Beyond Expertise and Roles: A Framework to Characterize the Stakeholders of Interpretable Machine Learning and their Needs

Harini Suresh
hsuresh@mit.edu
MIT

Steven R. Gomez
steven.gomez@mit.edu
MIT

Kevin K. Nam
kevin.nam@mit.edu
MIT

Arvind Satyanarayan
arvindsatya@mit.edu
MIT

ABSTRACT

To ensure accountability and mitigate harm, it is critical that diverse stakeholders can interrogate black-box automated systems and find information that is understandable, relevant, and useful to them. In this paper, we eschew prior expertise- and role-based categorizations of interpretability stakeholders in favor of a more granular framework that decouples stakeholders' knowledge from their interpretability needs. We characterize stakeholders by their formal, instrumental, and personal knowledge and how it manifests in the contexts of machine learning, the data domain, and the general milieu. We additionally distill a hierarchical typology of stakeholder needs that distinguishes higher-level domain goals from lower-level interpretability tasks. In assessing the descriptive, evaluative, and generative powers of our framework, we find our more nuanced treatment of stakeholders reveals gaps and opportunities in the interpretability literature, adds precision to the design and comparison of user studies, and facilitates a more reflexive approach to conducting this research.

CCS CONCEPTS

• **Human-centered computing** → **User models**; **HCI theory, concepts and models**.

KEYWORDS

interpretability; explainability; machine learning; expertise; knowledge; needs; goals; framework

ACM Reference Format:
Harini Suresh, Steven R. Gomez, Kevin K. Nam, and Arvind Satyanarayan. 2021. Beyond Expertise and Roles: A Framework to Characterize the Stakeholders of Interpretable Machine Learning and their Needs. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3411764.3445088

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

ACM ISBN 978-1-4503-8096-6/21/05.
https://doi.org/10.1145/3411764.3445088

1 INTRODUCTION

Automated systems based on machine learning (ML) and artificial intelligence (AI) are often described as "black boxes" due to the difficulty of extracting the logic behind their outputs in an understandable way. The inscrutability of such systems makes it difficult for them to facilitate effective trust-building and to be held accountable for the decisions they affect. A growing body of recent work has focused on tackling these issues through *model interpretability*, which involves producing visual explanations about a model's behavior for its users. But to design effective interpretability mechanisms, we need to first consider the question: who, exactly, are the stakeholders involved, and what are they trying to achieve?

Take, for example, an ML-based medical decision-making tool. The physicians using the system need to be able to align the output with their own clinical expectations and justify their recommendations to patients. Patients, then, need to have some confidence in the validity of these recommendations, and may want to explain decisions to family members. Other medical staff need to understand the decision-making processes insofar as it affects the treatment they administer to the patient. The developers who created the system should be able to monitor its performance and understand how to make improvements. Physicians and patients, as well, may want and be well-suited to provide feedback about on-the-ground errors the system makes. And there are undoubtedly other people involved in or affected by this system, from external legal agencies to all the people whose data went into training the ML model.

Existing methods for model interpretability, however, often do not explicitly identify or describe their intended user. As a result, many of these methods inadvertently end up being most understandable to the people that build them (i.e., ML researchers or developers). In other cases, the recipient of the interpretability system is described generally as a "layperson" or "end user"; resulting methods may produce simpler visuals, but experimental studies have shown they too often are not useful for people in practice [23, 72, 95, 110]. In our prior example, doctors, patients, and medical staff may all be considered "end users," but have significantly different needs and goals when interpreting, understanding, and reacting to the output of the ML model. Indeed, when it comes down to it, many organizations say they want to give users insight into ML systems through interpretability mechanisms, but these methods are only actually used internally by developers [17].

Part of this disconnect stems from the difficulty in identifying and characterizing different stakeholders and their interpretability needs. A growing body of work has engaged with this problem,

proposing an ecosystem of stakeholders [96, 113], and conducting literature surveys [44, 57, 87, 131] and interview studies [27, 58, 114] to understand their goals. Resultant frameworks typically adopt one of two approaches: they either categorize stakeholders by their expertise (using labels such as "experts", "novices", or "non-experts" [57, 87, 131]) or by their functional role in the ecosystem (e.g., "executives" and "engineers" [17], model "breakers" and "consumers" [58], or model "operators" and "executors" [113]). Stakeholder needs and goals then follow from these categories. While usefully advancing our understanding of the stakeholders involved, these initial frameworks are limited in their descriptive and generative powers [14]. For instance, role-based frameworks implicitly conflate a person's expertise with what they need from the system, with roles often depicted as a relatively static constructs. And expertise-based categories typically portray stakeholders lying on roughly linear scales that only account for cognitive notions of expertise and, thus, do not acknowledge the rich tacit knowledge and lived experience they may possess.

In response, we introduce a framework with a more granular and composable vocabulary to characterize the stakeholders of interpretable machine learning, and their needs. Our framework comprises two components. First, we decompose stakeholder expertise into two dimensions that describe the types of knowledge a stakeholder may possess (i.e., formal, instrumental, and personal knowledge), and the contexts in which this knowledge manifests (i.e., machine learning, the data domain, and the milieu). Second, we define stakeholder needs using a three-level typology of long-term goals, shorter-term objectives that target these goals, and immediate tasks that stakeholders perform to meet their objectives.

To understand the implications of our framework, we assess its descriptive, evaluative, and generative powers — three properties of interaction models first described by Michel Beaudouin-Lafon [14]. We code 58 papers describing interpretability systems or users, and find that our framework is consistently able to describe stakeholders' knowledge and interpretability needs while adding granularity and drawing new connections between them. We describe how our framework's abstractions can allow us to design more precise application-grounded evaluations [38], including bringing precision to participant recruiting and providing a structure to operationalize comparative studies. And finally, we demonstrate that our framework generates a rich intersection of user expertise and needs for study, and can also be turned inwards to facilitate a more reflexive design process.

2 BACKGROUND AND MOTIVATION

Researchers have recognized that more precisely defining "interpretability" or "explainability" is a key challenge for the field [80]. Although some work seeks to develop formal or technical definitions of interpretability [36, 38, 47], and though the burgeoning set of interpretability techniques often do not name specific target users or tasks [9, 42, 100, 133, 137], there is a growing recognition for the need to approach this problem space in a human-oriented manner. In this section, we motivate our contribution by surveying prior work and describing the limitations we observe with current approaches for defining the *why* and *who* of machine learning interpretability.

In early definitions, Lipton [80] and Doshi-Velez & Kim [38] identified that the need for interpretability primarily stems from a mismatch between the formal definition of the machine learning model, and its output and real-world impact. Lipton further expanded on this need by enumerating a set of desiderata for interpretability including building trust in the model, inferring causal relationships between the input and output, improving model tranferability and generalizability, providing introspection, and finally to facilitate fair and ethical decision-making. Others have since contributed to this list in a variety of ways including detailing how different interpretability methods can be chosen to mitigate particular cognitive biases [119], proposing taxonomies of questions used to arrive at an appropriate interpretability method [7], discussing how applications with different contexts or levels of automation might necessitate different design decisions [77, 106], and grounding the need for explanations in the social sciences [83].

Most relevant to our paper is a body of work that seeks to better define interpretability by studying the specific users involved. In surveying this work, we identified two distinct approaches to doing so. First are a group of papers that characterize users based on their expertise. For instance, both Yu & Shi [131] and Hohman et al. [57] classify users on roughly linear scales of machine learning expertise (from beginner to expert for Yu & Shi, while Hohman et al. adopt the terms "model developers and builders," "model users," and "non-experts"). Similarly, Mohseni et al. [87] identify "AI Novices" and "AI Experts" and add "Data Experts" to the mix. With all of these schemes, a user's needs then stem from their expertise. For instance, novices are typically described as needing educational or teaching tools, whereas experts require tools for debugging or deploying models, or assessing model performance.

The second category of papers characterize users based on their functional role instead. For instance, Tomsett et al. [113] posit an ecosystem of stakeholders including model creators, operators, executors and examiners, as well as the decision and data subjects that are affected by the model or whose data the model was trained with, respectively. Similarly, through semi-structured interviews, Bhatt et al. [17] identify four categories of stakeholders (executives, ML engineers, end users, and others) while Hong et al. [58] identify model builders, breakers, and consumers. Across this work, the role a person inhabits within an organization (or the role they play during the human-AI interaction) determines their interpretability needs. For example, model creators/builders/engineers are said to want introspection of the level of individual instances and features, model operators/breakers may wish to monitor the performance of the model including authoring test cases, and finally model executors/executives/consumers want to be able to have confidence and trust in the model. Cai et al. [27] and Tonekaboni et al. [114] follow this approach of role-based needfinding as well by interviewing clinicians.

While both expertise-oriented and role-oriented frameworks have usefully brought further definition to the problem of machine learning interpretability, we can observe limitations to their descriptive and generative powers (i.e., the degree to which they describe *existing* points, and help us identify *new* points in the problem space, respectively [14]). Role-based frameworks, for instance, do not break the problem space down into sufficiently granular and composable units. As a result, several roles appear to implicitly

conflate expertise and interpretability needs — for example, model "creators" are likely most expert with machine learning, and thus need debugging tools at the level of individual instances or features; but, one could imagine "auditors" appreciating insight at this level of abstraction even if they lack an equivalent level of machine learning expertise. Similarly, consider the domain of clinical diagnoses: model "consumers" could equally describe doctors and patients despite these users likely requiring different explanations of the model's output as a result of different levels of medical expertise. Here, model "executors" does not provide much more precision as both doctors and patients are tasked with making decisions informed by the model — doctors about what treatment to prescribe, and patients about whether they do indeed wish to proceed with the treatment. Finally, although most role-based frameworks explicitly note that roles are not mutually exclusive (i.e., a single role may map to more than one individual, and one person may play several roles), roles are nevertheless depicted as relatively static constructs. Not only might an individual user's role change over time but, even if they remain in the same role(s), their interpretability needs may change through repeated exposure to and increased familiarity with the models they are working with, or the situations in which these models are deployed.

Expertise-based frameworks exhibit similar limitations. In particular, a key concern is how these frameworks portray expertise as a linear scale from "novice" to "expert." Several external literatures have articulated concerns with this framing of expertise. For instance, in critiquing the influential Dreyfus linear model of skill acquisition [40], Dall'Alba & Sandberg note that *"[s]tage models of development appear to assume we know what skillful performance entails for each area of skill"* and that the *"focus on stages veils more fundamental aspects of development; it directs attention away from the skill that is being developed"* [34]. Moreover, Dall'Alba & Sandberg point to the fact that such models are primarily concerned with cognitive development and fail to acknowledge expertise gained through embodied practice of a skill [34]. We see a form of this latter critique in the literature on participatory design as well, which advocates that all stakeholders in a design process possess valuable expertise through their lived experience and tacit knowledge [69, 108]. Finally, although recent frameworks usefully consider domain expertise in addition to machine learning expertise, such a clean decoupling does not account for the ways expertise may transfer. For example, Cai et al. find that while medical practitioners express a desire for an "AI primer", they are nevertheless able to bring some of their training and experience working with other clinical technologies to bear — for instance, in understanding that the output of a model will not be perfect, or in enumerating "test cases" for an AI assistant [27]. Similarly, as AI/ML-enabled technologies increasingly permeate every day life, this ubiquity and familiarity will shape users' interpretability needs in ways that current expertise-based frameworks leave unaddressed.

