1) controlled, intentional sensemaking is triggered by a failure to confirm one's self; (2) sensemaking is grounded in the desire to maintain a consistent, positive self-conception; (3) people learn about their identities by projecting them into an environment—which includes their social, organizational, and cultural contexts—and observing the consequences; (4) sensemaking via identity construction is a mix of proaction and reaction; and (5) sensemaking is self-referential in that the self is what ultimately needs interpreting—what a given situation means is defined by the identity that an individual relies on while understanding it.

The relationship between identity and sensemaking is not limited to the individual sensemaker. The influence of social context can be seen in how identity is constructed. Weick describes this influence using three definitions of identity. First, Mead's claim that the mind and self are developed based on the communicative processes among people (i.e., social behaviorism). Individuals are comprised of "a parliament of selves" which reflect their various social contexts [74]. Second, Knorr-Cetina's inclusion of social contexts based on the larger tapestry of social, organizational, and cultural norms, i.e., the macro-social [53]. Finally, Erez and Earley's three self-derived needs that shape identity, which include intrapersonal and interpersonal dynamics: (1) the need for *self-enhancement*, seeking and maintaining a positive cognitive and affective state about the self; (2) the *self-efficacy* motive, desire to perceive oneself as competent and efficacious; and (3) the need for *self consistency*, desire to sense and experience coherence and continuity [28].

Sensemaking is made challenging by identity because the more identities that an individual has, the more ways they can assign meaning to something. Given the fluidity of identity construction, people have to grapple with several, sometimes contradicting, ways of understanding. Sometimes, this flexibility and adaptability in one's identity can be good. However, in most cases, this identity-based equivocality can lead to confusion, cognitive burden, and, in turn, lead people towards heuristics-based understanding [90].

3.1.2 *Identity Construction in the* ***Human-Machine*** *Context.*

Consider Platform X, a popular social media site which uses an ML model for content moderation, with two stakeholders in mind. First, Sharon, a 42 year old conservative in the U.S. who is against vaccination for COVID-19. Her recent posts include graphic descriptions and images of, what she claims, are the potential side-effects of getting vaccinated. Second, Avery, a 37 year old doctor who believes it is their responsibility to share unfiltered information about the COVID-19 pandemic. Several of their posts highlight the positives of getting vaccinated, and some of them present the rare potential side-effects that have been noted by medical professionals. For both Sharon and Avery, some posts have been removed by Platform X's content moderation model.

Social media platforms usually offer an explanation for post removal to maintain their user base and help people share content in line with their policies. With interpretability tools, these platforms can support richer explanations. Based on the local explanation from an interpretability tool, Sharon is told that her post was removed due to its content type, the number of her previously flagged posts, her predicted political affiliation based on her posting history, and the topic being COVID-19. She might immediately latch on to the predicted political affiliation as *the* explanation, and not try to understand the removal any further (i.e., sensemaking is not triggered because her identity remains intact). For Avery, who simply wants to share all relevant information given their identity as a doctor, the post removal might attack their needs for self-enhancement, self-efficacy, and self-consistency. As such, they might assume that the content type being graphic is the main reason for post removal—this would support their positive self-conception, and not require them to understand the model's reasoning any further.

Interpretability tools are designed to present information in a context-free, unbiased way. But, people rarely internalize information in this static way. Weick argues that whether or how people internalize an explanation is dependent on their identity as an individual and as a part of their varying social contexts.

Claim: Given multiple explanations, people will internalize the one(s) that support their identity in positive ways.

3.2 Social

AI/ML sensemaking is modified by social context because it represents the audience-oriented external factors that influence people as they try to understand the outputs of these systems. For interpretability, this suggests that explanations are internalized differently by people with different micro- and macro-social contexts.

3.2.1 *Social Elements of Sensemaking in the* ***Human-Human*** *Context.*

Sensemaking describes human cognition. This might give it the appearance of being about the individual, but it is not. Weick notes the work on socially shared cognition (e.g., [62, 91]) which shows that human cognition and social functioning are essential to each other. Specifically, an individual's conduct is dependent on their audience, whether this is an imagined, implied, or a physically present one [5, 15]. Regarding the lack of a need for a physically present audience, recall Weick's reference to Mead's work on the individual being "a parliament of selves" [74] (see Section 3.1 for details on socially-grounded identity construction).

A focus on social aspects of sensemaking naturally implies that modes of communication (e.g., speech, discourse) and tools that support these also get attention, since these represent the ways in which social contact is mediated. Weick describes their importance on three levels, which exist beyond the individual: (1) inter-subjective, the conversations with others that can lead to alignment; (2) generic subjective, the socially-established norms when alignment has been achieved; and (3) extra-subjective, the culturally-established norms that do not necessarily require communication anymore. As we go from inter- to extra-subjective, the role of the implied and invisible audience becomes increasingly prominent. This, in turn, shapes the modes and tools of communication necessary for sensemaking.

3.2.2 *Social Elements of Sensemaking in the* ***Human-Machine*** *Context.*

Consider the model developed for predicting diabetic retinopathy (DR) based on healthcare data (predominantly eye fundus photos) collected in the U.S. [11]. The U.S. healthcare system is consistent across organizations—there is low variability in how eye fundus photos are captured, how the medical records are stored, and who (a generalist or specialist doctor) makes a diagnosis. However, when the same model was applied to a different social and cultural context—in Thailand, where healthcare is dependent on individual providers and patient needs in different regions—it failed in unanticipated ways.

First, there is the issue with the data itself. Several countries in Southeast Asia, including Thailand, do not have dedicated rooms for capturing fundus photos, making the photos inconsistent in opacity and leading to potentially inaccurate predictions. Second, there are established norms around the results of a DR screening test. While it is often expected to receive results immediately in the U.S. healthcare system, this is less common in Thailand, with fewer technicians, doctors, and specialists. Patients living in smaller towns have to travel to larger cities for appointments with specialists. A patient who is anticipating their DR result 4-5 weeks later might not have budgeted enough time for travel, based on a referral on the same day as the DR screening test visit.

While interpretability tools may offer an explanation, these explanations are limited to the model and the training dataset. Weick's perspective suggests that it might not be enough to explain the prediction, due to the variability in people's social contexts when using predictions in real-world settings; recent work on domain and distributional shifts in ML datasets supports this perspective [55].

Claim: Differences in micro- and macro-social contexts affect the effectiveness of explanations.

3.3 Retrospective

Retrospection or reflective thinking influences AI/ML sensemaking by engaging people in deliberately thinking about the diverse interpretations of outputs when trying to understand these systems, instead of following the more automated, heuristics-based, reasoning pathways. For interpretability, this suggests that providing explanations before people can reflect on the model and its predictions negatively affects sensemaking.

