frontiers | Frontiers in Computer Science

TYPE Original Research
PUBLISHED 17 August 2023
DOI 10.3389/fcomp.2023.1117848

OPEN ACCESS

EDITED BY
Chathurika S. Wickramasinghe Brahmana,
Capital One, United States

REVIEWED BY
Federico Cabitza,
University of Milano-Bicocca, Italy
Ofra Amir,
Technion Israel Institute of Technology, Israel

*CORRESPONDENCE
Robert R. Hoffman
✉ rhoffman@ihmc.us

RECEIVED 07 December 2022
ACCEPTED 24 July 2023
PUBLISHED 17 August 2023

CITATION
Hoffman RR, Mueller ST, Klein G, Jalaeian M and Tate C (2023) Explainable AI: roles and stakeholders, desirements and challenges. *Front. Comput. Sci.* 5:1117848. doi: 10.3389/fcomp.2023.1117848

Explainable AI: roles and stakeholders, desirements and challenges

Robert R. Hoffman[1]*, Shane T. Mueller[2], Gary Klein[3], Mohammadreza Jalaeian[4] and Connor Tate[1]

[1]Institute for Human and Machine Cognition, Pensacola, FL, United States, [2]Department of Cognitive and Learning Sciences, Michigan Technological University, Houghton, MI, United States, [3]MacroCognition, LLC, Dayton, OH, United States, [4]Department of Integrated Systems Engineering, Ohio State University, Columbus, OH, United States

Introduction: The purpose of the Stakeholder Playbook is to enable the developers of explainable AI systems to take into account the different ways in which different stakeholders or role-holders need to "look inside" the AI/XAI systems.

Method: We conducted structured cognitive interviews with senior and mid-career professionals who had direct experience either developing or using AI and/or autonomous systems.

Results: The results show that role-holders need access to others (e.g., trusted engineers and trusted vendors) for them to be able to develop satisfying mental models of AI systems. They need to know how it fails and misleads as much as they need to know how it works. Some stakeholders need to develop an understanding that enables them to explain the AI to someone else and not just satisfy their own sense-making requirements. Only about half of our interviewees said they always wanted explanations or even needed better explanations than the ones that were provided. Based on our empirical evidence, we created a "Playbook" that lists explanation desires, explanation challenges, and explanation cautions for a variety of stakeholder groups and roles.

Discussion: This and other findings seem surprising, if not paradoxical, but they can be resolved by acknowledging that different role-holders have differing skill sets and have different sense-making desires. Individuals often serve in multiple roles and, therefore, can have different immediate goals. The goal of the Playbook is to help XAI developers by guiding the development process and creating explanations that support the different roles.

KEYWORDS

explainable AI, stakeholders, role-holders, explanatory reasoning, desirements

1. Introduction

The notion of AI systems that could explain themselves to end-users was stimulated in part by the Report of the European Union on the notion of a right to an explanation of the decision that an AI system makes (European Union Commission, 2016). Much of the emphasis was on the ethical implications of AI (Wachter et al., 2016; Goodman and Flaxman, 2017; Floridi et al., 2018). Research on the explanation needs of end-users has been a driving factor in explainable AI and the focus of a considerable amount of research.

But many researchers raised the possibility that other stakeholders besides end-users would not only need explanations but also would need different kinds of explanations (forms and contents) depending on their circumstances and responsibilities (e.g., Sheh and Monteath, 2018; Hepenstal and McNeish, 2020; Langer et al., 2021). A stakeholder is an individual who has some investment in AI, either in the form of direct support for

research and development or a vested interest in the success of the AI. The importance of considering stakeholder explanation needs has recently become salient (see Fazelpour, 2023; Shneiderman, 2023).

The focus of the present article is on the explanation needs of stakeholders. The purpose of the Stakeholder Playbook which we present in this article is to enable system developers to consider the different ways in which stakeholders or role-holders need to "look inside" the AI/XAI system. For example, some stakeholders might need to understand the boundary conditions of the system (its strengths and limitations).

This article begins by encapsulating the pertinent literature, and the opening question of how to taxonomize stakeholder groups. We then present the method and results of an empirical investigation that led to the Playbook. Senior and mid-career professionals having had direct experience either developing or using AI and/or autonomous systems were engaged in cognitive interviews concerning their roles and responsibilities. The results included many findings that came to us as surprises, which this article elaborates.

1.1. Background

Explainable AI has sparked interest in the study of how people explain complex systems, for themselves and to others (Hoffman et al., 2017; Klein et al., 2021). There have appeared many discussions of the quality of explanations, possible methodologies for evaluation (Doshi-Velez and Kim, 2017), and attempts to empirically evaluate explanation quality and the effectiveness of explanations (e.g., Hoffman et al., 2011, 2023; Miller, 2017; Mueller et al., 2019; Johs et al., 2020; Kenny et al., 2021).

There have been successful demonstrations that machine-generated explanations can have a positive impact on performance. Users sometimes desire explanations, but explanations can be particularly helpful when the AI is incorrect. Explanations can influence user trust in the AI and help users develop richer mental models of how the AI works (Gunning et al., 2021; Hoffman et al., 2023). Buçinca et al. (2020) found that explanations resulted in an improved performance on a simple AI-supported judgment task. Also using a proxy task, Lage et al. (2019) found that explanations helped people verify the AI's recommendation and helped them decide whether a change to input data would result in a change in the AI's recommendation. The explanatory human-machine dialog can enhance the process of knowledge acquisition for expert systems (e.g., Arioua et al., 2017). Strout et al. (2019) demonstrated that human "rationales" (annotations) of text material that a machine learning system presents as explanations for its classifications lead to judgments that the explanations are improved (see also Zaidan et al., 2007; Zhang et al., 2016; Nguyen, 2018).

The converse finding is that explanations do not always help. They can negatively impact user confidence in a final determination (Cabitza et al., 2023). The authors also raised issues about methods that might be used in XAI evaluation. Loyola-Gonzalez (2019) reviewed types of white box ("understandable") models and raised empirical questions about what it means for a white box explanation to be good (e.g., "Are decision trees self-explanatory?"). The author proposed a method in which domain experts would judge the goodness of white box models of various kinds (decision trees and decision rules). Rudin (2019) questioned the entire idea of explaining black box models with white boxes, in favor of simply creating back boxes that are interpretable in the first place.

In a recent review of the XAI literature, Cabitza et al. (2023) draw an important distinction between epistemological (formal/scientific) explanation and psychological (cognitive) explanation. It is this latter category that is the focus of the present study.

1.2. Types of explanations

The explanation of AI systems to users, stakeholders, or other beneficiaries of AI systems will almost inevitably differ from the explanations that satisfy a formalist understanding of explainability, interpretability, or transparency, in such modes as decision trees and logical programs (e.g., Tomsett et al., 2018; Calegari et al., 2019; Felzmann et al., 2019; Chari et al., 2020; Kaur et al., 2020; Tjoa and Guan, 2020).

Many articles have provided rosters of different types of explanations, framed as either a taxonomy of types or a classification based on multiple qualitative dimensions (e.g., Floridi et al., 2018; Sheh and Monteath, 2018; Hind et al., 2019; Arrieta et al., 2020; Dahan, 2020; Vermeire et al., 2021). Also, many articles refer to particular domains such as Health Care, Data Analytics, AI, and Verification and Validation (Preece et al., 2018; Arya et al., 2019; Chari et al., 2020; Mohseni et al., 2020) or legal domains (Al-Abdulkarim et al., 2016; Al-Abdulakarim et al., 2019; Felzmann et al., 2019; Atkinson et al., 2020).

Awareness of the importance of stakeholder dependence has led researchers in computer science to propose mappings of stakeholder groups onto explanation requirements. We now briefly review this literature.

1.3. Recent pertinent literature on stakeholder dependence

Many XAI researchers have emphasized the importance of stakeholder dependence or domain dependence on machine-generated explanations (Eiband et al., 2018; Sheh and Monteath, 2018; Hepenstal and McNeish, 2020; for reviews, see Mittelstadt et al., 2019; Langer et al., 2021).

Most of the discussions of the stakeholder relativity refer to a small set of stakeholder groups; sometimes only to a distinction among end-users, system developers, and "others." It has been asserted that developers need "scientific" explanations whereas end-users or lay persons need "everyday" explanations (Langer et al., 2021). Naiseh et al. (2020) offer a broader palette of stakeholder groups, integrating lists from Tomsett et al. (2018), Hind et al. (2019), and Ribera and Lapedriza (2019): creators/developers, AI researchers, lay users, operators, domain experts, decision-makers, affected parties, ethicists, theorists, external regulatory entities, or oversight organizations.

One could add to this list human factors/cognitive systems engineers, development team leaders, program managers, procurement officers, systems integrators, vendor managers, trainers, policymakers, regulators, legal practitioners, rights advocates, and guardians.

Speculations have been offered about the explanation requirements of stakeholder groups (Goodman and Flaxman, 2017; Floridi et al., 2018; Hind et al., 2019; Kaur et al., 2020; Tjoa and Guan, 2020). Arya et al. (2019) presented a taxonomy of types of explanation types and explanation methods and speculated on how the different types would be more helpful to loan officers vs. loan applicants vs. bank executives. Langer et al. (2021) listed 28 explanation requirements such as acceptance, confidence, fairness, and performance. These are mapped onto four stakeholder types: users, developers, regulators, and affected parties. Attesting to the burgeoning interest in this topic, Langer et al. (2021) cited over 100 articles that "claim, propose, or show that XAI-related research (e.g., on explainability approaches) and its findings and outputs are central when it comes to satisfying" explanation requirements (p. 7).

Hind et al. (2019, p. 124) offered a speculative mapping of stakeholder groups onto different explanation requirements: End-users (decision-makers) need explanations that let them build trust, and "possibly provide them with additional insight to improve their future decisions and understanding of the phenomenon." Affected parties need explanations that enable them to determine whether they have been treated fairly. Regulators need to know that decisions are fair and safe. System developers need to know if the AI is working as expected, how to diagnose and improve the AI, and possibly gain insight from its decisions. Hind et al. (2019) asserted that all explanations of AI systems must present a justification for the AI's decisions, contribute to user trust, include information that the user can verify, and rely on the domain concepts and terminology. They also asserted that the complexity of an explanation must match the "complexity capability of the user," but this interesting speculation is not analyzed.