And, across the two types of frameworks, interpretability needs or goals are determined primarily by the category a user falls within. While many frameworks allow for categories to overlap, this approach nevertheless obscures the fact that many goals can cut across several roles or expertise. For instance, almost every stakeholder involved will likely want to have trust in the model, and want to be able to assess the degree to which it may be biased — we see explicit evidence for this for machine learning experts [57] and data experts [87], model creators and breakers [58], model operators [27, 114], as well as for decision- and data-subjects who may wish to contest a decision or otherwise seek recourse [6, 69]. Similarly, while current frameworks primarily pose debugging and improving the model as goals model creators, builders, or any other traditionally-"expert" stakeholders may have, one could imagine that activists and other groups with non-traditional expertise may also wish to assess the outcome of domain-specific test cases.

In summary, recent work has recognized that better defining the problem space is a key challenge for machine learning interpretability, and has advanced our understanding by contributing frameworks for describing the stakeholders involved and their goals or needs. However, in analyzing the descriptive and generative powers of these frameworks, we see several limitations. In particular, by not providing a sufficiently granular or composable vocabulary, existing frameworks poorly distinguish the rich intersection that exists between attributes of the stakeholder (e.g., their expertise), the role that they may play (e.g., model creator or consumer), and their ultimate goals or needs with regards to interpretability (e.g., debugging the model, or building trust).

3 A FRAMEWORK TO CHARACTERIZE THE STAKEHOLDERS OF INTERPRETABLE ML

To develop a more granular and composable vocabulary for describing the stakeholders of interpretable machine learning, we engaged in an iterative process with alternating phases to diverge and converge our thinking. In particular, we began by surveying the literature on interpretability summarized in the previous section, and extracting passages that described users and stakeholders, as well as their needs, actions, and goals. To diverge our thinking, we looked to domains outside of interpretability and computer science, including the literatures on expertise and pedagogy [18, 34, 43, 45, 52, 53, 64, 120, 123, 128, 129], critical theory [6, 74, 85, 90, 111], law [31, 39, 55, 118, 132], and participatory action research [50, 60]. To converge our thinking, we reflected on how concepts from these external domains could be adapted within interpretability. This reflection process involved alternating phases of open coding to map external concepts to the passages we had initially extracted, affinity diagramming to identify recurring groupings and patterns between codes, analytic memo writing, and weekly hour-long conversations between all authors.

Our framework comprises two halves. First, it describes the knowledge stakeholders may possess and the contexts this knowledge may manifest in. And, second, it enumerates the long-term goals stakeholders may have, and breaks these goals down into shorter-term objectives and specific tasks they can perform.

3.1 Decomposing Stakeholder Expertise into Knowledge and Context

Prior work has identified, either explicitly [57] or implicitly via roles [58, 87, 113], that expertise is a defining attribute of interpretable ML stakeholders. To provide a more granular treatment

Table 1: Examples of how the three types of knowledge manifest in the three contexts described by our framework.

	Knowledge		
Context	**Formal**	**Instrumental**	**Personal**
ML	The math behind model architectures, optimization and training processes, etc.	Familiarity with ML toolkits, off-the-shelf models, etc.	Tricks of the trade (e.g., hyperparameter values, feature engineering, etc.)
Data Domain	Theories relevant to the data domain (e.g., symptoms and treatments, case law and legal precedent, etc.)	Experience working with other related technology (e.g., medical devices, document mining tools, etc.)	Lived experience (e.g., prior memories of similar events)
Milieu	Sociocultural theories (e.g., redlining, gerrymandering, mass incarceration, etc.)	Familiarity with broader ML-enabled systems (e.g., virtual assistants, recommendation algorithms, etc.)	Lived experience and cultural knowledge (e.g., values, attitudes)

of expertise, we adapt models of expertise from Fleck [45] and Eraut [43] to decompose a singular notion of expertise into three constituent types of *knowledge*. **Formal** knowledge comprises an understanding of codified theories, embodied in text or diagrams such as those found in textbooks, and is acquired through a prolonged educational process. **Instrumental** knowledge is an understanding of how to "apply" formal knowledge. It is embodied in the use of tools or other instruments, and is learnt through demonstration and practice. Finally, **personal** knowledge describes information that is entirely embodied in individual people, and is gained through their participation in specific domains. It is difficult to codify [43] as it consists of a person's lived experience (e.g., memories of specific events, self-knowledge about the way they may react in certain scenarios, etc.) as well as values that may be distributed in the cultures and societies they are a member of.

These types of knowledge manifest in *contexts*, or the domains or situations that determine what knowledge is relevant. We identify three contexts: **machine learning**, or the knowledge required to research, develop, operate, or deploy machine learning models; the **data domain**, or the knowledge necessary to collect, organize, analyze, and communicate the data the model was trained with or makes decisions about; and **milieus**, which refer to the environments that the human-AI interaction may be occurring within. These environments include both the physical surroundings (e.g., a home, bank, courthouse, doctor's office, etc.) as well as the broader sociocultural context (e.g., mass incarceration, redlining, gerrymandering, etc.).

Our framework provides a more expansive yet precise treatment of expertise in interpretable ML. While prior expertise- or role-based approaches latently encode notions of formal and instrumental knowledge, by explicitly articulating these concepts, our framework facilitates teasing apart differences and understanding the implications on interpretability design. For example, "model users" [57] and "model breakers" [58] cover an extremely broad range of possible stakeholders including model architects, trainers, engineers, data scientists, and machine learning artists [57], as well as domain experts, product managers, and auditors [58], respectively. These categories appear primarily focused on stakeholders' instrumental knowledge and, by analyzing contexts, we can separate machine learning instrumentalists (model architects, trainers) from data domain instrumentalists (artists, domain experts, product managers), and those that may span the two (data scientists, auditors). Doing so suggests that these stakeholder groups may have different interpretability needs that the broader categories of "model users" or "model breakers" obscured. For instance, perhaps interfaces for machine learning instrumentalists should be articulated in terms of the components exposed by popular toolkits. Similarly, for data domain instrumentalists, how might we analogize interpretability to tools and systems that they already work with in order to enable expertise transfer (akin to how Cai et al. found medical practitioners reasoning about uncertainty [27])?

Moreover, our framework explicitly recognizes the personal knowledge stakeholders may have — including "tricks of the trade" a person may have acquired, their experiences and memories, or the more distributed values of the cultures and societies they are a member of — as an important consideration when designing for interpretability. Critically, by placing it alongside formal and instrumental knowledge, our framework identifies it as an *equally important* form of knowledge. As a result, one might consider designing *for* stakeholders' personal knowledge — for instance, using example-based explanations such that a stakeholder might better "see themselves" in the data [2, 94, 99]. But our framework also suggests designing *with* stakeholders, to better account for personal knowledge that designers do not have — a position advocated for by various communities including participatory action research [50] & design [69, 108]. For instance, members of the general public might have different notions of what constitutes an "error" based on their personal knowledge [51].

Our framework also highlights that interpretability design must attend to more than the immediate contexts of machine learning and the data domain — explanations must be situated in stakeholders' milieu. Here, we draw an analogy to data visualization. Researchers and data journalists consider annotations to be a crucial component of effective visualization design because it helps readers understand the broader context associated with the visualized data. As Amanda Cox, Data Editor for The New York Times, says, *"the annotation layer is the most important thing we do ... otherwise it's a case of here it is, you go figure it out* [33]. We believe this property holds true for interpretability as well — it is insufficient for an explanation to be articulated purely in terms of the model or data if it misses critical aspects of the milieu. For instance, consider an ML-backed loan

evaluation system: explanations in the ML context would articulate the output decision in terms of model components, while explanations in the data domain might also discuss distributions in the training or test set and how this may lead to biased output. However, under our framework, we would consider these explanations to be incomplete if they were not situated the broader sociocultural milieu — for instance, how disparities in data distributions have occurred through policies such as redlining, or in the difficulty ex-offenders have in finding employment.

Finally, by decoupling knowledge and context as two orthogonal dimensions of the problem space, our framework enables a more systematic analysis of the stakeholders of interpretability. It eschews prior easily-quantifiable linear scales in favor of more descriptive treatments of expertise. Designers can work with each dimension individually — for example, how might interpretability help stakeholders formalize their personal knowledge in the data domain by scaffolding example-based explanation with featured-based saliency methods akin to faded worked examples [8]; or, as described previously, how might instrumental knowledge in the data domain transfer to the machine learning context [27]. And, by considering the intersection of the two dimensions as well, our framework can help us identify the ways in which expertise recurs in the interpretability ecosystem.

3.2 Distilling Stakeholder Needs into Goals, Objectives, and Tasks

Through our open coding and reflection process, we distilled a three-level typology of interpretability needs. The first level identifies two long-term *goals*: **understanding the model (G1)** and **building trust in the model (G2)**. These goals are high-level and difficult to define precisely, but we include them in our framework to acknowledge that they underlie almost every single piece of work we read. We do not expect stakeholders to be able to directly accomplish these goals, nor do we imagine that future methods or systems will address them squarely. Rather, these goals function like substrates which inform and influence the two lower levels of needs we describe below. For instance, we expect stakeholders to develop an understanding of models over time — through repeated exposure and interactions. And much work has framed trust as something society as a whole needs in order to accept new technologies [21, 118, 132] — indeed, trust may grow as stakeholders better understand the model, but may also develop in a proxied or deferred fashion through increased regulations and standards.

The second level of our typology describes the shorter-term *objectives* stakeholders might target to achieve their longer-term goals. We give real examples for each objective, and demonstrate how they can be relevant to a diverse range of stakeholders. These objectives are grounded in stakeholders' current real-world needs, but as ML tools continue to be deployed in new domains, we expect this typology will continue to evolve.

(O1) Debug or improve a model. The objective of improving a system or correcting its mistakes appears frequently in the literature, and is often posed in terms of the needs of developers [21, 54, 76]. For example, Bhatt et al. [17] describe how internal members of organizations try to use interpretability techniques to uncover inconsistencies between the model's logic and their intuition or expectations, in order to guide further improvements. However, it is critical to acknowledge that developers are not the only stakeholder group to which this need applies. For instance, Tonekaboni et al. [114] and Zarsky [132] both highlight the value of allowing a larger group of stakeholders, including the general public, to provide feedback for improving systems. Indeed, theories from Participatory Action Research also hold that people on the ground in a specific context are often much better suited to realize errors and devise appropriate fixes, as opposed to developers for whom the errors typically have no direct consequence [50].