3.3.1 *Retrospective Sensemaking in the* ***Human-Human*** *Context.*

Sensemaking is retrospective because the object of sensemaking is a *lived experience*. Weick describes the retrospective nature of

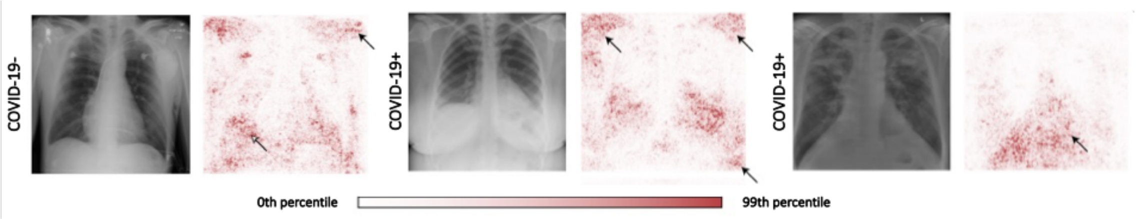


Figure 2: Saliency maps for chest radiographs, adapted from [20].

sensemaking as the most important, but perhaps the least noticeable, property. The reason it so frequently goes unnoticed is because of how embedded retrospection is in the sensemaking process. Retrospective sensemaking is derived from the work of Schutz, who believes that meaning is "merely an operation of intentionality, which. . . only becomes visible to the reflective glance" [97, 98]. The lived timeframe being considered for reflection can be the short- or long-term past, ranging from minutes, days, and years to "as I begin the latter portion of a long word, my utterance of the first part is already in the past" [36, p.44].

The retrospective process starts with an individual's present circumstances, and those shape the past experiences selected for sensemaking. Reflection happens in the form of a cone of light that starts with the present and spreads backwards. In this way, the cues of the past lived experience that are paid attention to for sensemaking depend on how the present is shaped. The challenge lies in *which* present to consider. People typically have several things on their mind at the same time, be it multiple projects at work or personal goals. With these, they have a multitude of lenses that they could apply for the reflective sensemaking process—the object of their sensemaking thus becomes equivocal. When dealing with equivocality, people are already overwhelmed with information and providing more details is often not helpful. "Instead, they need values, priorities, and clarity about preferences to help them be clear about which projects matter" [118, p.28]. In looking for clarity on which meaning to select, people are prone to a hindsight bias [107]. They select the most plausible story of causality for the outcome that they are trying to explain (Section 3.7 describes this property of sensemaking: being driven by plausibility over accuracy).

3.3.2 Retrospective Sensemaking in the ***Human-Machine*** *Context.* For ML-based systems, the model and its predictions are the "lived experiences." Consider a radiologist tasked with reading chest radiographs to determine if a patient has COVID-19. The hospital has purchased an ML-based image classification system. To help determine if the predictions makes sense, the software also provides saliency maps (an interpretability approach).

By immediately providing an explanation, the interpretability tool effectively disengages the retrospective process that helps with sensemaking. Figure 2 shows example explanations provided to the radiologist. As described in the accompanying research paper [20], these explanations show that the ML model sometimes relies on laterality markers to make the prediction. For example, in Figure 2, the saliency maps highlight not only the relevant regions in the lungs as being predictive, but also some areas (see pointers) that differ based on how the radiograph was taken. These, coincidentally, are also predictive of COVID-19 positive vs. negative results, leading to a spurious correlation.

Ideally, the radiologist evaluating the saliency map would be able to reach the same conclusion regarding these spurious correlations. However, the retrospective property would suggest that by providing this explanation without asking the radiologist to first think about what the explanation could be, the interpretability tool disengages their retrospective sensemaking process. This makes it easier for the radiologist to craft a plausible narrative that agrees with the model's prediction instead of analyzing the radiograph in detail and accurately understanding the model. When they immediately have the explanation, there is no cognitive need for the radiologist to understand the intricacies of the model, which increases the likelihood of them missing the issues with the model. Prior work on stakeholders' use of interpretability tools corroborates this perspective: people expect far more from interpretability tools than their actual capabilities and, in doing so, often end up over-trusting and misusing them [10, 52].

Claim: Providing explanations before people can reflect on the model and its predictions negatively affects sensemaking.

3.4 Enactive of Sensible Environments

Enactment is critical for AI/ML sensemaking because it represents how (much) people understand these systems—it reflects the parts of these systems that people understand, and then build on, over time. For interpretability, this suggests that the order in which explanations are seen affects how people understand a model and its predictions.

3.4.1 Enactment in the ***Human-Human*** *Context.* When we are tasked with making sense of something, it might appear to belong to an external environment that we must observe and understand. Weick argues that this is not the case, that sensemaking works such that "people often produce part of the environment they face" [118, p.30]. It is not just the person, rather, the person and their enacted environment that is the unit of analysis for sensemaking [87].

This environment that provides the necessary context for sensemaking is not a monolithic, fixed environment that exists external to people. Rather, people act, and their actions shape the environmental context needed for sensemaking: "they act, and in doing so create the materials that become the constraints and opportunities they face." [118, p.31]. Here, Weick is influenced by Follett, who claims that there is no subject or object in meaning-making. There is no meaning that one understands as the "result of the process;" there is just a "moment in process" [30, p.60]. As such, this meaning is inherently contextual in that it is shaped by the cycle of action-enaction between the human and their environment.

Weick cautions against two things with the enactive nature of sensemaking. First, to not restrict our definition of action in shaping our environment. Action here could mean creating, reflecting, or interpreting: "the idea that action can be inhibited, abandoned, checked, or redirected, as well as expressed, suggests that there are many ways in which action can affect meaning other than by producing visible consequences in the world" ([14], described by Weick [118, p.37]). Second, the enacted environments do not need to embody existing ones. People want to believe that the world is defined using pre-given features, i.e., knowledge and meaning exist, we just need to find them. This is called Cartesian anxiety: "a dilemma: either we have a fixed and stable foundation for knowledge, a point where knowledge starts, is grounded, and rests, or we cannot escape some sort of darkness, chaos, and confusion" [112, p.140]. When faced with equivocal meanings, people want to select ones that reduce Cartesian anxiety. But, in doing so, they also enable existing, socially constructed meanings to shape their sensemaking. This can be helpful in providing the clarity of values needed when faced with equivocality, or it can privilege some meanings over others, depending on agency and power [94].

3.4.2 Enactment in the ***Human-Machine*** *Context.* Enactment is most apparent when ML-based systems are used in urgent or reactive situations, such as predictive policing. Consider PredPol, which uses location-based ML models that rely on connections between places and their historical crime rates to identify hot spots for police patrol [1]. Say a police officer is monitoring PredPol to allocate patrol units to various neighborhoods. The model's predictions influence both the officer monitoring the software as well as those patrolling. Both will update their "environment" to be focused on certain neighborhoods. That is, they are primed to look for criminal activity in these neighborhoods. Additionally, when arrests are made using model predictions, they provide further evidence to the model that the patterns it has identified are accurate. In this way, the feedback loop causes the model to become increasingly biased [38]. If the police officers were also provided an explanation for the model's predictions, the type of explanation and the order in which they are seen (e.g., global vs. local explanation first) changes the enacted environment for the officers. The sensemaking perspective offers several properties for how the environment could be shaped (e.g., people's identity, social network).

