Proceedings of the Tenth AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2022)

Connecting Algorithmic Research and Usage Contexts: A Perspective of Contextualized Evaluation for Explainable AI

Q. Vera Liao [1]*, Yunfeng Zhang [2], Ronny Luss[3], Finale Doshi-Velez[4], Amit Dhurandhar[3]

[1]Microsoft Research, [2]Twitter Inc., [3]IBM Research, [4]Harvard University

Abstract

Recent years have seen a surge of interest in the field of explainable AI (XAI), with a plethora of algorithms proposed in the literature. However, a lack of consensus on how to evaluate XAI hinders the advancement of the field. We highlight that XAI is not a monolithic set of technologies—researchers and practitioners have begun to leverage XAI algorithms to build *XAI systems* that serve different usage contexts, such as model debugging and decision-support. Algorithmic research of XAI, however, often does not account for these diverse downstream usage contexts, resulting in limited effectiveness or even unintended consequences for actual users, as well as difficulties for practitioners to make technical choices. We argue that one way to close the gap is to develop evaluation methods that account for different user requirements in these usage contexts. Towards this goal, we introduce a perspective of contextualized XAI evaluation by considering the relative importance of XAI evaluation criteria for prototypical usage contexts of XAI. To explore the context dependency of XAI evaluation criteria, we conduct two survey studies, one with XAI topical experts and another with crowd workers. Our results urge for responsible AI research with usage-informed evaluation practices, and provide a nuanced understanding of user requirements for XAI in different usage contexts.

Introduction

The wide adoption of AI technologies in high-stakes domains, coupled with the proliferation of inscrutable "black-box" AI models, has spurred great interest in explainable AI (XAI) in academia and industry. Each year, hundreds of papers proposing various XAI algorithms are published. Unfortunately, a lack of consensus on what constitutes good explanations hinders the advancement of the field and real-world adoption of XAI. While practitioners recognize the value of explainability, they grapple with tremendous challenges in making appropriate choices of XAI techniques and creating effective XAI systems (Bhatt et al. 2020; Liao, Gruen, and Miller 2020; Hong, Hullman, and Bertini 2020). Researchers, especially those in the human-computer interaction (HCI) community, have begun to explore diverse XAI systems (e.g. (Kaur et al. 2020; Xie et al. 2020)). These studies pointed out that explainability is not a monolithic concept, and what users need to be explained varies across different types of systems and user tasks such as debugging a model, judging the reliability of model outputs, assessing regulatory compliance, or learning about a domain.

This recognition of the context-dependency of explanation "goodness" resonates with social science literature studying human explanations (Mueller et al. 2019; Vasilyeva, Wilkenfeld, and Lombrozo 2015). An explanation is often conceptualized as an attempt for the explainer to fill the gaps in the understanding of the explainee. Therefore its goodness should be relative to this understanding gap, which is determined by both the explainee's current understanding and the necessary understanding to achieve their given objective. Social science literature distinguishes different objectives for people to seek explanations, including predicting future events, diagnosis, assigning blame, resolving cognitive dissonance, rationalizing actions, and aesthetic pleasure (Lombrozo 2012; Keil 2006; Lombrozo 2006).

However, this context-dependent nature of explainability is not well-acknowledged in current XAI research. There is a fundamental disconnect between algorithmic research and downstream usage contexts. Algorithmic research is often not motivated by well-defined needs of intended users (Miller, Howe, and Sonenberg 2017). In fact, the intended use is often not made explicit, despite growing efforts in encouraging AI researchers to articulate the downstream impact of their research. This disconnect has been recognized to cause pitfalls of XAI methods (Liao and Varshney 2021; Ehsan and Riedl 2021)—unintended harmful consequences for users such as lacking actionability, cognitive burden and over-reliance on AI.

This disconnect is reflected in the dominant practices of how XAI algorithms are evaluated, which can profoundly shape the field. A major camp of XAI evaluation focuses solely on algorithmic criteria such as faithfulness and stability (Alvarez-Melis and Jaakkola 2018; Carvalho, Pereira, and Cardoso 2019), which are inadequate to capture the satisfaction of "users in context" (Hoffman et al. 2018). To move towards a rigorous science based on empirical evidence, Doshi-Velez and Kim (2017)'s foundational work called for application-grounded evaluation—with real humans and by the success of target tasks. When the resource is

*Part of the work was completed while the first and second authors were working at IBM Research

limited, compromises can be made with simplified or proxy tasks. However, how to design simplified or proxy evaluation tasks that capture the essential requirements for different application contexts remains an open question. Despite increasing effort to conduct human-subject studies to evaluate XAI algorithms, popular evaluation methods are often devoid of usage contexts and under-specify the criteria. Some asked participants to judge which explanation is "better" (Jeyakumar et al. 2020), without specifying an end-goal to ground the definition of "better". Another approach is to use a simplified task of "simulatability test" by asking participants to simulate the model output based on the explanation (Hase and Bansal 2020; Lucic, Haned, and de Rijke 2020). A recent study by Buçinca et al. (2020) pointed out that such tasks have limited evaluative power to predict the success in real tasks, as they do not capture what users need to be explained or how they process the explanations.