(O2) Ensure compliance with standards or regulations. Auditing, or ensuring that the development, deployment, and results of a certain system are compliant with a particular set of standards (whether they are legal, ethical, safety, or other) is already necessary in other areas such as finance or aerospace, and is emerging as an important objective for ML systems as well. The introduction of the GDPR, for example, has established a set of legal standards that automated systems must comply with. And it is not only external watchdog agencies or governments that are interested in ensuring such compliance. Individuals or groups within an organization may also have their own internal standards they want to ensure are met — for example, Raji et al. [97] describe the design of an internal auditing pipeline.

(O3) Understand how to incorporate the model's output into downstream actions. Several prior papers mention the need for guidance on whether and how to incorporate model predictions into further actions — whether that involves relating the model's output to relevant and actionable decisions, or understanding how much weight to place on the model's prediction [17, 21, 58, 76, 114]. This objective emerged as important for a number of different types of stakeholders, such as doctors using a diagnostic aid [58, 114], people applying for health insurance that involves automated screening [17], or for those subjected to automated decisions more generally [118].

(O4) Justify or explain actions influenced by a model's output. Through interviews with Intensive Care Unit and Emergency Department clinicians, Tonekaboni et al. [114] describe clinicians' desire to justify decisions influenced by a model's output to patients or colleagues. Similarly, by interviewing the head of AI at a bank using automated credit evaluations, Hong et al. [58] identify the need to justify to customers decisions that were influenced by the model. In addition to the immediate stakeholders acting on the model output (*executors* according to Tomsett et al. [113]), this goal can also stem from people about whom a decision was made (*decision subjects*). For example, Zarsky [132] frames the need to provide someone with an explanation of a decision or action that affects them as one that is necessary in order to respect their autonomy.

(O5) Understand how one's data is being used. Zarsky [132] grounds this objective in the theoretical premise of information privacy rights, framing it as an extension of individual autonomy. The need to have control over one's personal data

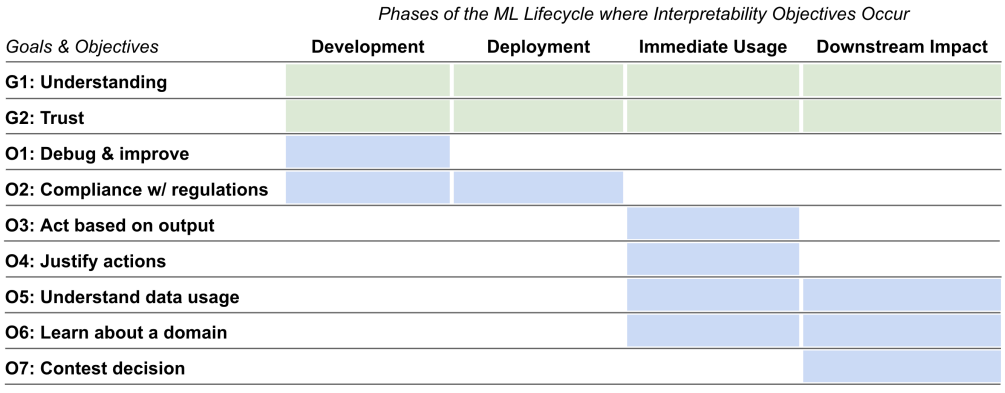

Figure 1: A visualization of the latent chronology in goals and objectives. Categories along the horizontal axis are relevant phases of the ML process. Colored cells indicate the phase in which a particular goal or objective typically occurs. Phases need not occur linearly, may be iteratively revisited, and many different stakeholders may be involved at any given phase.

has also been broadly accepted in European data protection law. Hildebrandt [55] makes a further argument that people should understand not only what data about them is being utilized, but the potential consequences of this usage as well. And Buneman et al. [22] distinguish between different types of data provenance users may be interested in. Clearly, given the prevalence of data mining, this objective is relevant to a wide range of stakeholders.

(O6) Learn about a domain. Through interview studies, both Hong et al. [58] and Liao et al. [76] describe how stakeholders across different domains use interpretability to generate new hypotheses or insights about a domain. For example, one participant aimed to use a model predicting surgeons' future performances as a tool to better understand what factors drive good performance, rather than using it as a predictive system. Hohman et al. [56] focus specifically on data scientists, describing how interpretability helped them find "valuable nuggets of information" in the data. Similarly, Doshi-Velez and Kim [38] identify the use of interpretability to advance scientific understanding. Indeed, there is a growing subfield of machine learning investigating how interpretability mechanisms can aid in scientific discovery [1].

(O7) Contest a decision made based on the model's output. Citron and Pasquale [31] posit that the right to challenge a decision affecting oneself should be ensured under due process, and Doshi-Velez et al. [39] draw a comparison to the legal system, where mechanisms for redress serve as a powerful form of accountability. Wachter et al. [118] also notes that the right to contest an automated decision is provided in Art. 22(3) of the European General Data Protect Regulation (GDPR). An individual affected by an algorithmic system may not be the only one who wishes to contest it, either. We can imagine that external stakeholders like lawyers, judges, or activists may also be interested in pushing back against model outputs that seem incorrect, arbitrary, or unfair.

The third and final level of our typology identifies the specific *tasks* a stakeholder can perform to achieve the goals described above. We break tasks out as a separate level of the typology to make clear that tasks do not map to objectives in a one-to-one fashion; rather, the same task may be used to accomplish several different objectives. For instance, detecting discrimination or other undesirable behavior in a model's prediction is likely to be a necessary task for both contesting a decision (O7) but also for understanding whether or how to incorporate model output in downstream actions (O3). Although the task is shared across these objectives, the specific type of discriminatory behavior a stakeholder may wish to detect, and the manner in which is it exposed and communicated, may differ based on the domain, the higher-level objective, and the stakeholder's knowledge. Here, we describe several such underlying tasks that we found to recur in the literature and give examples of how they can be relevant for multiple objectives. As with objectives, we expect this level of the typology to grow as ML continues to be deployed in new situations.

(T1) Assess reliability of a given prediction. Understanding the reliability of a given prediction is important for deciding how (or whether) to incorporate the model's output into further actions (O3), to prevent harmful outcomes or overreliance [23, 135]. Similarly, the ability to assess a given prediction and show, for example, that it may not have been reliable, is likely to provide important evidence for contesting a decision (O7).

(T2) Detect mistaken, discriminatory, or arbitrary behavior. The ability to detect discrimination or other unwanted logic codified in a model is considered a crucial tool for being able to contest an automated decision (O7) [39]. Similarly, ensuring that predictions are not being made arbitrarily is likely necessary to ensure compliance with ethical or legal standards (O2). In other cases, detecting incorrect reasoning was a way to guide model debugging and elucidate areas for improvement (O1) [27]. Some papers also frame this task as its converse, i.e., verifying that predictions are sensible and/or fair (by some definition) [101].

(T3) Understand the extent of the information the model is using. Understanding details and extent of features used

emerged as important for explaining actions influenced by the model (O4). Tonekaboni et al. [114], for example, describe how doctors felt that understanding the clinically relevant model features that were used was critical to first rationalizing the predictions to themselves, and then explaining them to patients. Depending on the context, recognizing higher-level groups of features (e.g., "demographic information," "patient medical history") may be more understandable and feasible than individual features. We can also imagine that developing this understanding will also be an important way for stakeholders to identify what aspects of their personal data are being incorporated into a specific system (O5) [55].

(T4) Understand the influence of different factors on the model's output. For stakeholders who are interested in generating new insights about a domain (O6), understanding how different factors influence the output is key. Roscher et al. [102] provide several examples of deriving scientific or medical insights by investigating the impact of scientifically-meaningful factors on predictive outcomes. This task is also important for ensuring compliance with particular standards or regulations (O2), which may detail when/how it is acceptable to use certain features. Unlike T3, this task may not provide a comprehensive understanding of the features used (e.g., perhaps just listing the most important) and is more focused on the ways those features influence the output.

(T5) Understand model strengths and limitations. Understanding the model's overall potential weaknesses is critical for understanding how to incorporate its output into further actions (O3). For example, Cai et al. [27] describe how doctors consistently wanted to know the proposed AI tool's specific limitations so that they could anticipate and account for them during decision-making. Understanding areas of weakness is likely to also be useful for debugging and improving the model (O1), e.g., by guiding additional data collection or training.

Note that specific implementations (e.g., counterfactual explanations [118]) are not included at the task-level; rather, they are used to *implement* a particular task. For example, counterfactual explanations might be one way to implement the task "detect discriminatory behavior" (T2), but might be more or less appropriate depending on the stakeholder's knowledge, their overarching objective, and the surrounding context. Section 4.3 (Generative Power) further discusses the implications our framework might have on choosing particular methods.

While several prior literature surveys have sought to collate and organize a list of interpretability needs, our framework makes some key advances to provide a more nuanced understanding these needs. First, where prior surveys focus primarily on computer science subdisciplines [44, 57, 87], our framework incorporates these insights and extends them by looking to the legal literature [31, 55, 118, 132] and research on participatory action and design [50, 108]. As a result, our framework is able to surface objectives such as "contesting a decision" (O7) or "understanding how one's data is being used" (O5) that prior surveys did not identify.

Second, and more importantly, where prior approaches define interpretability needs as a function of stakeholder expertise or role, our framework defines these needs as an independent component of the problem space. As a result, and as the examples above illustrate, our framework helps reveal that interpretability needs can cut across several different stakeholders. For instance, model debugging (O1) is one of the most frequently identified interpretability goals; but, prior work has primarily categorized it as a need machine learning experts (or model builders and developers) have. In contrast, our framework identifies that although certain stakeholders may not have much formal or instrumental machine learning knowledge, their personal knowledge may be crucial for identifying or fixing model errors. Similarly, while it may have previously been tempting to think that contesting a decision (O7) is a need primarily expressed by decision subjects, our framework highlights that other stakeholders (including lawyers, judges, and activists) may wish to do so as well to affect systematic change.

Finally, in contrast to the uniform treatment of prior interpretability surveys, our framework provides new levels of abstraction for discussing interpretability needs. In doing so, we can distinguish that these needs form a hierarchy: immediate tasks help stakeholders accomplish short-term objectives which, over time, achieve long-term goals. As with other multi-level typologies [20], this structure surfaces the compositionality latent in this space. For instance, as described above, there is a many-to-many relationship between goals, objectives, and tasks: one task may apply to several objectives; many tasks may be required to accomplish a single objective; and, together, they are all necessary to achieve goals. Similarly, our three-level sequence allows for describing interpretability needs as sequences of action. For example, to improve a model (O1), a stakeholder may wish to understand its strengths and limitations (T5) by repeatedly assessing the reliability of individual predictions (T1); or, a stakeholder's trust in the model (G2) may increase or decrease as a result of better understanding how it works (G1).

While we do not ascribe objectives to specific roles or expertise levels as in prior work, we note that they nevertheless exhibit a latent temporal structure — for example, the need to understand how a model's output should be incorporated into a decision (O3) occurs before someone wishing to contest that decision (O7). However, formalizing this latent chronology is not straightforward, as a given objective may (re)occur at several different stages during the ML process. And, there is a risk of unintentionally recapitulating prior stakeholder categorizations as particular roles or expertise may be implicitly associated with different stages of the ML process.