Interpretability tools offer different types of information (e.g., feature importances, partial dependency plots, data distributions), but do not impose an order on how this information is explored. End-users can take different paths to reaching conclusions about the model. Because sensemaking is sensitive to enacted environments, it is important to remember that any information or explanation about the model is not treated by people as static or isolated.

Claim: The order in which explanations are seen affects how people understand a model and its predictions.

3.5 Ongoing

The ongoing nature of AI/ML sensemaking highlights how interruptions and emotions can influence what is understood about these systems. For interpretability, this suggests that, if interrupted when viewing an explanation, the valence and magnitude of the resulting emotion can change what people understand about the model and its predictions.

3.5.1 Sensemaking as an Ongoing Activity in the ***Human-Human*** *Context.* Sensemaking never starts or stops; people are always in the middle of something. To think otherwise would suggest that people are able to chop meaningful moments from the flow of time, but that would be counter-intuitive because to determine whether something is "meaningful" would require sensemaking in the first place [22, 93]. Sensemaking is akin to being in situations of thrownness. Winograd and Flores describe these situations as having the following properties: (1) you cannot avoid acting; (2) you cannot step back and reflect on your actions, i.e., you have to rely on your intuitions; (3) the effects of action cannot be predicted; (4) you do not have a stable representation of the situation; (5) every representation is an interpretation, i.e., no objective analysis can be performed in the moment; and (6) language is action, i.e., people enact the situation via their descriptions of their environment, making it impossible to stay detached from it [123].

Emotion is embedded in sensemaking via the following process. Interruptions trigger arousal, i.e., a discharge in the autonomic nervous system, which convinces the individual that something in the environment has changed, that they must understand it and take appropriate action to get back to a state of flow [13, 73]. The higher the arousal post-interruption, the stronger the emotional response and, in turn, the stronger the affect of emotion on sensemaking. Why does it matter if there is an emotional response during an ongoing sensemaking process? Emotions affect sensemaking in that recall and retrospect are dependent on one's mood [105]. Specifically, people recall events that are congruent with their current emotional valence. Of all the past events that might be relevant to sensemaking in a current situation, the ones we recall are not those that look the same, but those that feel the same.

3.5.2 Sensemaking as an Ongoing Activity in the ***Human-Machine*** *Context.* Consider the PredPol example again. Let's assume the arrest record shows that the likelihood of a legitimate arrest in an area predicted as a hot spot by the model is 40%. The officer monitoring the model outputs is made aware of this number every time they log into the system. Imagine this happens one day: the patrol officers allocated to one of the hot spots make a legitimate arrest. The monitoring officer is commended for their role in anticipating the situation. This happens several times during the day. Thus, the monitoring officer associates positive feedback with arrests based on the model's predictions. When writing their report about the incidents, they use the explanations provided by the software to further justify their choices.

Next day, the patrol officers make another arrest in the same predicted hot spot. The monitoring officer is once again asked to record an explanation for selecting that area for patrol. Before they do so, they happen to look at social media and notice several posts showing outrage with regards to that arrest. This is an interruption, as described by the ongoing property of sensemaking. This time, when the monitoring officer is writing up their explanation, it could be that they mention that the model's predictions are not always right and highlight some other failure cases.

As we have noted before, information presented in explanations is rarely used in context-free settings. Despite being shown the same explanation, the monitoring officer could notice different aspects of it depending on whether they were interrupted, whether the interruption led to positive or negative emotional states, and the magnitude of those emotions.

Claim: The valence and magnitude of the emotion caused by an interruption during the process of understanding explanations from interpretability tools change what is understood.

3.6 Focused on and by Extracted Cues

Extracted cues modify AI/ML sensemaking because they represent the (incomplete) bits of information that people rely on when trying to understand these systems. For interpretability, this suggests that highlighting different parts of explanations can lead to varying understanding of the underlying data and model.

3.6.1 Extracting Cues in the ***Human-Human*** *Context.* Weick describes extracted cues as "simple, familiar structures that are seeds from which people develop a larger sense of what may be occurring" [118, p.50]. These extracted cues are important for sensemaking because they are taken as "equivalent to the entire datum from which they come" and in being taken as such, they "suggest a certain consequence more obviously than it was suggested by the total datum as it originally came" [48, p.340]. Sensemaking uses extracted cues like a partially completed sentence. The completed first half of the sentence constrains what the incomplete second half could be [101].

Extracting cues involves two processes—noticing and bracketing—which are both affected by context. First, context affects which cues are extracted based on what is noticed by the sensemaker. *Noticing* is an informal, even involuntary, observation of the environment that begins the process of sensemaking [106]. Cues that are noticed are either novel, unusual, or unexpected, or those that we are situationally or personally primed to focus on (e.g., recently or frequently encountered cues) [110]. Second, context affects how the extracted (noticed) cues are interpreted. Without context, any cues that are extracted lead to equivocal meanings [61]. These situations of equivocality need a clarity of values instead of more information for sensemaking (Section 3.3). Context can provide this clarity in the form of, for example, the social and cultural norms of the setting where sensemaking in happening. During the process of extracting cues, people are trying to form a cognitive reference map that presumes that there is a connection between the situation/outcome and the cue. However, important cues can be missed when people do not have any prior experience with the situation.

3.6.2 Extracting Cues in the ***Human-Machine*** *Context.* Consider the example where a company provides ML-based software to organizations to help them with hiring decisions. A marketing company uses this software to shortlist candidates by sending some questions in advance. The candidates answer these questions in a video format, and the ML-based software analyzes these videos and provides a hiring score along with an explanation. The kind of input data used by the model includes demographic information; prior experience from the candidate's resume; and tone of voice, perceived enthusiasm, and other emotion data coded by the software after analyzing the recorded video [51].

Let's say that the marketing company is using this software to shortlist candidates for the position of a sales representative. The software shows that A is a better candidate than B and explains its ratings (based on local explanations from interpretability tools). The HR folks see that A's rating is based on their facial expressions during the interview (they were smiling, not visibly nervous, and seemed enthusiastic). They consider these to be good attributes for a sales representative and hire A even though B is more qualified. Additional information about A's and B's qualifications is also noted in the local explanations but might not be the cues that are extracted or focused on in this instance.