Weller (2019, p. 2–3) presented a mapping of stakeholder groups onto different "types of transparency," which might be understood as types of explanations. Developers need to understand how their system is working, to debug or improve it, or to see what is working well or badly and get a sense of why. Users need to have a sense of what the AI is doing and why, need to feel comfortable with the AI's decision, and need to be enabled to perform some kind of action. Experts/regulators need to be able to audit a decision trail, especially when something goes wrong. Members of the general public need to feel comfortable so that they can keep using AI.

Although the majority of work in this area has centered on taxonomies, these notions have begun to make their way into products. For example, IBM's AI Explainability 360 demonstration (IBM, 2021) has distinct interfaces for three different "consumer" types (bank customer, loan officer, and data scientist), which offer increasingly complex perspectives on decision rules as well as moving from case-based examples as explanations to large-scale views of underlying data in a loan application scenario. Although the IBM demos work primarily to suggest potential interfaces and algorithms appropriate for different stakeholders and do not prescribe or dictate the interface, they are a concrete example of how explanations might be tailored to different stakeholders and roles.

All of these researchers recognize that explanations differing in form and content will have explanatory value for different stakeholders or role-holders. All of the researchers emphasize the need for XAI system developers to consider the intended beneficiaries of explanations, their domain, tasks, goals, and contexts.

2. The challenges of categorization

2.1. Requirements or desirements?

While the citations above refer to explanation requirements, the expressions are not anywhere close to build statements, that is, they are not directly actionable engineering guidance. Thus, in this article, we refer to "desirements" (Hoffman and Elm, 2006; Hoffman and McCloskey, 2013). This is somewhat different from the notion of "interpretability desiderata" (Lipton, 2018; Sokol and Flach, 2020; Langer et al., 2021). Desirements are expressions by a worker (user, operator, etc.) of functionalities or capacities that they wish they had, and that they believe would improve the work. Desirements are not immediate solutions to the problems of system design; they are pointers to the problems and suggestive of possible paths to solutions.

2.2. Stakeholder groups or roles?

The concept of the stakeholder was used as the entry point since that term was current in discussions of XAI at the time the interviews were conducted. XAI system development was focused on providing explanations to "end-users," and the question is called as to whether other stakeholders might need different explanations from those intended for end-users. However, many of the proposals for categorizing stakeholder groups have been speculative. Do different stakeholder groups actually require different kinds of explanations? As we will show, our results indicate that this is true only to a limited extent. Individuals who would fall into the same stakeholder group or category can nonetheless have different roles and responsibilities (which are typically in flux) and therefore would have different sensemaking needs and explanation requirements. The distinction between a stakeholder group or category and specific roles within or across categories is also important because an individual might serve in more than one role. A developer might also serve as a trainer. Furthermore, the roles adopted by an individual might cut across stakeholder groups. For example, an end-user might also have the skills and motivation to do some software development. While an individual serving in a role might have similar explanation requirements to an individual in some other role, their differing roles may bring with them different purposes, and these could color their sensemaking. This complexity may make it impractical to try and pin a single specific form or content of explanation onto a particular stakeholder group (see Hepenstal and McNeish, 2020).

As we will show, these considerations percolated in our findings.

2.3. Approaching the matter empirically

While there is value in all of the articles cited above—their rosters of stakeholder types and their sensible mappings of stakeholder types to explanation desirements—there have been only limited attempts to investigate these matters empirically. Hepenstal and McNeish (2020) derived explanation requirements from a focus group of professionals. Kaur et al. (2020) surveyed data scientists. Liao et al. (2020) surveyed user experience and design professionals. Bhatt et al. (2020) reported on the outcomes of a workshop involving representatives of industry, academia, jurisprudence, and policymakers. The focus of the workshop discussions was on identifying challenges and shortcomings with regard to machine transparency. Workshop participants presented speculations about "the capacities of different communities to engage with explainable AI" (p. 3). In general, the participants pointed to a number of needs:

- The need to determine the effectiveness of machine-generated explanations and how users understand the explanations;
- The need for participatory development (i.e., bringing developers together with experts in human–computer interaction);
- The need for community engagement (i.e., involving users in design and development);
- The need to educate stakeholders regarding machine explainability, especially its limits;
- The need for stakeholders to understand uncertainty and the failure modes of AI systems;
- The need for a capability to enable stakeholders to "toggle the information in an explanation" (p. 5);
- The need for explanations to match with the actions that the stakeholders might take or might need to take.

Rather than adopting a survey method, we conducted in-depth structured interviews with selected individuals. The Stakeholder Playbook is a synthesis of our results, and we present that following a discussion of our method.

3. Method

3.1. Participants

An attempt was made to solicit the participation of professionals who represent diverse stakeholder groups. Participants were 18 professionals (16 male and two female) who had experience with AI and/or autonomous systems (not just Machine Learning systems). The group included former military, civilian scientists working for the government, scientists working in the private sector, and scientists working as independent consultants. Participants were all either mid-career or senior professionals, and all had postgraduate degrees. Participants were recruited by soliciting individuals with appropriate experience and expertise, via relevant industry and professional contacts of the researchers. Participants were not paid to take part in the study.

Table 1 describes the kinds of AI/Autonomous systems with which the participants had experience. As this table shows, the pool of participants had experience with a diverse range of systems.

The Appendix presents the Participants' demographics: Age, degrees, current and previous job description or title, and current and previous self-identified role(s). The participants represented diverse roles. All but two of the participants had experience in more than one role. We had a participant who had risen to the post of development team leader and had been trained in cognitive science as well as computer science. Two participants had been trained in experimental psychology, but moved into applications, becoming cognitive systems engineers and system developers, and one participant had been trained in industrial systems engineering but moved into cognitive systems engineering and the role of systems developer. One participant had a background in human factors of AI applications. Another participant had been trained in mathematics but moved into system design. Four participants self-described as end-users, although their current primary role was not that of an end-user. Thus, occasional comments by a participant might be from their previous perspective as an end-user when their primary current role was, say, that of a developer.

3.2. Procedure

Participants were asked to participate in an interview concerning the explanation and understanding of AI systems. The consent form expressed the topic this way:

> "It is sometimes said that Artificial Intelligence systems are 'black boxes.' One cannot see their internal workings and develop trust in them by understanding how they work. The general goal of this research is to adduce information about how people understand the AI systems that are used in their workplace. This information will enable us to tailor the explanations of how the AI works, depending on the individual's role, responsibilities and needs."

The procedure was a structured interview. It was influenced by the methodology of Grounded Theory (Glaser, 1998), which emphasizes the importance of assessing interviewee statements from multiple perspectives. It was also influenced by cognitive task analysis (Crandall et al., 2006) in that the interview questions referenced the cognitive demands imposed on a given role. First, there was a brief discussion of the ambiguity of "AI", to achieve some common ground. Next, there was a brief discussion of the idea that various stakeholders might have particular needs for explanations of how an AI system works. This discussion had the purpose of determining which "hat" or "hats" the interviewee was comfortable representing (i.e., roles). Then, three demographic questions were asked (age, educational background, current job title or description, and current responsibilities).

The pre-planned questions were created by the researchers, with the intent of getting at three main aspects of AI use: Interviewee experience with AI systems, interviewee experience

TABLE 1 The types of AI/autonomous systems with which each participant had experience.

ID	Previous/current roles	Types of AI systems experienced
1	Supports legal professionals and regulators	NLP systems for dictation; spam filters
2	Supervisory role in human-systems integration and evaluation; acquisition support	Tactical radios, command and control systems; machine learning systems
3	Contract formation, defense in litigation	Command and control systems
4	Electromagnetic warfare; end-user; system evaluation; capabilities development	Battle management systems; tactical radios
5	Conflict analysis and resolution; data science legal issues	Legal applications for client decision-making
6	Data analytics; software development; end- user; guidance to large organizations on conflict resolution	Predictive modeling systems for intelligence analysis
7	System integration, AI system policy; manpower; human-machine teaming	Image recognition; UAV control
8	Guidance for businesses to implement strategy and tactics.	Aircraft systems; counterintelligence systems
9	Development team leader; systems analysis (human-automation systems); applications (of AI)	Simulation-based training systems; intelligence analysis; decision aids
10	System Development; Modeling and Simulation Systems;	Predictive modeling at the organizational level
11	Intelligence analysis	Decision aids; command and control systems; anomaly management
12	System development; system evaluation	Mission planning systems, visualization systems; planning optimization systems; systems evaluation
13	System development; system evaluation, development team lead; program management; user experience evaluation	Network systems, robotic systems; NLP systems; information management systems
14	Development team leader; design based on human factors; user experience analysis; system evaluation (usability)	Command and control systems; course of action analysis systems
15	Development team leader; system acquisition; system development (design)	Visualization systems, human-automation collaboration
16	Knowledge management	Autonomous systems; command and control systems; decision support systems; systems evaluation
17	Strategic planning system evaluation	Cyber defense systems; intelligence analysis systems
18	Development team lead; program evaluation	Anomaly detection systems; pattern recognition systems; NLP systems; machine learning systems

with explanations of AI systems that were provided, and interviewee desirements with regard to their understanding of the AI systems.

1. Are there any AI systems in use in your organization (briefly describe)?
2. Do you yourself rely on an AI system in order to do your job?
3. If so, how was it explained to you how it works?
4. Do you want to know more about how AI systems work in order to employ them better?
5. What do you feel you need to know about how an AI system?
6. What do you feel you need to know about an AI system in order to properly exercise your responsibilities?
7. Can you briefly describe any experiences you have had with AI systems where more knowledge would have helped?

The seven questions were all to the point, but somewhat abstract. As in other applications of cognitive interview methods, whenever a participant mentioned a particular experience or illustrative case, the interviewer encouraged the participant to provide more details. As each interview proceeded, there was an adjustment to the wording of some of the questions, to refer to the participant's role. For example, for Question 6, jurisprudence professionals were asked about legal issues regarding AI that were of concern. System developers, system integrators, and program managers were asked about their explanation needs to properly procure systems or manage system development.

The interviews took between 15 and 50 min, averaging 23 min. The interviews were audio recorded, with permission. Transcriptions were created from the re-codings, and then the audio files were erased. All personally identifying information was removed from the transcriptions. The participants were allowed to review the transcript of their recording and edit out anything they wished.

4. Results

The approach taken in qualitative analysis of ethnographic or cognitive interviews involves assessing participants' statements according to themes that emerge from more than one perspective or categorization scheme (Hutchins, 2003; Schoepfle, 2021).