As a result, the chronology we settle on, shown in Figure 1, is more flexible and refers to broad phases of the ML process. Rather than provide a precise ordering, it is meant to lend some helpful structure to the many stakeholder objectives. We indicate the phase(s) in which a particular objective typically occurs, and note that these phases are likely to unfold iteratively. For example, the development and deployment stages may be revisited after observing a system's downstream impact. Furthermore, many different stakeholders may be involved in each phase. For example, beyond engineers with formal ML or data knowledge, downstream users with significant personal knowledge may provide input to the development phase of a particular system if they report bugs or provide feedback that is used to retrain the model. We omit tasks from this chronology to preserve the many-to-many mapping between objectives and tasks.

Table 2: Knowledge types and contexts for interpretability stakeholders, with examples identified in our literature survey. We found a range of expertise and backgrounds under our framework, highlighting information that might be lost if XAI designers only consider a small set of roles like "ML expert" and "non-expert", as we have widely observed in past work.

	Knowledge		
Context	**Formal**	**Instrumental**	**Personal**
ML	model developers [107], computer science students [122], machine learning scientists and researchers [38, 83, 87, 101], model builders/engineers [17, 19, 44, 98, 102, 113, 127], model analyst [63], general [131] **(15/58)**	model users [107], data scientists/ML practitioners [7, 61, 122], autonomous robot developers [112], data scientists [3, 17, 21, 58], "medical experts' increasing familiarity with [computer-aided diagnosis] systems" [27], "domain experts who use machine learning for analysis" [44, 87, 114], practitioners [76, 131], optimization expertise [19], students with some ML familiarity [24], greater machine learning community [47], developers/implementers [48, 98, 105, 113, 121] **(23/58)**	"intuition of how the network looks" [131], "[ML] researchers' intuition of what constitutes a 'good' explanation" [83] **(2/58)**
Data Domain	biologists [107], robotics [112], scientists [102], doctors [27, 38, 63, 109, 117, 126], "enrolled in law school" [73], judges [12], agronomic engineers [19], regulators [17], HR managers who produce expert estimates [115], energy data operators [16], "business logic" [113], game theorists [121], general [11, 101] **(19/58)**	"model novices" interested in applying ML to specific domains [107], "deep knowledge of the circumstances for employee retention" [7], sign-language learners [93], domain knowledge to verify ML results qualitatively [104], "only specialists in part of the underlying process" [19], internal financial auditors [97], clinicians [27, 38, 105, 114], "increasingly adopt ML for optimizing and producing scientific outcomes" [102], operators [113, 130], peer grading in online education [65], general [29, 58, 115] **(17/58)**	"without accurate mental models, social factors can rationalize suspicious observations [about explanations]" [61], "how well the system's conceptual model fits their mental model" [30], mental models of the system to generalize the AI behavior [83, 130], patient/client/decision subject [21, 101, 105, 113, 117, 121], "hold preconception of what constitutes useful explanations for decisions" [76], "prognosticating their patient's condition in their personal experience" [114] **(12/58)**
Milieu	students studying information policy [122], ethicists [98], bodies like institutional review boards or ethics committees [117], "understanding requirements arising from social contexts other than just from usability or human cognitive psychology" [3] **(4/58)**	"community hospital small groups, to academic medical centers" [27], "use AI products in daily life" [87], UX/design practitioners [76, 127], data subject [98, 113], product managers [58], examiners/auditors [113], departments adopting decision-support technologies [114], use of AI in government and industry [21] **(10/58)**	loan applicant [7, 116], "different cultural, demographic or phenotypic groups" [97], recommender system users [68, 136], "actors bring their own points of view and own priorities" [125], "people employ certain biases and social expectations" [83], "anticipating the situated, user-encountered capability of AI is difficult" [127], familiarity with privacy and personal data issues [44], individual fatigue and workflow issues in heathcare [114], general [12, 105, 115] **(13/58)**

4 EVALUATION & EXAMPLE APPLICATIONS OF THE FRAMEWORK

To assess the implications of our framework, we look to the three powers of interaction models described by Beaudouin-Lafon [14]: the *descriptive* power, or how much coverage the framework achieves over existing points in the problem space; the *evaluative* power, or how well the framework helps us compare two points in the problem space; and, the *generative* power, or how the framework helps us envision new or previously unexplored points in the problem space. In addition to an evaluation of the framework, the evaluative and generative powers also serve as a demonstration of ways in which the framework can be used.

We find that our framework is able to describe over 50 existing papers on interpretability, and further provides a more granular treatment of relevant stakeholders. We then illustrate how our framework gives us a language with which to more carefully evaluate interpretability systems. Finally, we demonstrate how our framework can be used to generate new combinations of personas and needs, how it may suggest ways of designing future interpretability interfaces, and how it may be turned inwards to facilitate a more reflexive design process.

4.1 Descriptive Power

We assess our framework's descriptive power by using it to characterize the users and goals described by existing work on interpretable ML. We collected papers using a mix of explicit keyword searches in academic search engines and libraries (e.g., Google Scholar and arXiv), following the citation graph of collected entries, and by compiling the bibliographies of previous literature surveys [44, 57]. Our final list of papers span several research contribution types [75, 124] including frameworks that define interpretability desiderata or key considerations (e.g., Arya et al. [7], Lipton [80], Tomsett et al. [113]), evaluations of specific interpretability techniques (e.g., Balog and Radlinski [10], Cai et al. [26], Cheng et al. [29]) and user/case studies that provide insights into pertinent human factors in interpretability (e.g., Hong et al. [58], Liao et al. [76], Tonekaboni et al. [114]). We excluded any papers that introduced novel interpretability techniques without discussing target users or user-centric considerations (e.g., [133, 134]). Similarly, we excluded papers that included only a passing reference to user characteristics (e.g., "interpretability is important for doctors") without explicitly discussing them. We aimed collect a representative sample of current interpretability research directions, going beyond the papers we used to initially develop the framework, and sought out references across different applications, data domains, and computer science disciplines including machine learning, data

Table 3: Goals, objectives, and tasks for interpretability stakeholders. Our literature survey identified instances of theoretical and systems work that discuss or address these needs.

Stakeholder Need	References
G1: Understanding the model	"machine models" [86, 122], "understand the agent's behavior and responses enough to participate in the mixed-initiative execution process" [48], "to attain scientific outcomes with ML one wants an understanding" [102], "understand the 'algorithmic decision model'" [29], general [3, 11, 21, 26, 27, 35, 38, 47, 89, 121, 130] **(16/58)**
G2: building trust in the model	mechanisms for steering trust building [107], build appropriate trust [12, 24], mechanistic interpretation needed for trust building [117], trust for tool adoption and continued use [65], ensure that ML models reflect appropriate values [63], general [3, 10, 21, 25, 26, 29, 35, 44, 47, 48, 58, 73, 76, 80, 83, 87, 89, 113, 114, 130] **(26/58)**
O1: Debug or improve a model	model refinement [63, 107], help data experts to tune ML parameters for the data [87, 109, 131], "identify issues with a model and how to fix it" or debug and optimize [57, 58, 89, 103, 104, 112, 121, 122], improve an aspect or part of a system [44], general [5, 17, 30, 76, 98, 101, 102] **(21/58)**
O2: Ensure compliance with standards or regulations	adherence to standards and laws like GDPR and "right to explanation" [5, 38, 44, 47, 68, 97, 101, 103], forensics [113], justify clinical validation of ML in medical studies [117], facilitate monitoring for safety standards [121], general [17, 21, 80, 98, 102, 107] **(17/58)**
O3: Understand how to incorporate the model's output into downstream actions	learn "factors that could be changed to improve their profile for possible approval in the future" [7], learn how to correct actions based on model feedback [93], apply own domain-related decision-making using the XAI or not [24], make better or faster decisions [89], understand impact of prediction on other system components [76], understand how to get a desired outcome [116], understand consequences [101], understand errors for safety-oriented task [38], directing use in patient or medical work practice [109, 114, 126], general [10, 115] **(13/58)**
O4: Justify or explain actions influenced by a model's output.	justify the user's decision-making [7, 107, 114], reason about data outputs [61], explaining findings to collaborators [19], "enables the user to consider contrastive explanations... why one decision was made instead of another" [25], explain causes of an event [3, 83, 121], recommend treatment options to patient [113], "justify the result" [5, 24, 26, 136], general [12, 27, 35, 57, 58, 102, 131] **(21/58)**
O5: Understand how one's data is being used	"disclose what user data is being used in algorithmic decision-making" [87], know how one's data is being used to make decisions about others [113], understand why certain user data is collected [136], general [7, 30, 80] **(6/58)**
O6: Learn about a domain	learn about sign language and how to use it correctly [93], learning about ML [131], explanation "as a vehicle to generate insights about the phenomena described by the data" [58, 80], learn how to solve a task [105], learn game strategy (Go) [103], learn new facts/gain knowledge [5, 38, 101], learn design strategies [16] **(10/58)**
O7: Contest a decision made based on the model's output	"when I see things I don't completely agree with" [27], "present an incontestable subset of reasons to the bank employee" [30], contest a discriminatory decision [116], general [80, 113, 121, 125] **(7/58)**
T1: Assess reliability of a given prediction	identify and explain an outlier [19], increase or decrease trust in the model based on observed accuracy, relative or not to one's own performance [12, 130], "to ensure the scientific value of the outcome" [102], assess the AI's judgment [76] **(5/58)**
T2: Detect mistaken, discriminatory, or arbitrary behavior.	"anticipate ethics-related failures before launch" [97], bias or mistake detection [89, 101, 103, 125], understand skewness and biases in input data [35], find unknown vulnerabilities and flaws [5, 38] **(8/58)**
T3: Understand the extent of the information the model is using	data entanglement [97], be informed when the ML is not suitable for particular systems [102], understand "what the system was sensing to make its inferences" [29, 48, 115, 136] **(6/58)**
T4: Understand the influence of different factors on the model's output.	explore counterfactuals and how changes to data points affect predictions [122], understand model prediction mechanisms [29, 58], "factors influencing their individual decision" [26, 30, 116], "inspect how output changes with instance changes" [76, 80], "how drift in feature distributions would impact model outcomes" [17], "did the factor 'race' influence the outcome of the system" [98], "feedforward can help people understand and predict what is going to happen" [3] **(11/58)**
T5: Understand model strengths and limitations	understand model error from predictions [11], know when to trust the prediction or be skeptical [12, 58], understand limitations [76], "clarity around why the model under-performs" [114] **(5/58)**

visualization, human-computer interaction, and scientific computing.

In total, we selected 58 papers, and each paper was coded by at least two authors of this paper. Each coder used the framework to identify instances of stakeholder knowledge types and contexts, as well as goals, objectives, and tasks. When the coders disagreed on a designation for the entry, they discussed the conflicts until there was agreement on the code. Where possible, we collected snippets of the papers corresponding to a description or discussion of stakeholder knowledge or needs. The outcome of this coding process, including snippets[1], is shown in Tables 2 and 3.