Current interpretability tools present all types of information and let the user decide how to explore. Weick cautions against this unstructured exploration because it leads to equivocal alternatives for understanding an ML-based system. Which one of these alternatives is ultimately selected can be a reasonable, reflective process or entirely arbitrary.

Claim: Highlighting different parts of explanations can lead to varying understanding of the underlying data and model.

3.7 Driven by Plausibility rather than Accuracy

Recognizing that people are driven by plausibility rather than accuracy is critical for AI/ML sensemaking because we must account for people's inclination to only have a "good enough" understanding of these systems. For interpretability, this suggests that, given plausible explanations, people are not inclined to search for the accurate one amongst these.

3.7.1 Plausibility over Accuracy in the ***Human-Human*** *Context.* Weick argues that accuracy is nice but not necessary for sensemaking. Even when it is necessary, people rarely achieve it. Instead, people rely on plausible reasoning which is: (1) not necessarily correct but fits the facts, and (2) based on incomplete information [47]. When sensemaking, people can be influenced by what is "interesting, attractive, emotionally appealing, and goal relevant" [29].

Weick notes eight reasons for why accuracy is secondary to sensemaking. Most important among these, *first*, it is impossible to internalize the overwhelming amount of information available for sensemaking. To cope with this, people apply relevance filters to the information [31, 104]. *Second*, when people filter what they notice, this biased noticing can be good for action, though not for deliberation. But, deliberation is not the goal, it is "futile in a changing world where perceptions, by definition, can never be accurate" [118, p.60]. *Third*, at the time of sensemaking, it is impossible to tell if the sensemaker's perceptions will be accurate. It is only in retrospect—after the sensemaker has taken action based on their understanding—that they evaluate their perceptions for accuracy.

With accuracy not being necessary for sensemaking, it is only natural to ask: what is? Weick claims that what is necessary for sensemaking is a good story, "something that preserves plausibility and coherence, something that is reasonable and memorable, something that embodies past experiences and expectations, something that resonates with other people, something that can be constructed retrospectively but also can be used prospectively, something that captures both feeling and thought, something that allows for embellishment to fit current oddities, something that is fun to construct" [118, pp.60-61]. Stories help with sensemaking because they are templates from previous attempts at making sense of similar situations. Overall, this property is often amplified by the others in that the plausible narratives could depend on people's identity, implied or actual audience, extracted cues, emotional state, etc.

3.7.2 Plausibility over Accuracy in the ***Human-Machine*** *Context.* Interpretability outputs, such as text or visual explanations, inherently present a story. As long as this explanation / story is plausible, there is no reason for an individual to evaluate it for accuracy. Consider the example with the radiologist again, where they are tasked with deciding whether a chest radiograph shows that the patient has COVID-19. Their decision-making is supported by an ML-based software that has been trained on publicly available chest radiograph datasets. To help them understand the model's reasoning for a prediction, the radiologist has access to saliency maps as interpretable outputs (Figure 2).

According to Weick, when using the saliency map to determine whether the model's prediction makes sense, the radiologist is essentially searching for a plausible story that explains the prediction. The explanations in Figure 2 show some areas inside the lungs as relevant, a plausible reason for predicting COVID-19. The radiologist could believe this plausible explanation and choose to follow it. Human evaluations of interpretability tools show that this confirmatory use of explanations is often the case, even when explanations reveal issues with the underlying model [10, 16, 52].

Let's say that the radiologist was not immediately convinced that the prediction was accurate after seeing the saliency maps. Maybe they looked at one of them (e.g., Figure 2-Middle) and noticed that the radiograph's edges (by the person's shoulders and diaphragm) were also salient for the prediction. Even with this observation, the radiologist is looking for a plausible story. Perhaps the patient was coughing and could not stay still when the radiograph was being captured? That could explain the lateral markers for a COVID-19 positive patient. The model is relying on spurious correlations, but, with the role of plausibility in sensemaking, the radiologist might not try to accurately interpret the saliency map.

Claim: Given plausible explanations for a prediction, people are not inclined to search for the accurate one amongst these.

3.8 Summary

When designing solutions for promoting human understanding of ML models, we must consider the nuances of human cognition in addition to the technical solutions which explain ML models. Sensemaking provides a set of properties that describe these nuances—each of these can be seen as a self-contained set of research questions and hypotheses that relates to the other six. As the human-machine examples show, sensemaking properties could explain external factors that shape the information that is ultimately internalized by people when they use interpretability tools.

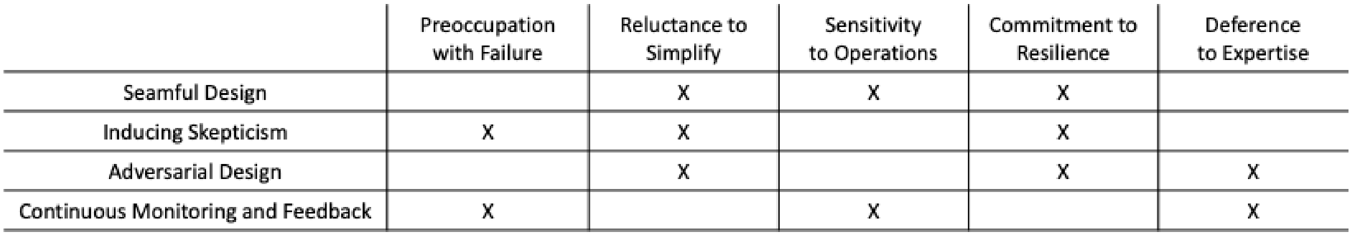
4 DISCUSSION

We propose a framework for Sensible AI to account for the properties of human cognition described by sensemaking. This has the potential to refine the explanations from interpretability tools for human consumption and to better support the human-centered desiderata of ML-based systems. How do we do this? Once again, Weick (along with his colleagues) proposes a solution: to explicitly promote or amplify sensemaking, we can follow the model of *mindful organizing* [119]. Sensemaking and *organizing* are inextricably intertwined. While sensemaking describes the meaning-making process of understanding, organizing describes the final outcome (e.g., a map or frame of reference) that represents the understanding. They belong to the same mutually interdependent, cyclical, recursive process—sensemaking is the process by which organizing is achieved [8, 120]. *Mindfulness* is expressed by actively refining the existing categories that we use to assign meaning, and creating new categories as needed for events that have not been seen before [59, 114, 119].