The first perspective was given in the topics of the interview questions themselves. The second perspective was that of the stakeholder groups or roles with which each participant had identified. That perspective revealed that some desirements were shared across the stakeholder categories. This suggested "role clusters"—where desirements are shared by individuals in different roles. Individuals in one role might want the same things explained and explained in the same way as an individual in another role.

The themes that emerged in the interviews were:

(1) Sensemaking desirements and the challenges faced by the Explainer,
(2) The challenges of self-explanation,
(3) The challenges of explaining AI systems to others,
(4) Trust and reliance issues, and
(5) Challenges for design and procurement.

These themes are illustrated in quotations from the interview transcripts, presented in Subsection 4.1 and Section 5. In addition to the themes, a distinction emerged between explanation desirements, access requirements, and cautions. Access requirements are actionable requirements for information access, such as reaching out and seeking explanations or further explanations. Cautions are things that the explainer (or the XAI system) needs to be cautious about. These three categories are utilized in the Playbook given below. A Supplement to this article is available from the authors upon request. It presents examples of quotations expressing the themes, categorized in terms of the role clusters.

The Playbook presented in Table 2 is a synthesis that references stakeholder groups, but as explained above, this is to be understood with respect to roles or role clusters, since an individual might self-identify as a representative of a particular stakeholder group but serve in more than one role.

The most meaningful way of conveying the results is to provide succinct and illustrative quotations from the interview transcripts. This is the approach taken in the following subsections.

4.1. Key finding: the importance of self-explanation

A simple model of the XAI explanation process is:

(1). The XAI system generates an explanation,
(2). The explanation is provided to the user,
(3). The user understands the explanation,
(4). Performance improves.

This "spoon feeding" model is incomplete. This is implied in The Playbook. Even when presented with good explanations, people engage in a deliberative process of sensemaking, or self-explanation (Chi et al., 2008; Lombrozo, 2016). The purpose of self-explanation is to enable the learner to develop a richer mental model, but also to satisfy curiosity, make more accurate predictions of the AI's behavior, achieve appropriate trust in the AI, and/or improve performance by using the AI as appropriate.

TABLE 2 The playbook.

Jurisprudence
Explanation desirement: analysis of system biases, assumptions, and bounding conditions.
Explanation desirement: description of the features upon which the system relies.
Explanation desirement: must be able to "look under the hood."
Explanation desirement: must be able to explain the system to clients.
Explanation desirement: must be able to explain the benefits of the system.
Access requirement: to the system development team—trusted software engineers, mathematicians.
Access requirement: to experienced and trusted domain practitioners.
Access requirement: to succinct background information on computer science and the pertinent AI technology.
Contracting; procurement
Explanation desirement: global explanation of "how it works" and how the data are processed.
Explanation desirement: global explanation of architecture and functionality (how the data are processed.
Explanation desirement: analysis of system biases, assumptions, and bounding conditions.
Explanation desirement: description and explanation of the data that were used to train the model.
Explanation desirement: analysis of the model's fitness for the data that are used in the operational environment.
Explanation desirement: analysis of the confidence or applicability of the system for the particular questions that are being asked.
Explanation desirement: assurance of data quality and curation.
Access requirement: trusted software engineers and domain practitioners.
Access requirement: access to trusted vendors, who do not "dumb things down."
Access requirement: leads with technical background need access to explanations of technical details.
Program manager; development team lead
Explanation desirement: global explanations of "how it works."
Explanation desirement: description of the data that the ai ingests; assumptions about the data.
Explanation desirement: analysis of how the system will be integrated with other systems in the broader work system.
Explanation desirement: analysis of system strengths, weaknesses, and bounding conditions (system assumptions).
Explanation desirement: analysis of how the system will be integrated with other systems in the broader work system.
Access requirement: to trusted software engineers, mathematicians.
Access requirement: to experienced and trusted domain practitioners.
Access requirement: to group discussions among developers and management.
Access requirement: to trusted vendors, who do not "dumb things down."
Caution: some program mangers and development team leads will need to know the details of the system processes and algorithms.
Developers
Explanation desirement: needs to determine the optimum balance between more data and its marginal value.
Access requirement: to a corpus of use cases that that are representative of the implementation contexts.
Access requirement: opportunity to explore the system by working between specific examples and global information; manipulate the inputs and see the outputs.
System integrator
Explanation desirement: explanation at the detailed technical level; needs to be able to "look under the hood."
Access requirement: to trusted software developers.
Trainer
Explanation desirement: rich corpus of edge cases.
Explanation desirement: needs to be able to achieve an understanding that is sufficient for them to be able to explain the system to trainees.

(Continued)

TABLE 2 (Continued)

Explanation desirement: needs to be able to achieve an understanding that is sufficient to allow them to help users understand the edge cases, and when a situation is approaching an edge.
System evaluator
Explanation desirement: explanation of the inputs, outputs and their relations. Explanation desirement: information of how the system manages the trade-offs in operational conditions. Explanation desirement: information supporting the design of usability and performance tests. Explanation desirement: operational definitions of the proposed "metrics" (measures) to be used in performance assessment. Access requirement: to trusted system designer. Access requirement: to the system developers when the system does something bizarre or unexpected. Access requirement: to an established network of experienced users to support self-explanation. Access requirement: feedback from prospective users. Caution: not all evaluators need to understand the technical detail (e.g., algorithms).
Policy maker
Explanation desirement: global explanation that is satisfying and consistent with how the system actually works. Explanation desirement: demonstrations that cover a range of examples to show the results based on different input conditions. Explanation desirement: descriptions of system biases, assumptions, and bounding conditions. Explanation desirement: descriptions of system limitations and weaknesses. Explanation desirement: demonstrations that include edge case scenarios.
End-user; adopter
Explanation desirement: global and local explanations that are satisfying and consistent with how the system actually works. Explanation desirement: explanation of data inputs and how the system processes the data. Explanation desirement: results of a cost-benefit analysis of different tools, with respect to the user's goals and responsibilities. Explanation desirement: explanations need to strike a balance between superficiality and technicality. Explanation desirement: explanations that support troubleshooting and system maintenance. Explanation desirement: intuitive displays for visualizing the data, to understand whether the data might be inadequate. Access requirement: to the system development team— trusted developers and software engineers. Access requirement: ability to explore the system behavior by "poking around." Caution: end-users sometimes do not care or don't need to know "how it works." Caution: end-users often desire better explanations than the ones that are provided. Caution: continuing explanation is required as the input data, the work system context, or the operational environment change.

Some of our Participants' comments point to the fact that the self-explanation is often triggered by the inadequacies of the explanations that are provided.

> *The superficial answer–it takes into account these variables, and chugga-chugga, and we're not going to explain that, and then the other [explanations] go way down into the weeds and not necessarily for you a direct path to understanding how it works. And so you just dive in and use it, right? Honestly, for me I'm just very interested in about how a lot of these tools work. So I don't like using something that I can't explain. So I am giving an explanation to, say, a commander. "I do not really know how this answer came out, but this is the answer." So a lot of times I'll either poke around with it about how it works, see if I can find out–if not exactly how it works, [I find] the things that make it tick. If I change x, what's going to change in the model? If I change y... A crude example of a kind of sensitivity analysis. See if anything is popping out at me. Sometimes I'll use it, back it up with my own analysis, to see if I can say "Hey, I looked at it both myself and I used this tool and we saw that either there are similar answers or not." Or, if I do not understand what it is I'll not use it [chuckles]* (P6).

Because spoon-fed explanations are often deficient and insufficient, people do not want to see more such explanations. What they want is a richer understanding. Understanding is the goal of the sensemaking process, not the comprehension of a piece of text or the perception of a saliency map. All of the participants referred to the need for knowledge, but this was not expressed as a need to be spoon-fed more or better explanations.

> *The universe in 30 seconds. [You] need explanation about the data that was used to train the model, the model's fitness for the data that is used in the production environment. Need to understand the fitness between what the model was trained on and what you are analyzing today. Need explanations of the confidence or applicability of the algorithm for the particular question that is being asked. Need automated support for visualizing the data, to understand how the data you are giving the tool might be inadequate* (P17).

Many of the Participants said they would obtain that understanding by exploration, reaching out, and self-training:

> *[I] test the tools to figure out how they work* (P6).
>
> *When I needed to I could go to them* [engineers] *for explanations* (P4).
>
> *It comes down to how much faith you have in the individual* (P16).
>
> *[I got] explanations in the past. I sought them. Face-to-face, verbal. For the most part, satisfying. Global. And you had to seek them out; they were not provided anywhere. Found help via professional networks* (P17).
>
> *I have to know how it works in order to get feedback from the operators. If I don't know how it works I do not know how to frame the usability test to collect the data in order to improve the system as we are developing it* (P17).
>
> *The training is baked into the system; training is embedded in the software. So if someone does not know how to use it, they need to be able to train themselves on how to use the tool* (P2).

A number of participants commented about how they preferred to manipulate ("poke around") and explore the AI system behavior under different scenarios, to "get a feel for it." Role-holders want to be provided with more examples of the AI encountering different situations; explanations that are exploratory rather than discursive: The visualization of tradeoffs (e.g., in a scheduling algorithm) would support appropriate reliance and the capacity to anticipate

when anomalous events occur and the recommendation may be misguided.

What role-holders want is empowerment. End-users and other role-holders need sufficient global and local (case exemplar) information to enable them to self-explain, to their satisfaction. They want to be able to discern where and how to explore and where and how to reach out to others. This key finding is manifest in the details in the Stakeholder Playbook.

We now discuss some results that were surprising, if not counterintuitive.

5. Surprises

We asked our participants *How was it explained to you how AI systems work?* and *Can you briefly describe any experiences you have had with AI systems where more knowledge would have helped?* The answers we received to these questions revealed a somewhat complex and even subtle picture, one that brings to the fore the cognitive aspects of machine-generated explanations. While some of the "surprises" in our results may not be surprises to some developers of AI and XAI systems, they are surprises with respect to the assumptions that have been made in the field of XAI. For example, it has been assumed in many XAI activities, including the DARPA Explainable AI Program, that end-users and other stakeholders all need explanations. Some XAI developers made a further assumption that explanations should always be provided. Indeed, many XAI systems have a persistent window in their interfaces, in which explanations are always presented. Another assumption that has been made is that once a sufficiently good explanation has been presented, that marks the end of the explanation process. As we now describe, these assumptions are at odds with our findings.