Our work is motivated to close these gaps. We propose to contextualize the evaluation of XAI by considering the relative importance of evaluation criteria for prototypical usage contexts of XAI systems. We use the term "contexts" to refer to situations where explanations are sought for distinct user objectives, which can appear in different applications. In Section 3, we first contribute taxonomies of XAI evaluation criteria and prototypical usage contexts by synthesizing related literature. In Section 4, we empirically explore the context-specific ranking of these evaluation criteria by surveying XAI experts (from both HCI and AI communities) and crowd workers as target users of an AI application. As a first step towards contextualized evaluation of XAI, we use a retrospective approach by surveying people's opinions to systematically explore this large problem space. We believe our results can contribute to four grounds:

- Our work aligns with responsible AI research—by explicating the context-dependency of evaluation criteria, we encourage XAI researchers to articulate the appropriate usage contexts for the algorithms they develop.
- By capturing the desired properties of XAI, our results also inform how to develop XAI techniques and systems that serve different usage contexts.
- Methodologically, we contribute a scenario-based survey approach to elicit crowd workers' opinions on the relative importance of evaluation criteria.
- By examining usage-informed evaluation criteria and the gaps in existing XAI evaluation practices, including nuanced differences between how end-users and XAI experts perceive the criteria differently, we call out implications to reflect on the common evaluation criteria as value-laden choices that will shape the field.

Related Work

XAI Methods and Usage Contexts

Although there is a lack of consensus on the definition of explainability, XAI works share a common goal of making AI models understandable by people. Many recent papers surveyed the *how* (methods) and *why* (objectives) of the XAI field (Guidotti et al. 2018; Arrieta et al. 2020; Adadi and Berrada 2018; Gunning et al. 2019; Lipton 2018; Gilpin et al. 2018; Molnar 2020). At a high level, the technical landscape of XAI can be divided into two camps: 1) directly interpretable models; 2) adopting "black-box" models such as deep neural networks and large tree ensembles, and then using post-hoc techniques to generate explanations. Guidotti et al. (2018) differentiate between model explanations (i.e. global explanations), local outcome explanations and model inspection. Under these categories, there are different explanation methods, such as leveraging features or examples to explain; and for each method, many techniques have been developed with differences in computational properties.

While some criticized that XAI techniques tend to be "developed in a vacuum" (Miller 2019), the field largely recognizes that there are "no one-fits-all solutions" (Arya et al. 2019) for the rich application opportunities of XAI. Recent works began to characterize the main user groups of XAI and postulate their different requirements (Preece et al. 2018; Hind 2019; Arrieta et al. 2020), including model developers, regulatory bodies, business owners, direct end-users, and impacted groups. Other studies identified similar roles based on empirical studies of ML practitioners (Bhatt et al. 2020; Hong, Hullman, and Bertini 2020).

A recent work by Suresh et al. (2021) points out that these frameworks lack granularity to distinguish between attributes of the users and their objectives to seek explanations. For example, people in any role may want to assess model biases or improve the model at certain usage points. Thereby Suresh et al. (2021) define stakeholders' knowledge and their objectives as two components that cut across to characterize the space of user needs for explainability. The authors further propose a multi-level typology to characterize XAI users' objectives, ranging from long-term goals (building trust and understanding the model), immediate objectives (debug, ensure regulatory compliance, take follow-up actions, justify actions influenced by a model, understand data usage, learn about a domain, contest model decisions), and specific tasks to perform with explanations.

Our definition of *XAI context* is similar to the "immediate objective" in Suresh et al.'s framework as characterizing a situation for which a user seeks explanations. In Section 3, we synthesize a list of prototypical XAI usage contexts with additional prior works reviewed. We choose to focus on XAI contexts following the objective-dependent stance of explainability in the social science literature (Lombrozo 2012; Keil 2006), along with a practical goal of informing context-specific design of XAI. We acknowledge that other factors such as user characteristics can further vary the evaluation criteria. Our work moves beyond the existing effort of characterizing the problem space to informing concrete context-specific requirements, aiming to provide actionable guidance for the evaluation and design of XAI.