Mindful organizing was proposed after observing high-reliability organizations (HROs). HROs are organizations that have successfully avoided catastrophes despite operating in high-risk environments [95, 117]. Examples of these include healthcare organizations, air traffic control systems, naval aircraft carriers, and nuclear power plants. Mindful organizing embodies five principles consistently observed in HROs: (1) **preoccupation with failure**, anticipating potential risks by always being on the lookout for failures, being sensitive to even the smallest ones; (2) **reluctance to simplify**, wherein each failure is treated as unique because routines, labels, and cliches can stop us from looking into details of an event; (3) **sensitivity to operations**, a heightened awareness of the state of relevant systems and processes because systems are not static or linear, and expecting uncertainty in anticipating how different systems will interact in the event of a crisis; (4) **commitment to resilience**, prioritizing training for emergency situations by incorporating diverse testing pathways and team structures, and when a failure occurs, trying to absorb strain and preserve function; and (5) **deference to expertise**, assuming that people who are in the weeds—often lower-ranking individuals—have more knowledge about the situation, and valuing their opinions. Our proposal for Sensible AI encompasses designing, deploying, and maintaining systems that are reliable by learning from properties of HROs. Table 2 presents the corresponding principles of HROs that serve as inspiration for each idea.

4.1 Seamful Design

We can help people understand AI and ML by giving them the agency to do so. Often, ML-based systems and interpretability tools are designed with seamless interaction and effortless usability in mind. However, this can engage people's automatic reasoning mode, leading them to use ML outputs without adequate deliberation [10, 16, 52]. Highlighting complex details of ML outputs and processes—seamful design [46]—can promote the reluctance to simplify that has helped HROs. It can also add a sensitivity to operations when changes to inputs for models can be clearly seen in the outputs. Enhancing reconfigurability of ML models and training people to understand their complexity can reduce automatic, superficial evaluations. Increasing user control in the form of seamful design has the added benefit of introducing opportunities for informational interruptions, which are helpful for the commitment to resilience seen in HROs. While current interpretability tools have interactive features that provide additional information as needed, contextualizing this information using narratives can help people maintain overall situational awareness and avoid dysfunctional momentum when using ML-based systems. For example, when a doctor is viewing a predicted diagnosis, a Sensible AI system could prompt



	Preoccupation with Failure	Reluctance to Simplify	Sensitivity to Operations	Commitment to Resilience	Deference to Expertise
Seamful Design		X	X	X	
Inducing Skepticism	X	X		X	
Adversarial Design		X		X	X
Continuous Monitoring and Feedback	X		X		X

Table 2: Principles of high-reliability organizations (columns) that inspired our design ideas (rows) for Sensible AI.

them to view cases with similar inputs but different diagnoses. Next, we discuss ways to design these systems without overloading the end-user with features, interactivity, and information.

4.2 Inducing Skepticism

One way to reduce over-reliance on generalizations and known information—both common outcomes of sensemaking—is to create situations in which people would ask questions. We call this inducing skepticism, an idea suggested in prior work as a strategy for promoting reflective design [100]. Inducing skepticism can foster a preoccupation with failure, an HRO principle that encourages cultivating a doubt mindset in employees. HRO employees are always on the lookout for anomalies, they interpret any new cues from their systems in failure-centric ways, and collectively promote wariness. This can be incorporated in ML-based systems, for example, by suggesting that end-users ask about how a particular prediction is unique or similar to other data points, questioning outputs of interpretability tools sometimes (e.g., "does this feature importance value make sense?"), presenting bottom-n feature importances in an explanation instead of top-n, highlighting cases for which the model is unsure of its predictions, etc. Inducing skepticism can also be accomplished in social ways, by promoting diversity in teams, both in terms of skillsets and experience. For example, novices can prompt experts to view an AI output in more detail when they ask questions about it. This diversity is a common way in which HROs maintain their commitment to resilience. These technical and social ways of inducing skepticism have a common goal, a reluctance to simplify by adding complexity and diversity to a situation.

4.3 Adversarial Design

No one person can successfully anticipate all failures, even when the system induces skepticism. Adversarial design suggests relying on social and organizational networks for this task. Adversarial design is a form of political design rooted in the theory of agonism: promoting productive contestation and dissensus [23, 80, 122]. By designing Sensible AI systems with dissensus-centric features, we can increase the likelihood that *someone* raises a red flag given early signals of a failure situation. Prior work has implemented adversarial design in the form of red teaming in technical and social ways (e.g., adversarial attacks for testing and promoting cybersecurity [2], and forming teams with collective diversity and supporting deliberation [33, 43, 44], respectively). Here, HRO principles of reluctance to simplify, commitment to resilience, and deference to expertise can be observed in practice. We propose technical redundancies and social diversity to reduce unanticipated failures in understanding AI outputs, as one way of operationalizing adversarial design. Technical redundancies can be implemented as system features wherein multiple people view the same output in different contexts, giving the team a better chance of finding potential issues. Social or organizational diversity can be expanded by including people with different roles, skillsets, and opinions. The more diversity in people viewing the outputs, the higher the likelihood that they collectively discover an issue, as long as deliberation is made easy [43].

4.4 Continuous Monitoring and Feedback

When ML-based systems are deployed in real-world settings, changes in data collection and distributional drifts are a given [55]. To manage these, researchers and practitioners have proposed MLOps—an extension of DevOps practices from software to ML-based settings—to include continuous testing, integration, monitoring, and feedback loops in maintaining the operation of ML-based systems in the wild [71]. We propose incorporating social features in this pipeline by designing for HRO principles such as preoccupation with failure, sensitivity to operations, and deference to expertise. For example, include (1) continuous failure monitoring, effectively serving as distributed fire alarms that can be engaged by people at varying levels in an organization, and (2) model maintenance, by relying on people on the ground for detailed understanding of failure cases, as seen in organizations that perform failure panels, audits, etc.

5 CONCLUSION

Interpretability and explainability approaches are designed to help stakeholders adequately understand the predictions and reasoning of an ML-based system. Although these approaches represent complex models in simpler formats, they do not account for the contextual factors that affect whether and how people internalize information. We have presented an alternate framework for helping people understand ML models grounded in Weick's sensemaking theory from organizational studies. Via its seven properties, sensemaking describes the individual, environmental, social, and organizational context that affects human understanding. We translated these for the human-machine context and presented a research agenda based on each property. We also proposed a new framework—Sensible AI—that accounts for these nuances of human cognition and presented initial design ideas as a concrete path forward. We hope that by accounting for these nuances, Sensible AI can support the desiderata (e.g., reliability, robustness, trustworthiness, accountability, fair and ethical decision-making, etc.) that interpretability and explainability are intended for.

ACKNOWLEDGMENTS

We thank our reviewers for their helpful comments. We are also grateful to Mitchell Gordon, Stevie Chancellor, and Michael Madaio for their feedback and support. Harmanpreet Kaur was supported by the Google PhD fellowship.