We found that the vocabulary provided by our framework was able to describe stakeholders and needs that appeared in prior work. All knowledge type-context intersections and goal/objective/task categories appeared in more than one paper. The most observed knowledge categories were ML-Instrumental (23 of 58 papers), Data Domain-Formal (19/58), and Data Domain-Instrumental (17/58),

[1]Quotations in the snippets may be paraphrased, and should be interpreted as describing a common theme in cases where multiple references are grouped together.

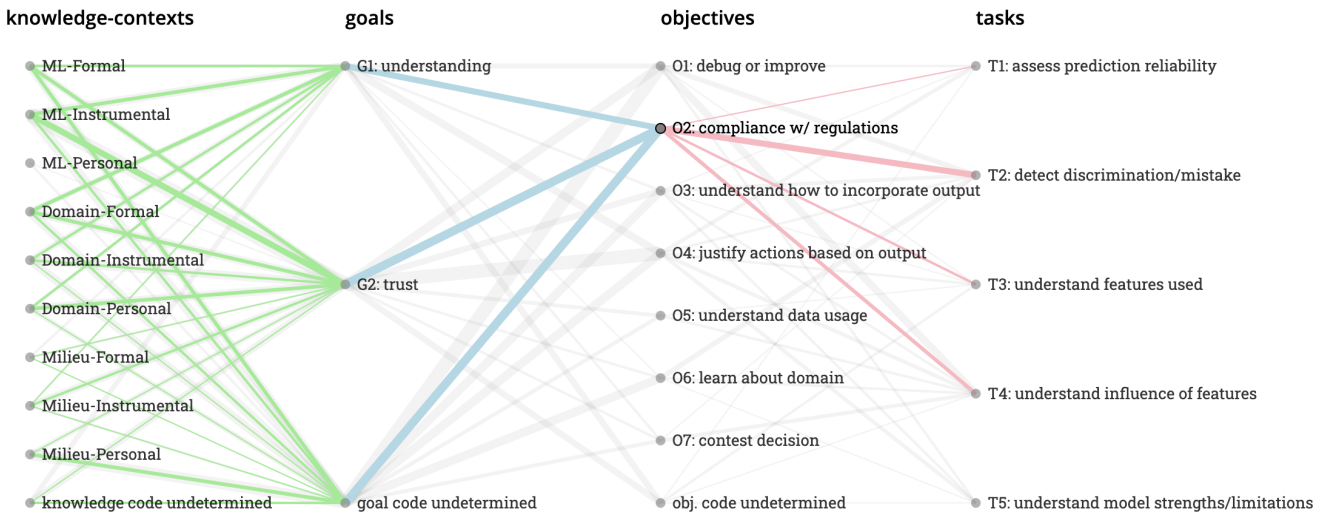

Figure 2: A state of an interactive figure, included in supplementary material[2], that visualizes the results of the analysis of our framework's descriptive power. We see how the two halves of the framework (knowledge-contexts and goals-objectives-tasks) provide a more granular and composable vocabulary with which to describe 58 papers from the literature on ML interpretability. Light grey links represent the set of all papers, and connect codes that appear together. The width of the link corresponds to the number of papers it represents. We use "code undetermined" to indicate cases where we were not able to code a particular category (e.g., if a paper did not explicitly specify a knowledge-context). In the interactive figure, hovering over a code selects all papers that contain the code, and highlights links to visualize the co-occurrence of other codes (e.g., "O2" shown here).

which capture substantial technical expertise. The most observed objectives were justifying or explain decisions influenced by model output (O4, 21/58) and debugging and model improvement (O1, 21/58). The most common tasks were understanding factors that influence the model output (T4, 11/58) and detecting mistaken/arbitrary behavior (T2, 8/58). We note, however, that there were some goals or objectives that arose which we were not able to categorize given our current framework, such as persuading the user [89, 105]. In the current version, we have chosen not to explicitly integrate this into the framework because it was presented as a need imposed upon a stakeholder by, e.g., the company deploying a particular system, while the current needs we describe are driven by the stakeholder themselves.

Beyond its comprehensiveness, the framework is often able to add an additional layer of granularity. For example, whereas many papers describe people with various types of milieu knowledge as "lay users," we are able to recognize and tease apart the different types of expertise they possess. In other cases, we are able to provide consistency and draw similarities between concepts that were previously obscured — for example, from looking at Table 3, we can see that there are many papers that use different terminology to ultimately refer to the same goal.

In using our framework as a descriptive instrument, we found that users' personal knowledge is often the most challenging to define due to the subtle ways it may interact with its context. For example, personal knowledge in a particular domain may cross over into the milieu in cases where a decision domain intersects with everyday functions in society, like finance ("loan applicants", who learn to interact with banking ecosystems). At the same time, another decision subject such as a medical patient might acquire personal domain knowledge by learning about a medical condition that affects them and relating that knowledge to their own symptoms and experiences. Personal knowledge may also be less observable by experimenters who are evaluating and developing systems using traditional processes; however, we believe developing methods for eliciting this knowledge is an under-explored opportunity for human-centered design.

Our framework also helps reveal patterns of under- and over-representation of certain stakeholders and needs in current interpretability research. For example, we noticed a relative lack of interpretability methods specifically focused on objectives like understanding how one's data is being used (O5) and contesting a decision based on the model's output (O7). The papers that did cover these areas were primarily *justifying* these needs from a legal perspective, rather than systems or methods development to help meet them. Similarly, while users with formal ML and data domain knowledge are mentioned frequently, there is significantly less attention paid to those with formal knowledge in the milieu. These stakeholders, however, are likely to have a deep understanding of the broader sociotechnical context in which systems are deployed. Our framework gives us the vocabulary to consciously recognize these gaps, and subsequently, work towards filling them.

Figure 2 uses a parallel coordinates display to summarize the framework's descriptive power, and connect its two halves. Axes correspond to the four components of the framework (knowledge-contexts, goals, objectives, and tasks) and nodes correspond to individual codes used during our qualitative process. Lines connect nodes to represent papers which contain these codes, with line

width encoding the number of papers. In a supplemental interactive version of the figure[2], hovering over specific nodes selects all papers that contain the code, and highlights links to visualize the co-occurrence of other codes. In doing so, the figure demonstrates the composable, many-to-many nature of our framework.

4.2 Evaluative Power

Our framework's evaluative power comes from it's ability to help design more ecologically valid and appropriately-scoped evaluations of interpretability systems/methods.

Similar to prior work [38], we noticed three methods that are primarily used to evaluate interpretability techniques when coding the interpretability literature. The most popular approach does not involve human evaluation. Instead, example outputs generated by the technique are used to illustrate its performance [28, 42, 46, 49, 78, 133, 134] and capabilities including how expressive the technique is [91, 92]. The next-most frequent style of evaluation are user studies conducted with *proxy stakeholders* (e.g., sourced from Amazon Mechanical Turk or a similar online platform) and/or *proxy tasks* (e.g., guessing a model's outcome) [26, 29, 37, 71, 79, 95, 130]. Recent work, however, has demonstrated the fallibility of using proxies — as Buçinca et al. [24] describe, proxies induce a different style of cognition, forcing participants to explicitly attend to explanations and the AI rather than implicitly incorporating them as part of the overall process.

The current gold standard evaluative methodology are application-grounded studies [38] in which domain experts engage with real-world tasks which include interpretable ML. For example, Bussone et al. [23], Wang et al. [119], and Lundberg et al. [81] evaluate interpretability methods by asking healthcare professionals to engage in hypothetical diagnostic scenarios. Although these studies often elicit richer and more relevant feedback about a system's real-world implications, they remain relatively rarely used. We posit that this lack of adoption is due, in part, to the difficulty of designing such studies — there is little principled guidance on how to recruit participants, particularly when the real-world situation requires specific types of expertise. Moreover, when comparing interpretability techniques, it can be difficult to design a task that is equitable for the various conditions. And, even when studies are successfully conducted, it can be difficult to understand how the results generalize or inform future work.

We believe our framework's vocabulary begins to make addressing these issues more tractable. In particular, stakeholder knowledge and context offers a more precise way of defining the participant pool, including identifying vectors along which it may be acceptable to introduce proxying. For instance, consider an ML-enabled clinical diagnosis; if it were difficult to recruit a sufficient number of doctors to participate in the study, our framework suggests that residents or medical students might also be viable participants because of their shared formal data domain knowledge (medicine). Similarly, consider evaluating explanations for loan applications; our framework helps identify that a study may not be ecologically valid if participants do not draw from similar pools of personal knowledge of the milieu.

[2]The interactive figure is also available online at vis.csail.mit.edu/pubs/beyond-expertise-roles/framework-connections.

Finally, our three-level typology of goals, objectives, and tasks provides a structure to operationalize comparative studies. For example, to evaluate the relative effectiveness of similar interpretability techniques (e.g., the plethora of saliency and attribution methods) in a human-centric manner, our framework's *tasks* may be the appropriate level of abstraction to target — they describe operations that can be performed directly and measured through quantitative and qualitative means, and are thus conducive for A/B testing style experiments. On the other hand, for larger-scale interpretability systems, it may be more appropriate to target *objectives* as the specific sequence of operations a participant performs will likely vary significantly between conditions; thus, results will likely be generated qualitatively, through observation and conversation.

4.3 Generative Power

Finally, we operationalize our framework to imagine either novel futures or futures underexplored by the existing work.

Generating new personas. Where prior literature has treated "non-experts" (i.e., stakeholders without expertise in either machine learning or the data domain) as a single, homogeneous group, and has designed for them as such, our framework makes explicit how heterogenous this group may be. By introducing the context of the general milieu, and considering how instrumental and personal knowledge may manifest in it, we can generate new stakeholder personas and imagine the implications on interpretability design. For instance, consider everyday people who have some familiarity or exposure to coding, or have tinkered with the Maker movement; we might describe them as having developed mental models for "computational thinking," or instrumental knowledge in the milieu under our framework. As a result of this knowledge, perhaps they would be more amenable to interpretability interfaces that promoted interactive question-answering by manipulating inputs. Similarly, consider someone who has been closely following mainstream media reporting on the propensity of social media recommendation algorithms to radicalize individuals — our framework would describe them as having rich personal knowledge in the milieu. Perhaps as a result of this knowledge, this person would be initially suspicious of an ML model. In this case, instead of starting with a tabula rasa, perhaps an interpretability interface would be initialized with summaries of the model's strengths and weaknesses (akin to a model card [84]).

Generating new persona-need combinations. Deriving needs from definitions of stakeholders has led to a relatively rigid set of interpretability needs that are recognized and developed for. But, by decoupling needs from stakeholder attributes, our framework allows for a much richer intersection of concerns than prior approaches. For example, consider the objective of debugging or improving a model (O1) — prior work has typically viewed this as a need faced by ML experts, and debugging tools are thus built to primarily serve them. Under our framework, we might describe these prior target users as having formal or instrumental ML knowledge; but our framework also exposes other stakeholders who might wish to address this objective as well: people with personal knowledge in the data domain or milieu. Indeed, this aligns with theories of personal and formal knowledge in Participatory Action Research. As Greenwood and Levin say, *"[p]recisely because local stakeholders*

take action in their own environments, the consequences of errors are both significant to them and often rapidly apparent" [50]. In contrast, researchers with more formal knowledge may *"rarely know whether they are right or not, as their findings seldom are acted upon and the practical results from their research rarely have direct consequences for them."* Feminist standpoint theory [51] would further posit that what is even considered an "error" or harmful might differ depending on the stakeholder's personal knowledge.