5.1. Not everyone actually needs or wants an explanation

Some participants commented that they have gotten "canned" explanations that are good. However, the surprise was that participants often said that they do not care or do not need to know "how it works" (P2, 5, 8, 10). Only three of the participants explicitly said that they did want explanations of how the AI works (P9, 11, 15). One participant (P2) asserted that they do not need explanations. This came as a surprise, insofar as it suggests not all roles entail a need for explanations. Four participants asserted that they (or other role-holders) do *not* need to be able to drill down into the technical details of how the system works (P2, 8, 12, 18). As we planned the interviewing, we expected that all of our participants would say they want and need explanations, and better explanations than the ones they are typically provided. This expectation on our part was largely due to the fact that the "explanations" we saw being generated by XAI systems seemed exclusively local and technocentric (e.g., heat maps, matrices of feature weights, decision trees, etc.).

5.2. Sometimes different stakeholders need the same explanations, not different ones

The commonly discussed motivation for making explanations role-specific is that different role-holders will need different things explained, and explained in different ways. One surprise in our findings was that this is not the case.

As can be seen in Table 2, above, every one of our participants wears at least two "hats" even though they identified as a representative of a particular stakeholder group. This is the main reason for referring to roles as well as stakeholder groups. More than this, there are what we call "role clusters." These emerged out of the participants' responses themselves when explanation desirements are shared by individuals in different roles. Individuals in one role might need the same things explained and explained in the same way as an individual in another role. These "clusters" cut across stakeholder groups and are utilized in Tables in the Supplement to this article.

5.3. After getting an explanation, many people want more

When we asked our participants, *Can you briefly describe any experiences you have had with AI systems development where more knowledge would have helped?*, three participants said *All of them* (P9, 11, 15), two participants said *Yes* (P10, 19), another participant said *Absolutely* (P13), and another participant said *Yes, always actually* (P15). These responses were terse, immediate, and strident. Other responses were less terse and more informative:

> *What I have found is it is usually either they are going to say the system just uses machine learning, or they say "We take these variables and we give you the answer"* (P6).
>
> *A lot of the explanations I have heard, always stop short. I need to know exactly what the AI is thinking when it gets into a scenario and something happens* (P16).

Another participant affirmed the question but was less conclusive about it:

> *[There] probably were times when I did have to deal with an AI and did not know a lot that was going on* (P8).

What was surprising was that we did not get strident affirmations of our probe question from all of our participants. In response to our probe questions, one participant responded by saying: *Nothing immediately comes to mind. It develops as you drill down into the details* (P7) and another participant denied the need for a deep explanation: *I don't care much about the details of the algorithms. No situation where I felt a need to understand the algorithm in detail* (P12).

5.4. Stakeholders need to think about the sensemaking needs of others

We were surprised to find that assertions about the sensemaking needs of stakeholders *other* than themselves were more frequent than affirmations of their own need for explanations (P1–7, 10, 12, 13, 14, 15, 18):

> *Now that I've worked with.. but I know lots of law firms who are working on products and is not obvious to the outside world how the AI works*) (P1).
>
> *The practitioners try to get into how the AI/ML system works* (P2).
>
> *This is so funny.* [Laughs] *Yes. I have seen, for example even with networking systems. If the users understand how they work, it is easier to do troubleshooting and preventative maintenance checks. If they don't understand it, they are not going to be able to use it* (P14).

And muddying the waters yet further, four participants denied that a certain other stakeholder needs their own particular kind of explanation (P2, 5, 8, 18).

5.5. Stakeholders need access to people

When explanatory information is not available, or is either superficial or so detailed that it is not useful, the challenge of self-explaining often becomes a matter of access to the right people who are in possession of the right information. The motivation to self-explain is manifest when the available local resources are inadequate:

> *I've gone to YouTube channels and had some random person from across the world explain some really complicated topic explanations and because their entire goal is to explain the modeling process, I understand it better than anyone else has ever explained it before. There's little things like that. You can see what other people are doing to explain it. It's like, "Oh, wow, I can learn this better from YouTube than from anybody else"* (P6).

The motivation to self-explain is often manifested as active reach-out to the developers:

> *I sought them face-to-face, verbal. For the most part, they were satisfying. Global explanations. And you had to seek them out; they were not provided anywhere. I found help via professional networks* (P17).
>
> *I have to know how it works in order to get feedback from the operators. If I don't know how it works I do not know how to frame the usability test to collect the data in order to improve the system as we are developing it* [P14].

A number of our participants commented that the establishment of a sufficient and satisfying understanding can *only* be achieved via discussions with the system's developers:

> *[I'm] glad to hear that other people have these experiences. I'll look around the room and think, "Am I the only one that's not getting this? Am I the only one who is confused?" One of the things that has been at least helpful for me is, a couple of things we have worked with the development people, and we were closely tied with the actual developers. Rather than interfacing through the salesmen, I would go and sit next to the developer and work with them. And they were very helpful. Even having a friendly relation with them, it's a lot easier to ask a lot of dumb questions, explain it very slowly, look at examples, stuff like that. That's been a good experience for me. I know it is not realistic in a lot of cases* (P6).

Participants expressed a need to be able to explain AI systems to others (other stakeholders); their effort at self-explanation often serves as useful feedback to developers:

> *Additional explanations, beyond the demos, provided by the developers. Wider range of examples to show the results based on different input conditions. Putting it through the rigors of the wider range of potential situations that might encounter. We often ended up doing it for the developers through our simulated exercise. Even repeated scenarios are not exactly the same every time* (P11).

Because of its importance and salience to our participants, the Stakeholder Playbook includes "Access Requirements" as well as "Explanation Requirements."

5.6. Sometimes you just have to trust

Our participants' responses seemed genuine: *I never totally understood how it worked, but I worked with it enough to trust it* (P7). All but one of the participants mentioned trust issues, which is not surprising given that trust issues are closely related to the challenges of explanation, and trust was mentioned in the interview instructions. But trust is not always something that develops, and trust is sometimes a "default":

> *A lot of it you take on face value because you are not going to become an expert End-user in all the beeps and squeaks that go into it. You can't have users dive into everything. They just want to know that it is going to do what they are asking it to do, and the more they see that, the more they develop trust* (P7).
>
> *It does things sometimes you don't know* (P4).

When trust is not default, it is very often tentative or skeptical, and different role-holders have different default stances along a spectrum of trust–distrust when using a new system:

> *People within our own organization say "Hey, this is really awesome." But then I would say "Yeah, but how much work am I going to put into that to get this tool to answer my question?"* (P6).
>
> *You have to overcome the initial trust hurdle early on, overcome it with training* (P7).
>
> *There are some scenarios where you know it was going to act but not which way it was going to act* (P7).

> *The challenge was figuring out what the system was going to do. The operator had to figure that out, to know what it was going to do (because you were not going to be able to affect it)* (P7).
>
> *Where you feel you have an input, and it does not give you back what you believe, it confuses more than helps* (P8).
>
> *I learned early on that all of the systems were brittle, even systems that were considered a success–some evidence that they were resilient, adaptable. But I knew they were brittle but did not really know why* (P9).
>
> *It comes down to how much faith you have in the individual [who is providing the AI system]. Is the decision the machine makes the same as I'd make, or does the AI think better because it is aware of other things? You need to trust the brain behind it, and their understanding of how it works, what its rules are* (P16).
>
> *I am willing to start out trusting, it, until it proves itself to not be reliable. Then it becomes trust but verify. Trust and Reliability feed each other. You start off trusting and watch reliability over time. Or you can start relying on it. Is a matter of loss of trust. But if it is life or death, you do not start with trust, but if it is a simple decision like data analysis, you can start with trust* (P16).

5.7. Trust is not just in the AI

Trusting extends across stakeholder groups. For example, trust in the vendors trumps the need to "poke around." Conversely, mistrust in the vendors (because of their over-promising) trumps adopters' attitudes about the AI's reliability and trustworthiness. Vendors need to instill confidence in the prospective adopters and users, even if the vendor cannot explain the "under the hood." Trust extends to other people and entire organizations.

> *Contractors promise us the world. And then it turns out, they do not necessarily have that. So some of the tools that we would work with didn't really end up doing anything* (P6).
>
> *People within our own organization say "Hey, this is really awesome." But then I would say "Yeah, but how much work am I going to put into that to get this tool to answer my question?"* (P6).
>
> *Companies were trying to push AI but no one wanted to hear about the limitations, constraints, and brittlenesses. The companies try to push things further, but no one wanted to talk about the warts or the limitations* (P11).

And trust is not just in the AI, it is especially in the data on which the AI was trained. Rather than expressing a desire to understand the inner workings of the AI itself, more common was participants' expression of a need to know about the input data or the data features the AI processes, or a need to see the input–output relations in additional scenario demonstrations (P1, 3–5, 7, 8, 11, 14, 16–18). Understanding the data the AI system uses would seem more helpful than poking under the hood to examine the innards of the system. They wanted to know what data were used to train the AI/ML. They want to know about any system biases attributable to the data. They want to know what data were used for a specific project or decision, and they want assurance that there is a match between the data inputs and the situation—if the AI/ML has been trained on or is using the wrong data, the outputs cannot be trusted.

> *Need to know whether there is bias, and then what those biases are. It is necessary to know about the data that go into the AI system* (P3).
>
> *You need to know what data it evaluates, in order to quantify its uncertainty. Data are not free. Data for AI ingestion is often not properly curated, tagged, etc. That is not cheap, it is certainly not free* (P7).
>
> *I'd seen that multiple times, and know I can't rely on that data but this other part of the data is valid.*

5.8. Explanations that are provided are rarely in a goldilocks zone

When an explanation is desired, the explanations that are provided are generally regarded as inadequate. Five participants asserted that the explanations they provided were always at either too low a level or too detailed (P6, 9, 11, 13, 15). Either way, the role-holder needs to reach out for more information. Many of our participants referred to their active reach-out to other people (e.g., Developers and other End-users) and other sources (social networks and YouTube) to enrich their understanding of AI systems (P4, 6, 7, 9, 10–17). XAI research has involved the creation of explanations that take many different forms and that express different kinds of content. Formats include diagrams, heat maps, matrices of feature weights, and logic trees. Our participants did not volunteer any opinions about such formats, except for the occasional reference to the need to "visualize" data. What they did refer to, and often, was a dialog with others, in which they sought out and received explanations.

For individuals who are not particularly computer savvy, global explanations can take the form of reductive but clear analogies. The value of explanation-by-analogy is crucial in scientific reasoning and problem-solving. Explanation by analogy has been underplayed in the XAI work. We were surprised that analogy only appeared once in our interviews when a Participant referred to a "sitting kids in a school bus" analogy to describe how ML systems work.