REFERENCES

[1] 2020. Law enforcement: Predpol Law Enforcement Intelligence led Policing Software: Predpol Law Enforcement Intelligence led Policing Software. https://www.predpol.com/law-enforcement/

[2] Hussein Abbass, Axel Bender, Svetoslav Gaidow, and Paul Whitbread. 2011. Computational Red Teaming: Past, Present and Future. *IEEE Computational Intelligence Magazine* 6, 1 (2011), 30–42. https://doi.org/10.1109/MCI.2010.939578

[3] 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 *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3173574.3174156

[4] Ali Alkhatib. 2021. To Live in Their Utopia: Why Algorithmic Systems Create Absurd Outcomes. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 95, 9 pages. https://doi.org/10.1145/3411764.3445740

[5] Gordon W Allport. 1985. The historical background of social psychology (Vol. 1). *The handbook of social psychology* (1985).

[6] David Alvarez-Melis and Tommi Jaakkola. 2017. A causal framework for explaining the predictions of black-box sequence-to-sequence models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 412–421. https://doi.org/10.18653/v1/D17-1042

[7] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *AI Magazine* 35, 4 (2014), 105–120. https://doi.org/10.1609/aimag.v35i4.2513

[8] Mary Ann Glynn and Lee Watkiss. 2020. Of Organizing and Sensemaking: From Action to Meaning and Back Again in a Half-Century of Weick's Theorizing. *Journal of Management Studies* 57, 7 (2020), 1331–1354. https://doi.org/10.1111/joms.12613

[9] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (6 2020), 82–115. https://doi.org/10.1016/J.INFFUS.2019.12.012

[10] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 81, 16 pages. https://doi.org/10.1145/3411764.3445717

[11] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M. Vardoulakis. 2020. A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376718

[12] Victoria Bellotti and Keith Edwards. 2001. Intelligibility and Accountability: Human Considerations in Context-Aware Systems. *Human–Computer Interaction* 16, 2-4 (2001), 193–212. https://doi.org/10.1207/S15327051HCI16234_05

[13] Ellen Berscheid. 1983. Emotion. In *Close Relationships*, H.H. Kelley, E. Berscheid, A. Christensen, J. Harvey, T. Huston, G. Levinger, E. McClintock, A. Peplau, and D.R. Peterson (Eds.). WH Freeman, 110–168.

[14] Herbert Blumer. 1969. *Symbolic interactionism*. Vol. 50. Englewood Cliffs, NJ: Prentice-Hall.

[15] T Bruns and GM Stalker. 1961. The management of innovation. *Tavistock, London* (1961), 120–122.

[16] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 188 (Apr 2021), 21 pages. https://doi.org/10.1145/3449287

[17] Jason W Burton, Mari-Klara Stein, and Tina Blegind Jensen. 2020. A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making* 33, 2 (2020), 220–239. https://doi.org/10.1002/bdm.2155

[18] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Sydney, NSW, Australia) *(KDD '15)*. ACM, New York, NY, USA, 1721–1730. https://doi.org/10.1145/2783258.2788613

[19] Randall Davis, Bruce Buchanan, and Edward Shortliffe. 1977. Production rules as a representation for a knowledge-based consultation program. *Artificial intelligence* 8, 1 (1977), 15–45. https://doi.org/10.1016/0004-3702(77)90003-0

[20] Alex J DeGrave, Joseph D Janizek, and Su-In Lee. 2021. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence* (2021), 1–10. https://doi.org/10.1038/s42256-021-00338-7

[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), 114–126. https://doi.org/10.1037/xge0000033

[22] Wilhelm Dilthey and Frederic Jameson. 1972. The rise of hermeneutics. *New literary history* 3, 2 (1972), 229–244.

[23] Carl DiSalvo. 2015. *Adversarial design*. Design Thinking, Design Theory.

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

[25] Paul Dourish. 2016. Algorithms and their others: Algorithmic culture in context. *Big Data & Society* 3, 2 (2016), 1–11. https://doi.org/10.1177/2053951716665128

[26] Upol Ehsan, Q Vera Liao, Michael Muller, Mark O Riedl, and Justin D Weisz. 2021. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–19. https://doi.org/10.1145/3411764.3445188

[27] Upol Ehsan, Samir Passi, Q Vera Liao, Larry Chan, I Lee, Michael Muller, Mark O Riedl, et al. 2021. The who in explainable ai: How ai background shapes perceptions of ai explanations. *arXiv preprint arXiv:2107.13509* (2021).

[28] Miriam Erez, P Christopher Earley, et al. 1993. *Culture, self-identity, and work*. Oxford University Press on Demand.

[29] Susan T Fiske. 1992. Thinking is for doing: portraits of social cognition from daguerreotype to laserphoto. *Journal of personality and social psychology* 63, 6 (1992), 877–889. https://doi.org/10.1037/0022-3514.63.6.877

[30] Mary Parker Follett. 1924. *Creative experience*. Longmans, Green and company.

[31] Gerd Gigerenzer. 1991. How to make cognitive illusions disappear: Beyond "heuristics and biases". *European review of social psychology* 2, 1 (1991), 83–115. https://doi.org/10.1080/14792779143000033

[32] Marco Gillies, Rebecca Fiebrink, Atau Tanaka, Jérémie Garcia, Frédéric Bevilacqua, Alexis Heloir, Fabrizio Nunnari, Wendy Mackay, Saleema Amershi, Bongshin Lee, et al. 2016. Human-centred machine learning. In *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems*. 3558–3565.

[33] Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. 2022. Jury Learning: Integrating Dissenting Voices into Machine Learning Models. In *CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 115, 19 pages. https://doi.org/10.1145/3491102.3502004

[34] Herbert P Grice. 1975. Logic and conversation. In *Speech acts*. Brill, 41–58. https://doi.org/10.1163/9789004368811_003

[35] David Gunning and David Aha. 2019. DARPA's explainable artificial intelligence (XAI) program. *AI Magazine* 40, 2 (2019), 44–58. https://doi.org/10.1609/aimag.v40i2.2850

[36] C. Hartshorne. 1962. Mind as Memory and Creative Love. In *Theories of the Mind*, Jordan M. Scher (Ed.). Free Press of Glencoe, 440–463.

[37] T J Hastie and R J Tibshirani. 1990. *Generalized Additive Models*. CRC Press.

[38] Will Douglas Heaven. 2020. Predictive policing algorithms are racist. They need to be dismantled. *MIT Technology Review* 17 (2020), 2020.