Generating new designs. One might initially consider our framework to be silent on *how* to design for particular stakeholders — for instance, it does not explicitly prescribe when to use local explanations [100], saliency maps, or feature visualization [91, 92]. However, our more granular definitions of expertise allows us to adapt theories of knowledge development from cognitive science and pedagogy. In particular, the cognitive science literature describes a process of "chunking", where people organize and think about information in terms of high-level concepts (or "chunks") which develop through experience, familiarity, and with increased knowledge [4, 82, 88]. Similarly, the literature on expertise offers several models of decision making that posit two modes: analytic or deliberative thinking that is based in formal or instrumental knowledge, and intuitive thinking that is based on informal or personal knowledge. Thus, when designing interfaces for ML interpretability, designers could begin by first eliciting and characterizing the knowledge their stakeholders have, and the associated cognitive chunks and/or modes of thinking. Standardized instruments — such as the Preference for Intuition and Deliberation (PID) scale [15] — could also be used. These results could then inform what features are used in an explanation (e.g., raw features or higher level combinations of features that align more with the stakeholder's cognitive chunks) or what types of explanations are given (e.g., more intuitive example-based explanations versus more analytical or mechanistic explanations). Prior work by Wang et al. [119] provides further guidance on how particular types of explanations can be more/less suited to different modes of reasoning.

Situating stakeholders' knowledge and goals within broader societal power dynamics can also help inform what sorts of interpretability methods do or do not work towards subverting existing hierarchies. Indeed, the literature on expertise from which we derive our framework inextricably links types of knowledge with issues of power. In particular, as Fleck notes, *"the view of knowledge as being disinterested or value neutral is idealistic"* and *"the possession of formal knowledge confers status and consequently a measure of power or influence within organizations"* [45]. Interpretability can play a key role here, addressing the *"pathology of beneficence"* that Yielder describes [128] — where experts have a tendency to make decisions *for* people rather than allowing them to decide for themselves — and reducing the ability of experts to merely "rent" out their knowledge [70]. However, this work must be conducted carefully for, as Thatcher et al. [111] observes, *"[t]he very obscurity of transformation from individual data point to commodified, aggregate big data also masks the asymmetrical power relations between users of technology and the almost exclusively corporate entities which algorithmically collect, link, and analyze the data points of many users."* Take the example of a mortgage applicant living in a redlined neighborhood, who wishes to contest the ML-based decision to reject their application. Their relative lack of power in this situation may be further compounded if they have little formal ML or data domain knowledge. We can begin to see, then, that interpretability methods that put the onus on the individual to change things about themselves in order to receive a better outcome (e.g., "Had your income been $3000 higher, you would have received the loan") help uphold, rather than subvert, existing hierarchies. Interpretability methods that instead shift their gaze upwards and focus on alerting affected stakeholders to potentially discriminatory or arbitrary behaviors by the algorithm might provide much stronger evidence to fight against the reigning power differentials [13].

Generating a more reflexive design process. The scope of interpretable ML should not be imagined by researchers or engineers alone — building interpretability systems that challenge power hinge on the involvement of stakeholders with different goals and knowledge. Indeed, people with formal knowledge (e.g., interpretability researchers, developers and designers, and the institutions within which they work) are often precisely the ones in positions of power over those with more personal knowledge (often those most directly affected by algorithmic systems). The concept of *interest convergence*, which stems from critical race theory, holds that those in power tend to support goals that serve their own interests [90]. In other words, without actively involving stakeholders whose interests are in opposition to existing power structures, and considering their input crucial, resultant interpretability systems will fit the standards and needs of those in power — for example, executives with a vested interest in maintaining the status quo, or engineers and researchers who might communicate about model decisions in a way that is not understandable to people without formal ML knowledge. Involving stakeholders with different interests first requires *reflexivity*, or explicitly acknowledging what *our own* backgrounds and interests are. However, doing so in the abstract can be difficult. While our framework was primarily designed to describe the external stakeholders of interpretable ML, we believe it can also be turned to focus internally on the participants of the interpretability design process. By using it to describe our knowledge and goals, we can more clearly recognize gaps in our own knowledge and, thus, the additional people we must deliberately include.

5 LIMITATIONS AND FUTURE WORK

In this paper, we present a framework to characterize the stakeholders of interpretable ML, and their needs. Our framework depicts stakeholder expertise as a two-dimensional space that describes the knowledge they possess (formal, instrumental, and personal knowledge), and the contexts in which this knowledge manifests (ML, the data domain, and the milieu). Our framework also details stakeholder needs as a three-level typology of long-term goals (understanding the model, and building trust in it), shorter-term objectives that build towards these goals (e.g., debugging a model, or contesting a decision), and finally immediate tasks that stakeholders can perform to meet their objectives (e.g., assess prediction reliability, and detect mistakes). In evaluating our framework, we find that it suitably covers a sample of 58 papers on ML interpretability, and its granular structure reveals gaps in the literature. Moreover, while speculative, we believe the framework offers the necessary vocabulary to assist in more precisely comparing and conducting

user-focused evaluations of interpretability systems. Finally, we find that the framework offers a richer intersection of stakeholder expertise and needs than prior approaches, and that it can be turned inward to facilitate a more reflexive design process.

Our framework takes the next step in better defining who the users of interpretable ML are, and its limitations point to promising opportunities for future work. In particular, we do not consider our framework to be an exhaustive description of the problem space, but rather a "living" artifact that will grow and adapt as interpretability matures as a research field. For example, we expect new goals, objectives, and tasks to be added to the framework as ML is deployed more deeply in existing domains, and as it reaches new domains. Indeed, when coding the interpretability literature, we found occasional instances of needs that do not precisely fit into our current framework (e.g., persuasion, adoption). But, more evidence is needed to determine at what level of the typology these needs fit into, and whether they are specific instances of a more general or fundamental need.

Similarly, while our framework begins to decompose expertise into knowledge and contexts, the models we base it on provide even more granularity. For example, Fleck [45] names several additional types of knowledge including informal knowledge, contingent knowledge, tacit knowledge, and meta-knowledge; and Eraut [43] identifies cultural and tacit knowledge, and the degree to which either have or have not been codified. Under our framework, these different types all lie within personal knowledge as we did not find sufficient evidence in the interpretability literature to warrant the additional granularity. The milieu context is similarly broad—covering physical, social, and cultural contexts in the literature. As additional work on interpretable ML is conducted, these two broad categories may come under the same pressure we initially identified with prior expertise- and role-based approaches: they may begin to conflate otherwise independent concerns. By identifying recurring instances of these tensions, we can begin to disentangle them and enumerate other knowledge types and contexts that are meaningful for interpretability.

Finally, our framework's model of expertise is grounded only in epistemology; but the literature on expertise has also argued that expertise is constructed rhetorically. As Johanna Hartelius describes, *"[a] speaker is only able to exercise expertise and enjoy expert status to the extent that she can persuade an audience to grant such things"* [52]. Rhetoric undoubtedly plays an important role in interpretability, and we can see evidence for this in the adjacent domain of data visualization [32]. Researchers have argued that the clean, minimalist aesthetic of modern visualizations lends them an air of authority and certainty [62] that contributes to their *"persuasive and seductive rhetorical force"* [41]. Through close readings of visualizations, researchers have shown how citing sources and representing uncertainty can signal transparency and impartiality [59], and with empirical studies, researchers have demonstrated that even seemingly-innocuous elements like titles can frame or slant reading visualizations [66] and can impact trust and recall [67]. How to adapt and replicate these findings for interpretability is a fertile ground for future work, and interpretability poses its own unique considerations. In particular, unlike visualizations, the rhetorical performance of an interpretability interface may sometimes be shared with or mediated by a human (e.g., an "operator" [113] or through reports [58], respectively).

ACKNOWLEDGMENTS

This research was sponsored by NSF Award #1900991, and by the United States Air Force Research Laboratory under Cooperative Agreement Number FA8750-19-2-1000. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the United States Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

REFERENCES

[1] 2020. ML Interpretability for Scientific Discovery (MLI4SD) Workshop. https://sites.google.com/view/mli4sd-icml2020/home. Accessed: 2020-09-16.

[2] Agnar Aamodt and Enric Plaza. 1994. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI communications* 7, 1 (1994), 39–59.

[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. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*. ACM Press, Montreal QC, Canada, 1–18. https://doi.org/10.1145/3173574.3174156

[4] 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.

[5] Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052 Conference Name: IEEE Access.

[6] Ali Alkhatib and Michael Bernstein. 2019. Street-level algorithms: A theory at the gaps between policy and decisions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[7] Vijay Arya, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilović, Sami Mourad, Pablo Pedemonte, Ramya Raghavendra, John Richards, Prasanna Sattigeri, Karthikeyan Shanmugam, Moninder Singh, Kush R. Varshney, Dennis Wei, and Yunfeng Zhang. 2019. One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques. *arXiv:1909.03012 [cs, stat]* (Sept. 2019). http://arxiv.org/abs/1909.03012 arXiv: 1909.03012.

[8] Robert K Atkinson, Sharon J Derry, Alexander Renkl, and Donald Wortham. 2000. Learning from examples: Instructional principles from the worked examples research. *Review of educational research* 70, 2 (2000), 181–214.

[9] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one* 10, 7 (2015), e0130140.

[10] Krisztian Balog and Filip Radlinski. 2020. Measuring Recommendation Explanation Quality:The Conflicting Goals of Explanations. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*. ACM, New York, NY, USA, Virtual Event, 10.

[11] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter Lasecki, Daniel S Weld, and Eric Horvitz. [n.d.]. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. ([n. d.]), 10.

[12] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel S. Weld. 2020. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. *arXiv:2006.14779 [cs]* (June 2020). http://arxiv.org/abs/2006.14779 arXiv: 2006.14779.

[13] Chelsea Barabas, Colin Doyle, JB Rubinovitz, and Karthik Dinakar. 2020. Studying up: reorienting the study of algorithmic fairness around issues of power. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 167–176.

[14] Michel Beaudouin-Lafon. 2004. Designing interaction, not interfaces. In *Proceedings of the working conference on Advanced visual interfaces*. 15–22.

[15] Cornelia Betsch. 2004. Präferenz für intuition und deliberation (PID). *Zeitschrift für Differentielle und Diagnostische Psychologie* 25, 4 (2004), 179–197.

[16] Aviruch Bhatia, Vishal Garg, Philip Haves, and Vikram Pudi. 2019. Explainable Clustering Using Hyper-Rectangles for Building Energy Simulation Data. *E&ES*

238, 1 (2019), 012068.