5.9. "Global vs. local" is not clear-cut

The distinction between global explanation (*How does it work?*) and local explanation (*Why did it make this particular decision?*) has been a key consideration in the literature on explanation and in the work on explainable AI (see Miller, 2017). One of our participants (P17) asserted that they only need global explanations, not local ones. This global focus appears widespread, as most comments by participants relate to a desire to know *how the system works,* rather than wanting to know *why it made a particular decision.* This is in stark contrast to the research in XAI, which focuses on local explanations and justifications of particular decisions or actions of an AI system (see Mueller et al., 2019).

Previous research has shown that global explanations are often accompanied by specific cases and that local explanations contain hints that contribute to global understanding (Klein et al., 2021). In other words, people benefit from having global and local explanations that are integrated. Participants in the present study made more reference to global explanation than to local explanation, yet the majority of XAI systems of which we are aware have focused on the delivery of local explanations.

Our participants' frequent reference to the need to see "edge" cases underscores our finding that explanatory value can derive from material that blurs the global–local distinction. Comments from our participants underscored this finding. For example, one participant (a developer) said he benefitted from having global and local explanations that are integrated.

> *Going from the specific allows me to go to the general. Enough specific examples allow me to accept that there is a general explanation (P10).*

5.10. Stakeholders can be quite interested in "deep dives"

It has been assumed in XAI activities that end-users and other stakeholders do not want and are not prepared to understand highly detailed or technical explanations of what goes on "under the hood." Our findings show that this is not always the case. Although end-users and user team leaders do not always have sufficient knowledge of computing concepts, our findings show that role-holders are often sufficiently versed in computer science to enable them to do deep dives inside the ML system (e.g., why it makes certain kinds of errors). Thus, some individuals sometimes want detailed technical explanations (P1, 4, 6, 12, 14, 15). They recognize the value in their being able to do deep dives. Our findings show that stakeholders' cross-disciplinary skill is perhaps more common than might be supposed. Yet, there are certain circumstances in which a stakeholder has little understanding of AI and is frustrated because they might not know exactly how the AI is making the decisions (e.g., P3).

5.11. Sensemaking by exploration is of greater importance than prepared explanations

Individuals in all roles actively self-explain. The end-user has to engage in an exploratory effort to self-explain the AI because inadequate information is provided or not enough information is available. Half of our participants asserted that they (or other stakeholders) are self-motivated to actively develop good explanations of "how it works" (P3, 4, 6, 7, 12, 13, 15, 17). Eleven of the 18 participants asserted that there are circumstances in which they (or other stakeholders) need to be able to actively seek and then drill down into the technical details of how the system works (P4–6, 10, 12–15, 17, 18).

It is widely recognized that contrastive explanations are valuable: "If X had been different, what would the AI have done?" or "Why did the AI decide A and not B?" (Miller, 2017). Our participants made two kinds of contrastive statements: (1) statements about the need to understand how the AI would perform if the input data were of questionable quality and (2) statements about the need to understand how the AI performs when dealing with edge cases. The participants took it for granted that the AI would do something differently if the data were different. The contrastive explanation did not take a logical form for our participants but was exploratory. A number of participants commented about how they preferred to manipulate ("poke around") and explore the AI system behavior under different scenarios, to "get a feel for it." Individuals want to be provided with more examples of the AI encountering different situations. End-users would benefit from local explanations that are exploratory: The visualization of tradeoffs (e.g., in a scheduling algorithm) would support appropriate reliance and the capacity to anticipate when anomalous events occur and the recommendation may be misguided. Global explanations are not just for understanding—they guide and enable the search for other information and resources.

5.12. Stakeholders are as likely to need to know about the data as they are to need to know about the AI system that processes the data

Rather than expressing a desire to understand the inner workings of the AI itself, more common was participants' expression of a need to know about the input data or the data features the AI processes, or a need to see the input–output relations in additional scenario demonstrations (P1, 3–5, 7, 8, 11, 14–18). Understanding the data that the AI system uses is felt to be more helpful than poking under the hood to examine the innards of the system. They wanted to know what data were used to train the AI/ML, about any system biases, what data were used for a specific project or decision, and they want assurance that there is a match between the data inputs and the situation—if the AI/ML has been trained on or is using the wrong data, the outputs cannot be trusted.

5.13. Stakeholders need to know about how the XAI interfaces with other systems

Participants expressed interest in explanations that are not about "the AI" but rather about the overall system architecture and business logic of the system. For example, what are the plug-ins? How is it interfacing with other systems? How is it connecting the different units within an organization? How is it synthesizing their data? What is the interplay of the components? What is the AI/ML accomplishing? What is the cost/benefit tradeoff of using the system?

5.14. Explaining is never a "one-off"

An additional limitation of prepared explanations derives from the assumption that an explanation, especially a global one, is provided once, likely during training or at the beginning of the operational experience. Explanations are often context-bound, trust is always tentative, and explanation is not a process that terminates. Explaining must be engaged frequently, especially for AI/ML systems that learn and change (improve).

> *Processing tools are very complicated but highly reliant on data streams and samples, need to understand what you are feeding it. For example, data collected from sensors undergoes initial processing that may result in gaps, may be collecting the wrong kinds of data, data may be skewed and mislead the AI. I need to understand what the data look like that are going into the model. I might be able to tune the sensors, or put context-specific labels on the data* (P17).

5.15. "We set a high bar for XAI systems"

Participants referred to stringent criteria. Incremental gains just are not worth it. The AI system has to be a game-changer. One participant asserted that if a novice user cannot perform 85% of the key tasks on the first test, the system will not be adopted. A number of the Participants' comments were candid expressions of the actual sentiments of operators:

> *Can a user, without training, figure out how to use the system within 10 min? If they fail at that, they don't use it* (P2).
>
> *I don't like using something that I can't explain. In some few cases I did get at least enough understanding that I felt comfortable for explaining to someone else. I give them an answer and I say, "I used this tool and got this answer. I'm not the expert obviously, but here's the general idea of what happened and why it makes sense."* (P6).
>
> *The tech has to be a game changer, not just an incremental improvement, it has to be a substantial improvement in performance because it is so hard to introduce and maintain new technology. The tools have to be a leap-ahead* (P5).

5.16. Trainers need to be able to train end-users on "how it fails" and "how it misleads" (limitations and weaknesses)

Trainers—and individuals in other Stakeholder roles—need to have an understanding that is sufficient to enable them to explain the AI system not to themselves but to other people. Trainers need access to a rich corpus of cases that are representative of implementation contexts, of course. But they also must be able to train end-users to engage in maintenance and troubleshooting activities. Trainers must be able to train end-users to sense how quickly they will enter a gray area in various scenarios. The XAI needs to do more than explain the "why" of particular decisions: It needs to be able to give the user advanced knowledge of when the work system is approaching an edge case.

Training on edge cases is crucial for the establishment of trust in AI:

> *What you are trying to find is where those edge cases where it hasn't been fully vetted... it's reaching the limits of its data that it using to inform its decisions. That's what everyone is going to be concerned about. If it lives in the heart of the black box the whole time, then you will have adequate trust it will perform the way it is expected to, but for your policy makers and your legal, they will certainly be interested in what those edge cases are. And the users will be, depending on how quickly they reach those edge cases in one of the scenarios* (P7).

The identification of edge cases is exemplified by studies of the errors made by Machine Translation (MT) systems. Knowledge about the kinds of things that cause problems for MT systems is a powerful enabler of how errors can be anticipated, for example, in the mistranslation of idioms and colloquialisms (Daems et al., 2017).

5.17. Stakeholders need to understand the design rationale

People need to know the rationale for the answer and why the answer makes sense (P6, 14). This is expressed in terms of the domain (concepts, principles, causes, etc.) and not just in terms of how AI works. Role-holders need to be able to explain the design rationale to vendors. Developers must understand the legacy work and how the end-users do what they do using their legacy system. The designer needs to get end-user reactions to the AI/ML system. End-users can help developers re-create the conditions that led to confusion and problems.

5.18. Operators are often left adrift

Our participants expressed the sentiment that even if a system is thought to be useful in general, the need to integrate a system into their organization's processes and mission is ignored and the integration burden falls on the operators.

> *The search for tools that really help with the job takes priority. Only after that is there a matter of exploration and self-explanation. Part of it is trying to find out how they work, part of it is trying to find out how they apply* (P6).
>
> *People would come to me with tools, even people within our own organization and say "Hey, this is really awesome." I would say "Yeah, but how much work am I really going to put into that to really get this tool working to answer your question?"* (P6).
>
> *I have to jump to see what if any of the tools might apply to any or all our organization's problem sets* (P6).

6. Concurrence with other findings

Our findings about explanation desirements are generally consistent with those expressed by Langer et al. (2021). They reviewed over 100 journal articles and conference presentations to roster researchers' speculations about desiderata. Their list includes qualities that emerged in our interviews but by design were tacit in our interview questions, such as acceptance, effectiveness, satisfaction, and usability. Their list also includes a few requirements that were not referenced in our interviews, such as "debugability" and "enable informed consent."

Our findings are in accord with other empirical research also conducted at about the same time. Factors similar to those we found also emerged in a study by Liao et al. (2020), who interviewed 20 user experience professionals asking questions about how AI systems work. They also go a step further in exploring the dependence on explanation requirements for different sorts of AI applications.

In a study by Hepenstal and McNeish (2020), a focus group of 12 professionals (representing both developers and users) worked on scenarios involving different kinds of AI systems. Their task was to think about their explanation needs. The factors that emerged from the group discussions included understanding the data, achieving appropriate trust, understanding the AI's limitations, methods for verification and validation of the AI, and the issue of responsibility. Furthermore, some of the dimensions were emphasized more by, say, developers rather than by users. These findings are consistent with ours.

7. Putting the playbook to work

The explanation desirements and the cautions are not solutions to the explanation challenges that face the system developer. The primary purpose of the Playbook is to provide guidance for creating explanations that support the different stakeholder groups and roles, enabling system developers to consider the different ways in which stakeholders or role-holders need to "look inside" the AI/XAI system. XAI system development efforts rarely consider the need to provide explanations that express the design rationale, explanations of the source and quality of the data on which the AI was trained, explanations of how the AI was tested for usability, and so forth. Such cautions should be considered at the point in system development when the explanation capabilities of the AI system are being envisioned.