[39] Carl G Hempel and Paul Oppenheim. 1948. Studies in the Logic of Explanation. *Philos. Sci.* 15, 2 (1948), 135–175. https://doi.org/10.1086/286983

[40] Denis J Hilton. 1996. Mental models and causal explanation: Judgements of probable cause and explanatory relevance. *Thinking & Reasoning* 2, 4 (1996), 273–308. https://doi.org/10.1080/135467896394447

[41] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M. Drucker. 2019. Gamut: A Design Probe to Understand How Data Scientists Understand Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. ACM, New York, NY, USA, Article 579, 13 pages. https://doi.org/10.1145/3290605.3300809

[42] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. 2019. Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland UK) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3290605.3300830

[43] Lu Hong and Scott Page. 2020. The Contributions of Diversity, Accuracy, and Group Size on Collective Accuracy. *Accuracy, and Group Size on Collective Accuracy (October 15, 2020)* (2020). https://doi.org/10.2139/ssrn.3712299

[44] Lu Hong and Scott E Page. 2004. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences* 101, 46 (2004), 16385–16389. https://doi.org/10.1073/pnas.0403723101

[45] 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. https://doi.org/10.1145/3392878

[46] Sarah Inman and David Ribes. 2019. "Beautiful Seams": Strategic Revelations and Concealments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland UK) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300508

[47] Daniel J Isenberg. 1986. The structure and process of understanding: Implications for managerial action. In *The Thinking Organization*, H.P. Sims Jr. and D.A. Gioia (Eds.). Jossey Bass, San Fransisco, 238–262.

[48] William James. 2007. *The principles of psychology*. Vol. 1. Cosimo, Inc.

[49] P Devereaux Jennings and Royston Greenwood. 2003. Constructing the iron cage: Institutional theory and enactment. *Debating organization: point-counterpoint in organization studies* 195 (2003).

[50] Jongbin Jung, Connor Concannon, Ravi Shroff, Sharad Goel, and Daniel G Goldstein. 2017. Simple rules for complex decisions. *Available at SSRN 2919024* (2017). http://dx.doi.org/10.2139/ssrn.2919024

[51] Jeremy Kahn. 2021. HireVue drops facial monitoring amid A.I. algorithm audit. *Fortune* (2021).

[52] 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*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376219

[53] Karin D. Knorr-Cetina. 1981. The micro-sociological challenge of macro-sociology : towards a reconstruction of social theory and methodology. In *Advances in social theory and methodology: toward an integration of micro- and macro-sociologies*, K. Knorr-Cetina and A. V. Cicourel (Eds.). Routledge & Kegan Paul, Boston, 1–47.

[54] Rafal Kocielnik, Saleema Amershi, and Paul N Bennett. 2019. Will you accept an imperfect AI? exploring designs for adjusting end-user expectations of ai systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14. https://doi.org/10.1145/3290605.3300641

[55] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. 2021. WILDS: A Benchmark of in-the-Wild Distribution Shifts. In *Proceedings of the 38th International Conference on Machine Learning*, Marina Meila and Tong Zhang (Eds.), Vol. 139. PMLR, 5637–5664. https://proceedings.mlr.press/v139/koh21a.html

[56] Ravi S Kudesia. 2017. Organizational sensemaking. In *Oxford research encyclopedia of psychology*.

[57] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on visual languages and human centric computing*. IEEE, 3–10. https://doi.org/10.1109/VLHCC.2013.6645235

[58] Himabindu Lakkaraju, Stephen H Bach, and L Jure. 2016. Interpretable Decision Sets: A Joint Framework for Description and Prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery. https://doi.org/10.1145/2939672.2939874

[59] Ellen J Langer. 1989. Minding matters: The consequences of mindlessness–mindfulness. In *Advances in experimental social psychology*. Vol. 22. Elsevier, 137–173.

[60] David B Leake. 1991. Goal-based explanation evaluation. *Cognitive Science* 15, 4 (1991), 509–545.

[61] Kenneth Leiter. 1980. *A primer on ethnomethodology*. Oxford University Press, USA.

[62] John M Levine, Lauren B Resnick, and E Tory Higgins. 1993. Social foundations of cognition. *Annual review of psychology* 44, 1 (1993), 585–612.

[63] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3313831.3376590

[64] Peter Lipton. 1990. Contrastive explanation. *Royal Institute of Philosophy Supplements* 27 (1990), 247–266.

[65] Zachary C. Lipton. 2018. The mythos of model interpretability. *Commun. ACM* 61 (9 2018), 36–43. Issue 10. https://doi.org/10.1145/3233231

[66] Stine Lomborg and Patrick Heiberg Kapsch. 2020. Decoding algorithms. *Media, Culture & Society* 42, 5 (2020), 745–761. https://doi.org/10.1177/0163443719855301

[67] Tania Lombrozo. 2006. The Structure and Function of Explanations. *Trends in cognitive sciences* 10, 10 (2006), 464–470. https://doi.org/10.1016/j.tics.2006.08.004

[68] Tania Lombrozo. 2012. Explanation and abductive inference. *The Oxford Handbook of Thinking and Reasoning* (2012), 260–276.

[69] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 4765–4774. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[70] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376445

[71] Sasu Mäkinen, Henrik Skogström, Eero Laaksonen, and Tommi Mikkonen. 2021. Who Needs MLOps: What Data Scientists Seek to Accomplish and How Can MLOps Help? *arXiv preprint arXiv:2103.08942* (2021).

[72] Bertram F Malle. 2006. *How the mind explains behavior: Folk explanations, meaning, and social interaction*. Mit Press.

[73] George Mandler. 1984. *Mind and body: Psychology of emotion and stress*. WW Norton & Company Incorporated.

[74] George Herbert Mead. 1934. *Mind, self and society*. Vol. 111. Chicago University of Chicago Press.

[75] David Alvarez Melis, Harmanpreet Kaur, Hal Daumé III, Hanna Wallach, and Jennifer Wortman Vaughan. 2021. From Human Explanation to Model Interpretability: A Framework Based on Weight of Evidence. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 9. 35–47.

[76] Matthew B Miles and A Michael Huberman. 1994. *Qualitative data analysis: An expanded sourcebook*. sage.

[77] Tim Miller. 2019. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence* 267 (2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[78] Tim Miller. 2021. Contrastive explanation: A structural-model approach. *The Knowledge Engineering Review* 36 (2021). https://doi.org/10.1017/S0269888921000102

[79] Tim Miller, Piers Howe, and Liz Sonenberg. 2017. Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. In *IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI)*.

[80] Chantal Mouffe. 2013. *Agonistics: Thinking the world politically*. Verso Books.

[81] Richard E Nisbett and Timothy D Wilson. 1977. Telling more than we can know: Verbal reports on mental processes. *Psychological review* 84, 3 (1977), 231–259. https://doi.org/10.1037/0033-295X.84.3.231

[82] Donald A Norman. 2014. Some observations on mental models. In *Mental models*. Psychology Press, 15–22.

[83] Samir Passi and Steven J Jackson. 2018. Trust in data science: Collaboration, translation, and accountability in corporate data science projects. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–28. https://doi.org/10.1145/3274405

[84] Charles Sanders Peirce. 1878. Illustrations of the Logic of Science: IV The Probability of Induction. *Popular Science Monthly* 12 (Apr 1878), 705–718.

[85] Peter Pirolli and Stuart Card. 2005. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of international conference on intelligence analysis*, Vol. 5. McLean, VA, USA, 2–4.