[17] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and Peter Eckersley. 2020. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, Barcelona, Spain, 648–657. https://doi.org/10.1145/3351095.3375624

[18] Stephen Billett (Ed.). 2010. *Learning Through Practice*. Springer Netherlands, Dordrecht. https://doi.org/10.1007/978-90-481-3939-2

[19] Nadia Boukhelifa, Anastasia Bezerianos, Ioan Cristian Trelea, Nathalie Méjean Perrot, and Evelyne Lutton. 2019. An Exploratory Study on Visual Exploration of Model Simulations by Multiple Types of Experts. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*. ACM Press, Glasgow, Scotland Uk, 1–14. https://doi.org/10.1145/3290605.3300874

[20] Matthew Brehmer and Tamara Munzner. 2013. A multi-level typology of abstract visualization tasks. *IEEE transactions on visualization and computer graphics* 19, 12 (2013), 2376–2385.

[21] Andrea Brennen. 2020. What Do People Really Want When They Say They Want "Explainable AI?" We Asked 60 Stakeholders.. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (CHI EA '20)*. Association for Computing Machinery, Honolulu, HI, USA, 1–7. https://doi.org/10.1145/3334480.3383047

[22] Peter Buneman, Sanjeev Khanna, and Tan Wang-Chiew. 2001. Why and where: A characterization of data provenance. In *International conference on database theory*. Springer, 316–330.

[23] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. In *2015 International Conference on Healthcare Informatics*. IEEE, Dallas, TX, USA, 160–169. https://doi.org/10.1109/ICHI.2015.26

[24] 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. *Proceedings of the 25th International Conference on Intelligent User Interfaces* (March 2020), 454–464. https://doi.org/10.1145/3377325.3377498 arXiv: 2001.08298.

[25] Ruth M. J. Byrne. 2019. Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, Macao, China, 6276–6282. https://doi.org/10.24963/ijcai.2019/876

[26] Carrie J Cai, Jonas Jongejan, and Jess Holbrook. 2019. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 258–262.

[27] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 1–24. https://doi.org/10.1145/3359206

[28] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. 2019. This looks like that: deep learning for interpretable image recognition. In *Advances in neural information processing systems*. 8930–8941.

[29] 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.

[30] Michael Chromik, Malin Eiband, Sarah Theres Völkel, and Daniel Buschek. 2019. Dark Patterns of Explainability, Transparency, and User Control for Intelligent Systems. *Los Angeles* (2019), 6.

[31] Danielle Keats Citron and Frank Pasquale. 2014. The scored society: Due process for automated predictions. *Wash. L. Rev.* 89 (2014), 1.

[32] Michael Correll. 2019. Ethical dimensions of visualization research. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[33] Amanda Cox. 2011. Shaping Data for News. https://vimeo.com/29391942

[34] Gloria Dall'Alba and Jörgen Sandberg. 2006. Unveiling professional development: A critical review of stage models. *Review of educational research* 76, 3 (2006), 383–412.

[35] Arun Das and Paul Rad. 2020. Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey. *arXiv:2006.11371 [cs]* (June 2020). http://arxiv.org/abs/2006.11371 arXiv: 2006.11371.

[36] Amit Dhurandhar, Vijay Iyengar, Ronny Luss, and Karthikeyan Shanmugam. 2017. A formal framework to characterize interpretability of procedures. *arXiv preprint arXiv:1707.03886* (2017).

[37] 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.

[38] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv:1702.08608 [cs, stat]* (March 2017). http://arxiv.org/abs/1702.08608 arXiv: 1702.08608.

[39] Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Stuart Schieber, James Waldo, David Weinberger, and Alexandra Wood. 2017. Accountability of AI under the law: The role of explanation. *arXiv preprint arXiv:1711.01134* (2017).

[40] Hubert L Dreyfus and Stuart E Dreyfus. 1986. The power of human intuition and expertise in the era of the computer. *Mind over machine. Nueva York: The Free Press* (1986).

[41] Johanna Drucker. 2012. Humanistic theory and digital scholarship. *Debates in the digital humanities* 150 (2012), 85–95.

[42] Mengnan Du, Ninghao Liu, and Xia Hu. 2019. Techniques for interpretable machine learning. *Commun. ACM* 63, 1 (2019), 68–77.

[43] Michael Eraut. 2010. Knowledge, working practices, and learning. In *Learning through practice*. Springer, 37–58.

[44] Juliana J. Ferreira and Mateus S. Monteiro. 2020. What Are People Doing About XAI User Experience? A Survey on AI Explainability Research and Practice. In *Design, User Experience, and Usability. Design for Contemporary Interactive Environments (Lecture Notes in Computer Science)*, Aaron Marcus and Elizabeth Rosenzweig (Eds.). Springer International Publishing, Cham, 56–73. https://doi.org/10.1007/978-3-030-49760-6_4

[45] James Fleck. 1998. Expertise: knowledge, power and tradeability. In *Exploring expertise*. Springer, 143–171.

[46] Ruth C. Fong and Andrea Vedaldi. 2017. Interpretable Explanations of Black Boxes by Meaningful Perturbation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[47] 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)*. 80–89. https://doi.org/10.1109/DSAA.2018.00018

[48] Alyssa Glass, Deborah L. McGuinness, and Michael Wolverton. 2008. Toward establishing trust in adaptive agents. In *Proceedings of the 13th international conference on Intelligent user interfaces - IUI '08*. ACM Press, Gran Canaria, Spain, 227. https://doi.org/10.1145/1378773.1378804

[49] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Counterfactual Visual Explanations. In *Proceedings of the 36th International Conference on Machine Learning*, Vol. 97. Long Beach, California, USA. http://proceedings.mlr.press/v97/goyal19a.html arXiv: 1904.07451.

[50] Davydd Greenwood and Morten Levin. 2007. *Introduction to Action Research*. SAGE Publications, Inc., 2455 Teller Road, Thousand Oaks California 91320 United States of America. https://doi.org/10.4135/9781412984614

[51] Donna Haraway. 1988. Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist studies* 14, 3 (1988), 575–599.

[52] Johanna Hartelius. 2008. *The Rhetoric of Expertise*. Ph.D. Dissertation. The University of Texas at Austin.

[53] Johanna Hartelius. 2011. Rhetorics of Expertise. *Social Epistemology* 25, 3 (July 2011), 211–215. https://doi.org/10.1080/02691728.2011.578301 Publisher: Routledge _eprint: https://doi.org/10.1080/02691728.2011.578301.

[54] Sam Hepenstal and David McNeish. 2020. Explainable Artificial Intelligence: What Do You Need to Know?. In *Augmented Cognition. Theoretical and Technological Approaches*, Dylan D. Schmorrow and Cali M. Fidopiastis (Eds.). Springer International Publishing, Cham, 266–275.

[55] Mireille Hildebrandt. 2012. The Dawn of a Critical Transparency Right for the Profiling Era. *Astronomy & Astrophysics - ASTRON ASTROPHYS* (06 2012). https://doi.org/10.3233/978-1-61499-057-4-41

[56] 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)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300809

[57] Fred Hohman, Minsuk Kahng, Robert Pienta, and Duen Horng Chau. 2019. Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers. *IEEE Transactions on Visualization and Computer Graphics* 25, 8 (Aug. 2019), 2674–2693. https://doi.org/10.1109/TVCG.2018.2843369

[58] 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 (May 2020), 1–26. https://doi.org/10.1145/3392878

[59] Jessica Hullman and Nick Diakopoulos. 2011. Visualization rhetoric: Framing effects in narrative visualization. *IEEE transactions on visualization and computer graphics* 17, 12 (2011), 2231–2240.

[60] Janet Jull, Audrey Giles, and Ian D. Graham. 2017. Community-based participatory research and integrated knowledge translation: advancing the co-creation of knowledge. *Implementation Science* 12, 1 (Dec. 2017), 150. https://doi.org/10.1186/s13012-017-0696-3

[61] 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. ACM, Honolulu HI USA, 1–14. https://doi.org/10.1145/3313831.3376219

[62] Helen Kennedy, Rosemary Lucy Hill, Giorgia Aiello, and William Allen. 2016. The work that visualisation conventions do. *Information, Communication & Society* 19, 6 (2016), 715–735.

[63] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*. PMLR, 2668–2677.

[64] Ian M. Kinchin and B. Cabot. 2010. Reconsidering the dimensions of expertise: from linear stages towards dual processing. *London Review of Education* (July 2010). https://doi.org/10.1080/14748460.2010.487334

[65] René F Kizilcec. 2016. How much information? Effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 2390–2395.

[66] Ha-Kyung Kong, Zhicheng Liu, and Karrie Karahalios. 2018. Frames and slants in titles of visualizations on controversial topics. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.

[67] Ha-Kyung Kong, Zhicheng Liu, and Karrie Karahalios. 2019. Trust and recall of information across varying degrees of title-visualization misalignment. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[68] Luciana Monteiro Krebs, Oscar Luis Alvarado Rodriguez, Pierre Dewitte, Jef Ausloos, David Geerts, Laurens Naudts, and Katrien Verbert. 2019. Tell me what you know: GDPR implications on designing transparency and accountability for news recommender systems. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–6.

[69] Bogdan Kulynych, David Madras, Smitha Milli, Inioluwa Deborah Raji, Angela Zhou, and Richard Zemel. 2020. Participatory Approaches to Machine Learning. International Conference on Machine Learning Workshop.

[70] David F Labaree. 2000. On the nature of teaching and teacher education: Difficult practices that look easy. *Journal of teacher education* 51, 3 (2000), 228–233.

[71] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Samuel J. Gershman, and Finale Doshi-Velez. 2019. Human Evaluation of Models Built for Interpretability. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7, 1 (Oct. 2019), 59–67. https://www.aaai.org/ojs/index.php/HCOMP/article/view/5280 Number: 1.

[72] Vivian Lai and Chenhao Tan. 2019. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*. ACM Press, Atlanta, GA, USA, 29–38. https://doi.org/10.1145/3287560.3287590

[73] Himabindu Lakkaraju and Osbert Bastani. 2020. "How do I fool you?": Manipulating User Trust via Misleading Black Box Explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)*. Association for Computing Machinery, New York, NY, USA, 79–85. https://doi.org/10.1145/3375627.3375833

[74] Christopher A. Le Dantec and Sarah Fox. 2015. Strangers at the Gate: Gaining Access, Building Rapport, and Co-Constructing Community-Based Research. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '15*. ACM Press, Vancouver, BC, Canada, 1348–1358. https://doi.org/10.1145/2675133.2675147

[75] Bongshin Lee, Kate Isaacs, Danielle Albers Szafir, G Elisabeta Marai, Cagatay Turkay, Melanie Tory, Sheelagh Carpendale, and Alex Endert. 2019. Broadening intellectual diversity in visualization research papers. *IEEE computer graphics and applications* 39, 4 (2019), 78–85.