Using the Playbook, developers can also anticipate the access requirements of adopters. For example, they can choose to include capabilities to enable adopters to contact experienced users and they can choose to include capabilities for adopters to get guidance when something goes wrong.

The Playbook offers a considerable number of desirements and cautions. Not all of them need to be considered for every AI application. The Playbook presents suggestions for how developers might focus on what might be important requirements for their particular application and intended adopters.

Furthermore, the access requirements are not about the design of explanations *per se*, but about things that developers need to consider during their development process. For example, our program manager participants expressed a need to access "trusted vendors." Developers expressed a need to access individuals who are experienced at the work for which the AI is designed. Development team leaders emphasized the need to consider how the AI system will be integrated with other systems when it is embedded in the full work context and so forth.

The Playbook also has applications with regard to management and policy for organizations that develop XAI systems. The Playbook reaches the management and policy levels since some of the stakeholder roles fall in those arenas, especially the desirements and access requirements of system developers and program managers. But at an even higher level, organizational leadership and culture can impose constraints on project development, and can even be an obstacle to successful adoption. Is management willing to listen to what role-holders and beneficiaries are saying? Will the system development process embrace an engineering requirement that the system must be explorable by end-users? Management can be invited to consider the Playbook as a whole; indeed there should be intrinsic motivation to hear what developers desire and what end-users actually need. At the detailed level, project and higher level managers can consider whether particular desirements and access requirements fall within the intended scope of their XAI development project. Will there be adequate support for developers to involve domain practitioners and intended beneficiaries in the system development process? Will management be supportive of attempts to reveal and specify the limitations and biases of the XAI system? Will there be sufficient support for getting adopters, end-users, and trainers access to system engineers and developers? Or will end-users and adopters simply be set adrift? Will there be support for follow-on activity to use end-user feedback in refining and improving the XAI system?

Most of the access requirements in the Playbook can be directly implemented.

- There can be a policy that the systems development process must engage experienced domain practitioners. This has been accomplished by having end-users comment on the explanation capabilities of prototype systems (see, for example, Jacobs et al., 2021).
- There can be a policy that the system deliverable must include links that enable end-users to submit queries to system developers or software engineers. Working in the other direction, targeted end-users or end-user groups could be asked to identify "trusted developers" whom they would like to be able to consult. The access capability needs to be efficient, to enable end-users to rapidly get information when they experience something anomalous.
- There can be a policy that the deliverable must include in its instructional material a precis explaining the workings of the AI system and its particular architecture. XAI systems should certainly include such instructional material but often do not (see Mueller and Klein, 2011).

Some of the explanation desirements in the Playbook can be satisfied, at least in part, by a "Cognitive Tutorial" (Mueller and Klein, 2011). Although initially intended as a cognitive tutorial for end-users, the concept applies to the provision of explanatory information to stakeholders. The tutorial consists of instructions and exercises that:

- Describe the data used in training the AI (which would be a corpus of data in the case of ML systems). This speaks to the desirement for an explanation of the data and inputs.
- Describe the system representation, modeling mechanisms, and algorithms. This speaks to the desirement for a global explanation.
- Present incident-based interviews with system developers and trainers, illustrating what happens when the ML system is "stretched." This speaks to the desirement for examples or demonstrations that show the results based on different input conditions.
- Present incident-based interviews with system adopters, discussing issues in usability and performance. This can speak to the desirement for guidance in anticipating the trade-offs that arise in operational contexts.
- Present end-user notes about their experiences with troubleshooting. This speaks to the desirement for guidance in knowing what to do when something goes wrong.

Exercises in the tutorial include scenarios accompanied by a forced choice between alternative solutions. Such exercises can show positive and negative usage quickly. Exercises can present scenarios that contain errors or misinterpretations and then allow the learner to discover the problems. Exercises can present a scenario and have the learner construct a plan or model of the scenario, which is then compared to the plan or model created by an expert. Exercises intended to promote global understanding can take a "back door" approach in which learners are asked a question about system operation and then have to develop and demonstrate their understanding via the use of the XAI system. Such exercises can enable the learner to discover for themselves how the system accomplishes a task.

Individual stakeholders can choose among the tutorial modules to satisfy their particular desirements. Mueller and Klein (2011) illustrate a cognitive tutorial created for explaining how to use the JAVA Causal Analysis Toolkit.

8. Discussion

8.1. Reconciling the paradoxes

Do stakeholders and role-holders want explanations? Do they want more explanations? Do they want better explanations? The answer is Yes or No. Some individuals want explanations, and some do not. Some want detailed technical explanations, and some do not. Some expressed a need to explore, and some did not. Some stakeholders say they only need a global understanding, but some say they do need to look under the hood.

How can such contradictory findings be reconciled? A given individual may not need an explanation, either global or local, depending on their style and circumstance (e.g., they can rely on trusted developers). But individuals in all roles do want and need a satisfying understanding of something, either the AI or the data that were fed to it, at least some of the time. Their expression of a need for explanation (or self-explanation) is often subtle and indirect. But not everyone needs explanations, and explanations are not needed all the time. Only half of our participants expressed an interest in receiving explanations. The other participants were either indifferent or explicitly disinterested in receiving explanations. These results paint a much more complex and subtle picture than appears in much of the previous theoretical and taxonomic research, or the assumptions made about the XAI notion.

The contradictions can be resolved by acknowledging that different individuals have different capabilities, different sensemaking requirements, different immediate goals, and typically serve in more than one role. Different cognitive styles are also a factor, as suggested by participant comments to the effect that they preferred to dive in and play with the system, rather than getting an explanation of how it works. These factors combine to define what, for each individual, constitutes satisfactory and actionable understanding.

8.2. Rethinking the stakeholder concept

Since most of our participants served or had served in multiple roles, assigning participants to role categories required multiple tabulations, depending on the participant's perspective when making particular comments. For example, four of the 18 Participants self-described as end-users, although their current primary role was that of a developer. Thus, occasional comments by a participant might be from their perspective as a former or sometimes end-user when their primary current role was, say, that of a developer. Our participants frequently referenced the explanation requirements of other stakeholders. For example, end-users sometimes play a key role in procurement decision-making. As another example, iurisprudence specialists are more like end-users than developers or system integrators, because (as one Participant put it) they are "OK with some of the stuff being a black box". But they need to have appropriate trust and appropriate mistrust and be able to determine when they have entered a gray area.

Roles are not equivalent to professional identity. Roles are fluid, and sometimes people serve in several roles simultaneously. As this research topic moves forward, it may be more useful to talk about roles than to pretend that people consistently inhabit distinct stakeholder groups. Based on their own research, both Hepenstal and McNeish (2020) and Liao et al. (2020), also moved from a consideration of stakeholder categories to a consideration in terms of roles.

Regarding *roles* as the critical determiners of an individual's explanation requirements has implications for the mapping of explanations types onto XAI methods for generating explanations. The Supplement to this article presents "clusters" of participants' comments, clusters that cut across traditional stakeholder categories. Participants' comments are categorized as expressing sensemaking requirements, sensemaking challenges, issues in explaining to others, trust and reliance issues, and challenges for design and procurement. Within each of those categories are clusters that cut across roles. For example, jurisprudence professionals and contracting professionals may have similar needs in terms of the explanations they require. Those same two groups of professionals may share the challenge engendered by their lack of computer science knowledge. As another example, end-users and policy analysts may both need to understand edge cases.

8.3. Human-centered approach for integration and evaluation

In the design and development of AI systems, including XAI systems, developers must have a deep and thorough understanding of the cognitive work (Wachter et al., 2016). One of our participants commented: *The integrator needs to lead people who go into the field and see how users use network the technology*. Yet, vendors are not always required to deeply study the legacy work and work context prior to designing the tools. How will the insertion of technology change cognitive work? What new forms of error might it trigger?

It is generally assumed that the main goal of XAI is to promote better performance by the user at the particular task that the AI is designed to accomplish or help the human accomplish. But explanations need to support the process of human–machine integration as a contextualized work system in which the human and the machine are interdependent. The findings presented here call out the importance of a Human–Computer Interdependence approach for the design and evaluation of AI systems, including XAI systems.

> *AI/ML is early in the Tech maturation pipeline so, it is not in the systems that are being fielded and tested. But it's gonna be a big problem, how you assess this technology* (P2).
>
> *The challenge is how do you get AI/ML into the requirements documents; how do you design AI technology well for the user and get it into the requirement document* (P2).
>
> *We have limited time to evaluate all the time. We have novices who have never used the legacy system, but [on the new system] outperform the experts on the legacy system. They do not have the negative transfer and all that baggage. Experts on legacy systems are often slower on the new system because they are trying to figure things out, or it is not where they expected* (P14).

This is not just a technology assessment, it is an issue of human–technology–work integration. It is necessary to find the gaps in the user's needs and understand the methods to address those via human–system interdependence analysis. The system developer and integrator need to understand AI and ML systems from the standpoint of the cognitive work the technology is to help with. The end-user, especially the subject-matter expert, needs to be able to do what the systems integrator does. The integrator has to explain to the user experts what the human factors person is doing. The explanation is about the process and rationale for the human–machine integration activities. The user expert needs to be able to do what the human–system integrator does.

Our interview participants called for a procurement requirement to conduct systems integration. They called for training in the acquisition community about usability and assessment methodologies.

> *For me, the overarching questions we should always be asking are: How do I use this? How will it help the task that has to be accomplished? How will it help the human accomplish that task? When we put AI in just because we can, we miss the overall objective. The objective is the task goal, why is it hard for humans to do, how will the AI make it better, and then based on that how will the human use the AI easier for the human do to. It may evolve the task or the task may become slightly different. But that human-centered or task-oriented design is at the center of designing any AI application is critical. This is often missed* (P7).
>
> *If it's really AI, it is supposed to be operating more in the complex cognitive realm. And then my skepticism kicks in, that the AI can really keep up and be useful to people. I would want to know from there, how much opportunity is there to allow the users finish the design* (P9).
>
> *I need to know all the things that the operators need to know. Oh, yes. Absolutely. I spend a lot of time understanding what they do, and I don't have their training. I get SMEs to work with me. And I mentor them to do what I am doing, So that they can do my job like I do my job, but they help me all along the way to understand everything and the "why" behind it* (P14).

Current government procurement requirements were not designed with AI systems in mind, and they seem inadequate for the task. How can AI requirements (including explanation desirements) get into the procurement process? How will AI/ML require a change to the current "human readiness levels" scheme? How can we create procurement standards or "best practice" guidance for evaluating AI systems? These questions have become a focus point in discussions of AI procurement issues (Tate et al., 2016; Russell et al., 2021).