[86] Joseph C Pitt. 1988. *Theories of explanation*. Oxford University Press.

[87] Louis R Pondy and Ian I Mitroff. 1979. Beyond open system models of organization. *Research in organizational behavior* 1, 1 (1979), 3–39.

[88] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 237, 52 pages. https://doi.org/10.1145/3411764.3445315

[89] J R Quinlan. 1986. Induction of Decision Trees. *Mach. Learn.* (1986). https://doi.org/10.1023/A:1022643204877

[90] James Reason. 1990. *Human Error*. Cambridge university press.

[91] Lauren B Resnick, John M Levine, and Stephanie D Teasley. 1991. *Perspectives on socially shared cognition* (1st ed. ed.). Number Washington, DC :. American Psychological Association.

[92] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144. https://doi.org/10.1145/2939672.2939778

[93] Hans Peter Rickman. 1979. Dilthey selected writings. (1979).

[94] Peter S Ring and Andrew H Van de Ven. 1989. Formal and informal dimensions of transactions. *Research on the management of innovation: The Minnesota studies* 171 (1989), 192.

[95] Karlene H Roberts. 1990. Some characteristics of one type of high reliability organization. *Organization Science* 1, 2 (1990), 160–176. https://doi.org/10.1287/orsc.1.2.160

[96] Daniel M. Russell, Mark J. Stefik, Peter Pirolli, and Stuart K. Card. 1993. The Cost Structure of Sensemaking. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems* (Amsterdam, The Netherlands) *(CHI '93)*. Association for Computing Machinery, New York, NY, USA, 269–276. https://doi.org/10.1145/169059.169209

[97] Alfred Schutz. 1972. *The phenomenology of the social world.* Northwestern University Press.

[98] Alfred Schutz and Fred Kersten. 1976. Fragments on the Phenomenology of Music. (1976).

[99] Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. In *ICCV*.

[100] 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* (Aarhus, Denmark) *(CC '05)*. Association for Computing Machinery, New York, NY, USA, 49–58. https://doi.org/10.1145/1094562.1094569

[101] John Shotter. 1983. Duality of "structure" and "intentionality" in an ecological psychology. *Journal for the Theory of Social Behaviour* 13, 1 (1983), 19–44.

[102] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312. 6034* (2013).

[103] Ben R Slugoski, Mansur Lalljee, Roger Lamb, and Gerald P Ginsburg. 1993. Attribution in conversational context: Effect of mutual knowledge on explanation-giving. *European Journal of Social Psychology* 23, 3 (1993), 219–238.

[104] James F Smith and Thomas Kida. 1991. Heuristics and biases: Expertise and task realism in auditing. *Psychological bulletin* 109, 3 (1991), 472.

[105] Mark Snyder and Phyllis White. 1982. Moods and memories: Elation, depression, and the remembering of the events of one's life. *Journal of personality* 50, 2 (1982), 149–167. https://doi.org/10.1111/j.1467-6494.1982.tb01020.x

[106] William H Starbuck and Frances J Milliken. 1988. Executives' perceptual filters: What they notice and how they make sense. (1988).

[107] Barry M Staw. 1975. Attribution of the "causes" of performance: A general alternative interpretation of cross-sectional research on organizations. *Organizational behavior and human performance* 13, 3 (1975), 414–432. https://doi.org/10.1016/0030-5073(75)90060-4

[108] Simone Stumpf, Vidya Rajaram, Lida Li, Weng-Keen Wong, Margaret Burnett, Thomas Dietterich, Erin Sullivan, and Jonathan Herlocker. 2009. Interacting meaningfully with machine learning systems: Three experiments. *International journal of human-computer studies* 67, 8 (2009), 639–662. https://doi.org/10.1016/j.ijhcs.2009.03.004

[109] William R Swartout. 1983. XPLAIN: A system for creating and explaining expert consulting programs. *Artificial intelligence* 21, 3 (1983), 285–325. https://doi.org/10.1016/S0004-3702(83)80014-9

[110] Shelley E Taylor. 1991. Asymmetrical effects of positive and negative events: the mobilization-minimization hypothesis. *Psychological bulletin* 110, 1 (1991), 67–85.

[111] Bas van Fraassen. 1988. The Pragmatic Theory of Explanation. In *Theories of Explanation*, Joseph C Pitt (Ed.). Oxford University Press.

[112] Francisco J Varela, Evan Thompson, and Eleanor Rosch. 2016. *The embodied mind: Cognitive science and human experience.* MIT press.

[113] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3174014

[114] Timothy J Vogus and Kathleen M Sutcliffe. 2012. Organizational mindfulness and mindful organizing: A reconciliation and path forward. *Academy of Management Learning & Education* 11, 4 (2012), 722–735. https://doi.org/10.5465/amle.2011.0002c

[115] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.* 31 (2017), 841.

[116] 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* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3290605.3300831

[117] K.E. Weick, K.M. Sutcliffe, and D. Obstfeld. 1999. Organizing for high reliability: Processes of Collective Mindfulness. *Research in Organizational Behaviour* 21 (1999), 81–123.

[118] Karl E Weick. 1995. *Sensemaking in organizations.* Vol. 3. Sage.

[119] Karl E Weick and Kathleen M Sutcliffe. 2015. *Managing the unexpected: Sustained performance in a complex world.* John Wiley & Sons.

[120] Karl E Weick, Kathleen M Sutcliffe, and David Obstfeld. 2005. Organizing and the process of sensemaking. *Organization science* 16, 4 (2005), 409–421. https://doi.org/10.1287/orsc.1050.0133

[121] Daniel S Weld and Gagan Bansal. 2019. The challenge of crafting intelligible intelligence. *Commun. ACM* 62, 6 (2019), 70–79. https://doi.org/10.1145/3282486

[122] Mark Wenman. 2013. *Agonistic democracy: Constituent power in the era of globalisation.* Cambridge University Press.

[123] Terry Winograd, Fernando Flores, and Fernando F Flores. 1986. *Understanding computers and cognition: A new foundation for design.* Intellect Books.

[124] Jiaming Zeng, Berk Ustun, and Cynthia Rudin. 2017. Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 180, 3 (2017), 689–722. https://doi.org/10.1111/rssa.12227

[125] Wencan Zhang and Brian Y Lim. 2022. Towards Relatable Explainable AI with the Perceptual Process. In *CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, Article 181, 24 pages. https://doi.org/10.1145/3491102.3501826

[126] Jichen Zhu, Antonios Liapis, Sebastian Risi, Rafael Bidarra, and G Michael Youngblood. 2018. Explainable AI for designers: A human-centered perspective on mixed-initiative co-creation. In *2018 IEEE Conference on Computational Intelligence and Games (CIG)*. IEEE, 1–8. https://doi.org/10.1109/CIG.2018.8490433