XAI Evaluation

Our approach is informed by a bulk of prior research on XAI evaluation. Our focus is on "XAI evaluation criteria"—normative properties of "what constitute good explanations" (Hoffman et al. 2018)—which should be distinguished from outcome measurements of using XAI such

as identification of bugs or improvement of decisions. For normative criteria, earlier works focused on *model intrinsic* criteria such as faithfulness (how well the explanations approximate the original "black-box" model's decisions) and stability (how consistent the explanations are for similar cases). Recent work began to define criteria that capture aspects of human perception of "good" explanations, such as comprehensibility, actionability, and interactivity (Carvalho, Pereira, and Cardoso 2019; Sokol and Flach 2020).

However, we must recognize that there are varying priorities, even trade-offs, between these criteria depending on the context. For example, while an ML engineer might demand faithful explanations to engage in model debugging tasks, a layperson using a decision-support AI may be willing to sacrifice some degree of faithfulness for compactness. We can find support for this context-dependency in recent HCI works studying different XAI systems. For example, model debugging tools often integrate detailed local and global explanations (Narkar et al. 2021; Hohman et al. 2019). Decision-makers were found to have less desire for global explanations during time-constrained decisions, and prefer less distracting information (Xie et al. 2020). For AI capability assessment, a study hinted on that example-based explanations may have an advantage to expose users to the AI limitations (Buçinca et al. 2020).

As mentioned, a widely cited XAI evaluation framework is the taxonomy by Doshi-Velez and Kim (2017), proposing three categories with decreasing specificity and cost: application-grounded evaluation with humans and real tasks, human-grounded evaluation with humans and simplified tasks, and functionally grounded evaluation with no humans and proxy tasks (e.g., quantifying with some formal definition of human desired property). Our perspective builds on this framework and extends it by calling out the need to design simplified and proxy evaluation tasks based on evaluation criteria that are important to the target application context. For example, in a decision-making context with AI assistance, a common user objective is to have appropriate reliance, knowing when not to rely on the AI when it is likely to err. Recent work highlights the communication of model *uncertainty* as a desired property of XAI (Wang and Yin 2021; Carvalho, Pereira, and Cardoso 2019). However, this criterion cannot be captured by the commonly used simplified evaluation task, "simulatability test" (Buçinca et al. 2020). We speculate that methods that directly measure the success of uncertainty communication—such as by whether people can correctly judge if a model prediction is uncertain based on the explanation, or by quantifying the correlation between some notion of uncertainty salience in the explanations and ground-truth uncertainty—would make more effective evaluation tasks for this usage context.

Lastly, researchers have begun to explore a growing number of XAI systems (Hohman et al. 2019; Zhang, Liao, and Bellamy 2020; Kaur et al. 2020; Xie et al. 2020). However, the design choices and evaluation measurements used were largely inconsistent and ad-hoc, which hinders the development of scientific knowledge about human-XAI interaction and principled design guidelines. By exploring the prioritized evaluation criteria, i.e. desired properties, of explanations for prototypical XAI usage contexts, we also aim to inform the design choices and evaluation practices for researchers and practitioners working on XAI systems.

Taxonomies Development

We conducted a literature search to consolidate taxonomies of evaluation criteria and prototypical usage contexts of XAI. With Google Scholar (retrieved by December 2020), we used search terms "explainable/ interpretable AI/ML" + "evaluation/ assessment/ metrics" for the former, and + "goal/ objective/ context/ motivation/ use case" for the latter. After an initial review, we focused on a subset of papers with taxonomies or frameworks proposed. As enumerated below, we note that both taxonomies are necessarily incomplete given the fast advancement of the field. However, we consider them as sufficiently comprehensive, and contend our methodology can be used to extend the results.

XAI Evaluation Criteria

As mentioned, we focus on normative explanation "goodness" criteria, which can be further differentiated between *model intrinsic* properties of explanations (e.g., faithfulness), and *human-centered* properties that reflect the perception of the explainee (e,g., comprehensibility). Model-intrinsic properties can usually be measured by computational metrics while human-centered properties are best measured by human responses with questionnaires or behavioral measures. This differentiation is not binary, as it is possible to devise proxy measures to assess human-centered criteria. For example, while the "compactness" criterion (being succinct and not overwhelming) is contingent on the explainee's perception, it is possible to define some notion of "information units" to quantify the compactness of XAI output (Abdul et al. 2020). Lastly, we differentiate between an evaluation *construct* (*what* criterion) and an evaluation *method* or metric (*how* to measure). While our studies focus on the constructs, as we enumerate the criteria below, we also discuss existing methods to measure them, if any, or potential directions to develop new methods.