9. Limitations of the present work

Our interviewing was constrained in terms of the number of questions that could be asked and answered in about an hour. An attempt was made to gain coverage of the key issues of explanation and understanding. Alternative questions could certainly have been utilized.

The method used in the identification of the themes and role clusters depended, of course, on the researcher's subjective judgments. Might other individuals find other themes or role clusters? Undoubtedly. But we would expect considerable concurrence by individuals who are versed in ethnographic interviewing and qualitative data analysis. Furthermore, the finding of alternative themes or role clusters would not negate the informativeness of the themes and clusters that were identified. A categorized listing of interviewee statements that illustrate all of the themes and clusters is available from the authors upon request. This extended set of interviewee statements makes manifest the content validity of the identified themes and role clusters.

Our findings are based on a limited sample, of course. While this practical constraint may be regarded as a shortcoming, there are mitigating factors that set the present work apart: (1) All of the participants were experienced, senior, or mid-career professionals; (2) All had advanced degrees and experience with a variety of AI systems; and (3) All were engaged in an in-depth discussion, as opposed to simply providing ratings, questionnaire responses, or assertions in an open forum. Of course, an extension of this sort of data set will be valuable. The involvement of more interviewees representing all of the roles would be very important. Thus, we regard the Stakeholder Playbook as a first pass. We look forward to further empirical efforts to collect more evidence, elaborate on the desirements, requirements, and challenges, and include other roles.

The Playbook conflates some desirements for general AI system development with desirements for augmenting AI models with explanations. For example, trust is an issue that pertains to all AI systems. Trainers need to be able to train end-users on how any AI system fails and how it misleads. Partnerships between end-users and developers are a must for the successful development of any AI system, regardless of the level of explainability it provides. However, AI systems that do not involve the automatic generation of explanations nevertheless come with instructions and training materials. Thus, explanation desirements pertain.

The Playbook is not intended to solve the challenges in the design and conduct of experiments that are aimed at evaluating AI or XAI systems. Amarasinghe et al. (2022) have detailed some of the issues in the design and methodology of evaluation experiments, including the failure to engage the participation of experienced users or domain experts, and the failure to engage participants in tasks with real-world relevance. Many evaluation studies rely on single measures (of such things as performance and trust) and many studies rely on the use of Likert scales rather than in-depth cognitive interviews. Many studies rely on unnecessarily large samples of Mechanical Turkers, for the sake of achieving statistically significant effects rather than to seek practically significant effects. Recent attempts at evaluation have included some methodological elements that might lend rigor to the effort (see, for instance, Buçinca et al., 2020; Dodge et al., 2021; Klein et al., 2021; Rosenfeld, 2021; Hoffman et al., 2023).

10. Prospects

Our results reveal issues with the general XAI tactic for providing explanations. Individuals prefer to *engage* in explanation, rather than being passive recipients of explanatory materials. In this spirit, the Stakeholder Playbook is intended to expand the horizons of XAI systems. It must be possible to create better XAI systems by cutting loose from the assumption that explanations are simply provided and then directly empowering the stakeholder's sensemaking.

Data availability statement

The data presented in this study can be provided by the authors upon request. A supplement to this article presents examples of quotations expressing the themes, categorized in terms of the role clusters.

Ethics statement

The studies involving humans were approved by Institutional Review Board, Florida Institute for Human and Machine Cognition. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their informed consent to participate in this study. No potentially identifiable images or data are presented in this study.

Author contributions

RH was lead author and conducted the interviews. MJ and CT assisted with the analysis. GK and SM had the initial concept and method. All authors contributed to the article and approved the submitted version.

Funding

This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA) under agreement number FA8650-17-2-7711.

Acknowledgments

The authors would like to acknowledge William J. Clancey of the Florida Institute for Human and Machine Cognition for his contributions to the conceptual foundations of XAI. The authors thank Timothy Cullen (Col., US Army, Ret.) for his consultation on this project. The authors thank the three reviewers for their helpful suggestions on the submission. This material is approved for public release. Distribution is unlimited. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon.

Conflict of interest

GK was employed by MacroCognition, LLC.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Author disclaimer

The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of AFRL or the U.S. Government.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomp.2023.1117848/full#supplementary-material

References

Al-Abdulakarim, L., Atkinson, K., Bench-Capon, T., Whittle, S., Williams, R., and Wolfenden, C. (2019). Noise induced hearing loss: Building an application using the ANGELIC methodology. *Argu. Comput.* 10, 5–22. doi: 10.3233/AAC-181005

Al-Abdulkarim, L., Atkinson, K., and Bench-Capon, T. (2016). A methodology for designing systems to reason with legal cases using abstract dialectical frameworks. *Artif. Intell. Law* 24, 1–49. doi: 10.1007/s10506-016-9178-1

Amarasinghe, K., Rodolfa, K. T., Jesus, S., Chen, V., Balayan, V., Saleiro, P., et al. (2022). On the importance of application-grounded experimental design for evaluating explainable ml methods. *arXiv*:2206.13503.

Arioua, A., Buche, P., and Croitoru, M. (2017). Explanatory dialogs with argumentative faculties over inconsistent knowledge bases. *J. Expert Syst. Applic.* 80, 9. doi: 10.1016/j.eswa.2017.03.009

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* 58, 82–115. doi: 10.1016/j.inffus.2019.12.012

Arya, V., Bellamy, R. K., Chen, P. Y., Dhurandhar, A., Hind, M., Hoffman, S. C., et al. (2019). One explanation does not fit all: A toolkit and Taxonomy of AI explainability concepts. arXiv:1909.03012.v2.

Atkinson, K., Bench-Capon, T., and Bollegala, D. (2020). Explanation in AI and law: Past, present and future. *Artif. Intell.* 22, 103387. doi: 10.1016/j.artint.2020.103387

Bhatt, U., Andrus, M., Weller, A., and Xiang, A. (2020). Machine learning explainability for external stakeholders. arXiv:2007.05408v1.

Buçinca, Z., Lin, P., Gajos, K. Z., and Glassman, E. L. (2020). "Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems," in *Proceedings of the ACM International Conference on Intelligent User Interfaces* (New York: Association for Computing Machinery) 454–464. doi: 10.1145/3377325.3377498

Cabitza, F., Campagner, A., Malgieri, G., Natali, C., Scheenberger, D., Stoeger, K., and Holzinger, A., (2023). Quod erat demonstrandum? - Toward a typology of the concept of explanation for the design of explainable AI. *Expert Syst. Applic.* 313, 118888 doi: 10.1016/j.eswa.2022.118888

Calegari, R., Ciatto, G., Dellaluce, J., and Omicini, A. (2019). "Interpretable narrative explanation for ML predictors with LP: A case study for XAI," in *Workshop 'From Objects to Agents' (WOA 2019)*. Available online at: https://www.semanticscholar.org/paper/Interpretable-Narrative-Explanation-for-ML-with-LP%3A-Calegari-Ciatto/1e345972e7625c771554c15d362b98fd2e86d8f4 (accessed 29 March 2023).

Chari, S., Seneviratne, O., Gruen, D. M., Morgan, A., Das, A. K., and McGuiness, D. L. (2020). "Explanation ontology: A model of explanations for user-centered AI," in *International Semantic Web Conference* (Cham: Springer International Publishing) 228–243. doi: 10.1007/978-3-030-62466-8_15

Chi, M. T. H., Roy, M., and Hausmann, R. G. M. (2008). Observing tutorial dialogues collaboratively: Insights about human tutoring effectiveness from vicarious learning. *Cogn. Sci.* 32, 301–341. doi: 10.1080/03640210701863396

Crandall, B., Klein, G., and Hoffman, R. R. (2006). *Working Minds: A Practitioner's Guide to Cognitive Task Analysis*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/7304.001.0001

Daems, J., Vandepitte, S., Hartsuiker, R. J., and Macken, L. (2017). Identifying the machine translation error types with the greatest impact on post-editing effort. *Front. Psychol.* 8, 1282. doi: 10.3389/fpsyg.2017.01282

Dahan, S. (2020). *AU-powered trademark dispute resolution*. Report to the European Union Intellectual Property Office (EUIPO). Available online at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3786069 (accessed July 31, 2023).

Dodge, J., Anderson, A., Khanna, R., Irvine, J., Dikkala, R., Lam, H. K.-H., et al. (2021). From "no clear winner" to an effective explainable Artificial Intelligence process: An empirical journey. *Appl. AI Lett.* 2, e36. doi: 10.1002/ail2.36

Doshi-Velez, F., and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv:1702.08608v2.

Eiband, M., Schneider, H., Bilandzic, M., Fazekas-Con, J., Haug, M., and Hussmann, H. (2018). "Bringing transparency design into practice," in *23rd International Conference on Intelligent User Interfaces* 211–223. doi: 10.1145/3172944.3172961

European Union Commission (2016). "*General Data Protection Regulation Article 22, Recital 71.*" Available online at: https://www.privacy-regulation.eu/en/recital-71-GDPR.htm (accessed July 31, 2023).

Fazelpour, S. (2023). *Disciplining deliberation: Interpreting machine learning trade-offs in sociotechnical systems. AI Metrology Colloquia Series.* National Institute of Standards and Technology. Available online at: https://www.nist.gov/node/1698041/ai-metrology-colloquia-series (accessed July 3, 2023).

Felzmann, H., Villaronga, E. F., Lutz, C., and Tamo-Larrieux, A. (2019). Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data Soc.* 6, 2053951719860542. doi: 10.1177/2053951719860542

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., et al. (2018). AI4people—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds Mach.* 28, 689–707. doi: 10.1007/s11023-018-9482-5

Glaser, B. G. (1998). *Doing Grounded Theory - Issues and Discussions*. Mill Valley, CA: Sociology Press.