[76] 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 (CHI '20)*. Association for Computing Machinery, Honolulu, HI, USA, 1–15. https://doi.org/10.1145/3313831.3376590

[77] Brian Y. Lim and Anind K. Dey. 2009. Assessing Demand for Intelligibility in Context-Aware Applications. In *Proceedings of the 11th International Conference on Ubiquitous Computing* (Orlando, Florida, USA) *(UbiComp '09)*. Association for Computing Machinery, New York, NY, USA, 195–204. https://doi.org/10.1145/1620545.1620576

[78] 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* (Copenhagen, Denmark) *(UbiComp '10)*. Association for Computing Machinery, New York, NY, USA, 13–22. https://doi.org/10.1145/1864349.1864353

[79] 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* (Boston, MA, USA) *(CHI '09)*. Association for Computing Machinery, New York, NY, USA, 2119–2128. https://doi.org/10.1145/1518701.1519023

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

[81] Scott M Lundberg, Bala Nair, Monica S Vavilala, Mayumi Horibe, Michael J Eisses, Trevor Adams, David E Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim, et al. 2018. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature biomedical engineering* 2, 10 (2018), 749–760.

[82] George A Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review* 63, 2 (1956), 81.

[83] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (Feb. 2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[84] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.

[85] Shakir Mohamed, Marie-Therese Png, and William Isaac. 2020. Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. *Philosophy & Technology* (July 2020). https://doi.org/10.1007/s13347-020-00405-8 arXiv: 2007.04068.

[86] Sina Mohseni, Jeremy E. Block, and Eric D. Ragan. 2020. A Human-Grounded Evaluation Benchmark for Local Explanations of Machine Learning. *arXiv:1801.05075 [cs]* (June 2020). http://arxiv.org/abs/1801.05075 arXiv: 1801.05075.

[87] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2020. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *arXiv:1811.11839 [cs]* (April 2020). http://arxiv.org/abs/1811.11839 arXiv: 1811.11839.

[88] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2018. How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1802.00682* (2018).

[89] Ingrid Nunes and Dietmar Jannach. 2017. A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction* 27, 3-5 (2017), 393–444.

[90] Ihudiya Finda Ogbonnaya-Ogburu, Angela D.R. Smith, Alexandra To, and Kentaro Toyama. 2020. Critical Race Theory for HCI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–16. https://doi.org/10.1145/3313831.3376392

[91] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. 2017. Feature visualization. *Distill* 2, 11 (2017), e7.

[92] Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. 2018. The building blocks of interpretability. *Distill* 3, 3 (2018), e10.

[93] Prajwal Paudyal, Junghyo Lee, Azamat Kamzin, Mohamad Soudki, Ayan Banerjee, and Sandeep KS Gupta. 2019. Learn2Sign: Explainable AI for Sign Language Learning.. In *IUI Workshops*.

[94] Evan M Peck, Sofia E Ayuso, and Omar El-Etr. 2019. Data is personal: Attitudes and perceptions of data visualization in rural pennsylvania. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[95] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Manipulating and Measuring Model Interpretability. *arXiv:1802.07810 [cs]* (Nov. 2019). http://arxiv.org/abs/1802.07810 arXiv: 1802.07810.

[96] Alun Preece, Dan Harborne, Dave Braines, Richard Tomsett, and Supriyo Chakraborty. 2018. Stakeholders in Explainable AI. *arXiv:1810.00184 [cs]* (Sept. 2018). http://arxiv.org/abs/1810.00184 arXiv: 1810.00184.

[97] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 33–44.

[98] 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*, Hugo Jair Escalante, Sergio Escalera, Isabelle Guyon, Xavier Baró, Yağmur Güçlütürk, Umut Güçlü, and Marcel van Gerven (Eds.). Springer International Publishing, Cham, 19–36. https://doi.org/10.1007/978-3-319-98131-4_2

[99] Alexander Renkl. 2014. Toward an instructionally oriented theory of example-based learning. *Cognitive science* 38, 1 (2014), 1–37.

[100] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386* (2016).

[101] Mireia Ribera and Agata Lapedriza. 2019. Can we do better explanations? A proposal of user-centered explainable AI.. In *IUI Workshops*.

[102] Ribana Roscher, Bastian Bohn, Marco F. Duarte, and Jochen Garcke. 2020. Explainable Machine Learning for Scientific Insights and Discoveries. *IEEE Access* 8 (2020), 42200–42216. https://doi.org/10.1109/ACCESS.2020.2976199

[103] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. 2017. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning

models. *arXiv preprint arXiv:1708.08296* (2017).

[104] Udo Schlegel, Hiba Arnout, Mennatallah El-Assady, Daniela Oelke, and Daniel A. Keim. 2019. Towards a Rigorous Evaluation of XAI Methods on Time Series. *arXiv:1909.07082 [cs]* (Sept. 2019). http://arxiv.org/abs/1909.07082 arXiv: 1909.07082.

[105] Johannes Schneider and Joshua Handali. 2019. PERSONALIZED EXPLANATION FOR MACHINE LEARNING: A CONCEPTUALIZATION. In *Proceedings of the 27th European Conference on Information Systems (ECIS)*. Stockholm & Uppsala, Sweden. https://aisel.aisnet.org/ecis2019_rp/171

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

[107] Thilo Spinner, Udo Schlegel, Hanna Schäfer, and Mennatallah El-Assady. 2020. explAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (Jan. 2020), 1064–1074. https://doi.org/10.1109/TVCG.2019.2934629 Conference Name: IEEE Transactions on Visualization and Computer Graphics.

[108] Clay Spinuzzi. 2005. The methodology of participatory design. *Technical communication* 52, 2 (2005), 163–174.

[109] Mukund Sundararajan, Jinhua Xu, Ankur Taly, Rory Sayres, and Amir Najmi. 2019. Exploring Principled Visualizations for Deep Network Attributions.. In *IUI Workshops*, Vol. 4.

[110] Harini Suresh, Natalie Lao, and Ilaria Liccardi. 2020. Misplaced Trust: Measuring the Interference of Machine Learning in Human Decision-Making. In *WebSci '20: 12th ACM Conference on Web Science, Southampton, UK, July 6-10, 2020*, Emilio Ferrara, Pauline Leonard, and Wendy Hall (Eds.). ACM, 315–324. https://doi.org/10.1145/3394231.3397922

[111] Jim Thatcher, David O'Sullivan, and Dillon Mahmoudi. 2016. Data colonialism through accumulation by dispossession: New metaphors for daily data. *Environment and Planning D: Society and Space* 34, 6 (Dec. 2016), 990–1006. https://doi.org/10.1177/0263775816633195

[112] Andreas Theodorou, Robert H Wortham, and Joanna J Bryson. 2017. Designing and implementing transparency for real time inspection of autonomous robots. *Connection Science* 29, 3 (2017), 230–241.

[113] Richard Tomsett, Dave Braines, Dan Harborne, Alun Preece, and Supriyo Chakraborty. 2018. Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems. In *arXiv:1806.07552 [cs]*. http://arxiv.org/abs/1806.07552 arXiv: 1806.07552.

[114] Sana Tonekaboni, Shalmali Joshi, Melissa D. McCradden, and Anna Goldenberg. 2019. What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. In *Machine Learning for Healthcare Conference*. 359–380. http://proceedings.mlr.press/v106/tonekaboni19a.html ISSN: 1938-7228 Section: Machine Learning.

[115] Joe Tullio, Anind K. Dey, Jason Chalecki, and James Fogarty. 2007. How it works: a field study of non-technical users interacting with an intelligent system. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. Association for Computing Machinery, San Jose, California, USA, 31–40. https://doi.org/10.1145/1240624.1240630

[116] Berk Ustun, Alexander Spangher, and Yang Liu. 2019. Actionable Recourse in Linear Classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 10–19. https://doi.org/10.1145/3287560.3287566

[117] Effy Vayena, Alessandro Blasimme, and I. Glenn Cohen. 2018. Machine learning in medicine: Addressing ethical challenges. *PLOS Medicine* 15, 11 (Nov. 2018), e1002689. https://doi.org/10.1371/journal.pmed.1002689

[118] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2018. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology* 31, 2 (March 2018), 841–887. http://arxiv.org/abs/1711.00399 arXiv: 1711.00399.

[119] 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.

[120] Ryan Weber. 2012. Review of The Rhetoric of Expertise. *Rhetoric and Public Affairs* 15, 1 (2012), 193–196. https://www.jstor.org/stable/41955617 Publisher: Michigan State University Press.

[121] Adrian Weller. 2019. Transparency: Motivations and Challenges. *arXiv:1708.01870 [cs]* (Aug. 2019). http://arxiv.org/abs/1708.01870 arXiv: 1708.01870.

[122] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, and J. Wilson. 2020. The What-If Tool: Interactive Probing of Machine Learning Models. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2020), 56–65.

[123] Robin Williams, Wendy Faulkner, and James Fleck (Eds.). 1998. *Exploring Expertise*. Palgrave Macmillan UK, London. https://doi.org/10.1007/978-1-349-13693-3

[124] Jacob O Wobbrock and Julie A Kientz. 2016. Research contributions in human-computer interaction. *interactions* 23, 3 (2016), 38–44.

[125] Christine T. Wolf. 2019. Explainability scenarios: towards scenario-based XAI design. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, Marina del Ray California, 252–257. https://doi.org/10.1145/3301275.3302317

[126] Yao Xie, Ge Gao, and Xiang 'Anthony' Chen. 2019. Outlining the Design Space of Explainable Intelligent Systems for Medical Diagnosis. arXiv:1902.06019 [cs.HC]

[127] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, Honolulu, HI, USA, 1–13. https://doi.org/10.1145/3313831.3376301

[128] Jill Yielder. 2001. *Professional Expertise: A Model for Integration and Change*. Thesis. ResearchSpace@Auckland. https://researchspace.auckland.ac.nz/handle/2292/2340 Accepted: 2008-01-30T02:04:57Z.

[129] Jill Yielder. 2004. An integrated model of professional expertise and its implications for higher education. *International Journal of Lifelong Education* 23, 1 (Jan. 2004), 60–80. https://doi.org/10.1080/0260137032000172060

[130] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, Glasgow, Scotland Uk, 1–12. https://doi.org/10.1145/3290605.3300509

[131] Rulei Yu and Lei Shi. 2018. A user-based taxonomy for deep learning visualization. *Visual Informatics* 2, 3 (Sept. 2018), 147–154. https://doi.org/10.1016/j.visinf.2018.09.001

[132] Tal Z Zarsky. 2013. Transparent predictions. *U. Ill. L. Rev.* (2013), 1503.

[133] Matthew D. Zeiler and Rob Fergus. 2014. Visualizing and Understanding Convolutional Networks. In *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Vol. 8689. Springer International Publishing, Cham, 818–833. https://doi.org/10.1007/978-3-319-10590-1_53 Series Title: Lecture Notes in Computer Science.

[134] Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. 2018. Interpretable Convolutional Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[135] 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 2020 Conference on Fairness, Accountability, and Transparency*. 295–305.

[136] Ruijing Zhao, Izak Benbasat, and Hasan Cavusoglu. 2019. Transparency in Advice-Giving Systems: A Framework and a Research Model for Transparency Provision.. In *IUI Workshops*.

[137] Jan Ruben Zilke, Eneldo Loza Mencía, and Frederik Janssen. 2016. Deepred–rule extraction from deep neural networks. In *International Conference on Discovery Science*. Springer, 457–473.