Goodman, B., and Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation." *AI Mag.* 38, 50–57. doi: 10.1609/aimag.v38i3.2741

Gunning, D., Vorm, E., Yunyan, J.W., and Turek, M. (2021). DARPA's explainable AI Program: A retrospective. *Appl. AI Lett.* 19, 1727. doi: 10.22541/au.163699841.19031727/v1

Hepenstal, S., and McNeish, D. (2020). "Explainable artificial intelligence: What do you need to know?" in *Augmented Cognition. Theoretical and Technological Approaches*, eds. D. Schmorrow and C.M. Fidopiastis (Cham: Springer). doi: 10.1007/978-3-030-50353-6_20

Hind, M., Wei, D., Campbell, M., Codella, N. C., Dhurandhar, A., Mojsilović, A., et al. (2019). "TED: Teaching AI to explain its decisions," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (New York: Association for Computing Machinery) 123–129. doi: 10.1145/3306618.3314273

Hoffman, R. R., and Elm, W. C. (2006). "HCC implications for the procurement process. *IEEE Intell. Syst.* 21, 74–81. doi: 10.1109/MIS.2006.9

Hoffman, R. R., Klein, G., and Miller, J. E. (2011). Naturalistic investigations and models of reasoning about complex indeterminate causation. *Inf. Knowl. Syst. Manag.* 10, 397–425. doi: 10.3233/IKS-2012-0203

Hoffman, R. R., and McCloskey, M. J. (2013). Envisioning desirements. *IEEE Intell. Syst.* 26, 82–89. doi: 10.1109/MIS.2013.108

Hoffman, R. R., Mueller, A. S. T., Klein, G., and Litman, J. (2023). Measures for explainable AI: Explanation goodness, User satisfaction, mental models, curiosity, trust and human-AI Performance. *Front. Comput. Sci.* 5, 1096257. doi: 10.3389/fcomp.2023.1096257

Hoffman, R. R., Mueller, S. T., and Klein, G. (2017). Explaining Explanation, Part 2: Empirical Foundations. *IEEE Intell. Syst.* 34, 78–86. doi: 10.1109/MIS.2017.3121544

Hutchins, E. (2003). "Cognitive ethnography," in *Proceedings of the Annual Meeting of the Cognitive Science Society* 25.

IBM (2021). *IBM Research Trusted AI*. Available online at: http://aix360.mybluemix.net/consumer (accessed January 30, 2023).

Jacobs, M., He, J., F., Pradier, M. F., Lam, B., Ahn, A. C., et al. (2021). "Designing AI for trust and collaboration in time-constrained medical decisions: a sociotechnical lens," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (New York: Association for Computing Machinery) 1–14. doi: 10.1145/3411764.3445385

Johs, A. J., Agosto, D. E., and Weber, R. O. (2020). Qualitative investigation in explainable Artificial Intelligence: A bit more insight from social science. arXiv:2011.07130 doi: 10.22541/au.163284810.09140868/v1

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., and Wortman Vaughan, J. (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* (New York: Association for Computing Machinery) 1–14. doi: 10.1145/3313831.3376219

Kenny, E., Ford, C., Quinn, M., and Keane, M. (2021). Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies. *Artif. Intell.* 294, 103459. doi: 10.1016/j.artint.2021.103459

Klein, G., Hoffman, R. R., and Mueller, S. T. (2021). Modeling the process by which people try to explain things to others. *J. Cogn. Eng. Deci. Mak.* 15, 213–232. doi: 10.1177/15553434211045154

Lage, I., Chen, E., He, J., Narayanan, M., Kim, B., Gershman, S., et al. (2019). An evaluation of the human-interpretability of explanation. arXiv:1902.00006

Langer, M., Oster, D., Speith, T., Hermanns, H., Kastner, L., Schmidt, E., et al. (2021). What do we want from explainable artificial intelligence (XAI): A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary research. *Artif. Intell.* 296, 103473. doi: 10.1016/j.artint.2021.103473

Liao, Q. V., Gruen, D., and Miller, S. (2020). "Questioning the AI informing design practices for explainable AI user experiences," in *Proceedings of CHI 2020* (New York: Association for Computing Machinery). doi: 10.1145/3313831.3376590

Lipton, Z. C. (2018). The mythos of model interpretability. *Queue* 16, 31–57. doi: 10.1145/3236386.3241340

Lombrozo, T. (2016). Explanatory preferences shape learning and inference. *Trends Cogn. Sci.* 20, 748–759. doi: 10.1016/j.tics.2016.08.001

Loyola-Gonzalez, O. (2019). Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE Access* 7, 154096–154113. doi: 10.1109/ACCESS.2019.2949286

Miller, T. (2017). Explanation in Artificial Intelligence: Insights from the social sciences. *Artif. Intell.* 267, 1–38. doi: 10.1016/j.artint.2018.07.007

Mittelstadt, B. D., Russell, C., and Wachter, S. (2019). "Explaining explanations in AI," in *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency* (New York: Association for Computing Machinery) 279–288. doi: 10.1145/3287560.3287574

Mohseni, S., Zarel, N., and Ragan, E. D. (2020). A multidisciplinary survey and framework for design and evaluation of explainable AI Systems. *ACM Trans. Inter. Intell. Syst.* 11, 1–45. doi: 10.1145/3387166

Mueller, S. T., Hoffman, R. R., Clancey, W., Emrey, A., and Klein, G. (2019). *Explanation in human-AI systems: a literature meta-review, synopsis of key ideas and publications and bibliography for explainable AI.* Technical Report from Task Area 2 to the DARPA Explainable AI Program. Available online at: https://apps.dtic.mil/sti/citations/AD1073994 (accessed January 29, 2023).

Mueller, S. T., and Klein, G. (2011). Improving users' mental models of intelligent software tools. *IEEE Intell. Syst.* 26, 77–83. doi: 10.1109/MIS.2011.32

Naiseh, M., Jiang, N., Ma, J., and Ali, R. (2020). "Personalizing explainable recommendations: literature and conceptualization," in *Trends and Innovations in Information Systems and Technologies,* eds. Á., Rocha, H., Adeli, L. P., Reis, S., Costanzo, I., Orovic, and F., Moreira (Cham: Switzerland: Springer International Publishing) 518–533. doi: 10.1007/978-3-030-45691-7_49

Nguyen, D. (2018). "Comparing automatic and human evaluation of local explanations for text classification," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 1069–1078. Available online at: https://www.semanticscholar.org/paper/Comparing-Automatic-and-Human-Evaluation-of-Local-Nguyen/3785f9083dcb46d2bea7ee771c2a513bb43917a6 (accessed March 19, 2023).

Preece, A., Harborne, D., Braines, D., Tomsett, R., and Chakraborty, S. (2018). Stakeholders in explainable AI. arXiv preprint arXiv:1810.00184.

Ribera, M., and Lapedriza, A. (2019). "Can we do better explanations? A proposal of user-centered AI," in *Proceedings of the ACM IUI 2019 Workshop* (New York: Association for Computing Machinery).

Rosenfeld, A. (2021). "Better metrics for evaluating explainable artificial intelligence," in *AAMAS '21: Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems* (New York: Association for Computing Machinery). Available online at: https://www.ifaamas.org/Proceedings/aamas2021/pdfs/p45.pdf$\delimiter"026E30F$ (accessed July 31, 2023).

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* 1, 206–215. doi: 10.1038/s42256-019-0048-x

Russell, S., Jalaian, B., and Moskowitz, A. S. (2021). "Re-orienting towards the science of the artificial: Engineering AI systems," in *Systems Engineering and Artificial Intelligence* 149–174. doi: 10.1007/978-3-030-77283-3_8

Schoepfle, M. (2021). *Introduction to Cognitive Ethnography and Systematic Field Work*. Thousand Oaks, CA: Sage Publications.

Sheh, R., and Monteath, I. (2018). Defining explainable AI for requirements analysis. *KI - Künstliche Intell.* 32, 261–266. doi: 10.1007/s13218-018-0559-3

Shneiderman, B. (2023). "Human-centered ai: ensuring human control while increasing automation." in *Proceedings of the 5th Workshop on Human Factors in Hypertext* 1–2. doi: 10.1145/3538882.3542790

Sokol, K., and Flach, P. (2020). "A framework for systematic assessment of explainable approaches," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (New York: Association for Computing Machinery) 56–67.

Strout, J., Zhang, Y., and Mooney, R. J. (2019). "Do human rationales improve machine explanations?" in *Proceedings of the Second BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP* (Stroudsburg, PA: Association for Computational Linguistics) 56–62. doi: 10.18653/v1/W19-4807

Tate, D. M., Grier, R. A., Martin, C. A., Moses, F. L., and Sparrow, D. A. (2016). *A Framework For Evidence-Based Licensure Of Adaptive Autonomous Systems*. Alexandria, VA: Institute for Defense Analysis. Available online at: https://www.ida.org/-/media/feature/publications/a/af/a-framework-for-evidence-based-licensure-of-adaptive-autonomous-systems/p-5325.ashx (accessed July 31, 2023).

Tjoa, E., and Guan, C. (2020). "A survey on explainable Artificial Intelligence (XAI): Toward medical XAI," in *IEEE Transactions on Neural Networks and Learning Systems.* Available online at: https://www.researchgate.net/publication/346017792_A_Survey_on_Explainable_Artificial_Intelligence_XAI_Toward_Medical_XAI (accessed July 31, 2023).

Tomsett, R., Braines, D., Harborne, D., Preece, A., and Chakraborty, S. (2018). "Interpretable to whom? A role-based model for analyzing interpretable machine learning systems," in *Proceedings of the 2018 ICML Workshop on Human Interpretability in Machine Learning (WHI 2018)* (Stockholm, Sweden).

Vermeire, T., Laugel, T., Renard, X., Martens, D., and Detyniecki, M. (2021). "How to choose an explainability method? Towards a methodical implementation of XAI in practice," in *Machine Learning and Principles and Practice of Knowledge Discovery in Databases. (ECML PKDD 2021)*, eds. M. Camp (Cham, Switzerland: Springer). doi: 10.1007/978-3-030-93736-2_39

Wachter, S., Mittelstadt, B., and Floridi, L. (2016). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *Int. Data Priv. Law* 72, 76–99. doi: 10.2139/ssrn.2903469

Weller, A. (2019). "Transparency: Motivations and challenges," in *Explainable AI: Interpreting, explaining and visualizing deep learning,* eds. W., Samek, G., Montavon, A., Vedalti, L.K., Hansen, and K.-R., Muller (Cham, Switzerland: Springer Nature) 23–40. doi: 10.1007/978-3-030-28954-6_2

Zaidan, O., Eisner, J., and Piatko, C. (2007). "Using annotator rationales to improve machine learning for text categorization," in *Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference* 260–267. Available online at: https://aclanthology.org/N07--1033.pdf (accessed July 31, 2023).

Zhang, Y., Marshall, I., and Wallace, B. C. (2016). *Rationale-Augmented Convolutional Networks for Text Classification.* Available online at: https://arxiv.org/pdf/1605.04469.pdf