The papers we reviewed are distributed in the AI and HCI communities (Sokol and Flach 2020; Carvalho, Pereira, and Cardoso 2019; Yeh et al. 2019; Alvarez-Melis and Jaakkola 2018; Murdoch et al. 2019; Schneider and Handali 2019; Mohseni, Zarei, and Ragan 2018; Guidotti et al. 2018; Miller, Howe, and Sonenberg 2017; Jesus et al. 2021; Lakkaraju and Bastani 2020; Hancox-Li 2020; Gilpin et al. 2018; Doshi-Velez and Kim 2017; Kulesza et al. 2013, 2015; Hoffman et al. 2018; Hsieh et al. 2020). We found most criteria are covered by Carvalho, Pereira, and Cardoso (2019) and an "explainability requirements fact sheet" by Sokol and Flach (2020) (criteria under "usability requirements" instead of developer requirements). We develop our list based mainly on these papers, supplemented with additional items and definitions from others (sources are cited for each criterion below) We arrived at the following list of evaluation criteria, with definitions used in the survey in *italic*.

- **Faithfulness**: *The explanation is truthful to how the AI gives recommendations* (Alvarez-Melis and Jaakkola

2018), also referred to as fidelity (Carvalho, Pereira, and Cardoso 2019; Ras, van Gerven, and Haselager 2018) or soundness (Kulesza et al. 2013). It is a commonly used criterion, especially to evaluate post-hoc explanations, with computational metrics proposed in the literature (Alvarez-Melis and Jaakkola 2018; Yeh et al. 2019).

- **Completeness**: *The explanation covers all components that the AI uses to give recommendations, or can generalize to understand many AI recommendations* (Sokol and Flach 2020; Kulesza et al. 2013; Gilpin et al. 2018), also referred to as representativeness (Carvalho, Pereira, and Cardoso 2019). It is considered an orthogonal aspect to faithfulness for an explanation to accurately reflect the underlying model. According to Sokol and Flach (2020), it can be quantified by metrics that reflect the coverage or generalizability across sub-groups of a data set.
- **Stability**: *The explanation remains consistent for similar cases I ask about* (Carvalho, Pereira, and Cardoso 2019; Alvarez-Melis and Jaakkola 2018), also referred to as robustness (Hancox-Li 2020). While in some cases it can be at odds with faithfulness (if the model decision itself is unstable), stability is argued to be important if the goal is to understand not just the model, but true patterns in the world (Hancox-Li 2020; Alvarez-Melis and Jaakkola 2018). Prior works proposed several metrics (Alvarez-Melis and Jaakkola 2018; Hsieh et al. 2020).
- **Compactness**: *The explanation gives only necessary information and does not overwhelm*, also referred to as parsimony (Sokol and Flach 2020; Ras, van Gerven, and Haselager 2018). It reflects the design principles of "providing appropriate details" (Mueller et al. 2019; Kulesza et al. 2015). This criterion can be at odds with completeness. Computational metrics with some notion of information units or cognitive chunks have been proposed (Abdul et al. 2020; Doshi-Velez and Kim 2017) to proximate compactness, but capturing the essence of "necessary information" may require task-specific definitions or subjective responses.
- **(Un)Certainty (communication)**: *The explanation reflects the (un)certainty or confidence of the AI in its recommendations* (Carvalho, Pereira, and Cardoso 2019). Recent studies underscored this criterion to support appropriate reliance on AI (Zhang, Liao, and Bellamy 2020; Bansal et al. 2021). To our knowledge, there is no established method to measure this criterion. It is possible to devise survey scales to measure certainty perception or computational metrics with some definition of certainty representation, and then compare them against the true certainty or correctness of model predictions.
- **Interactivity**: *The explanation is interactive and can answer my follow-up questions* (Sokol and Flach 2020). Recent work considers interactivity as a necessary requirement for XAI, given the diverse knowledge gaps people have (Liao, Gruen, and Miller 2020; Weld and Bansal 2019). While being interactive can encompass many capabilities, we adopt a broad definition of allowing users to specify their explainbility needs by asking questions interactively. Behavioral measurements and computational metrics of interactivity can be conceived based on the scope of interactions the XAI can support.
- **Translucence**: *The explanation is transparent about its limitations, for example, the conditions for it to hold* (Carvalho, Pereira, and Cardoso 2019). It is referred to as contextfulness in Sokol and Flach (2020), which discussed multiple types of explanation limitations such as ambiguity and a lack of generalizability. While underexplored in the current literature, it is a criterion that could critically impact people's trust in the explanation itself. Translucence should be measured with regard to exposing the true limitations of the explanations.
- **Comprehensibility**: *The explanation is easy to understand (e.g., intuitive, taking less time to understand)* (Carvalho, Pereira, and Cardoso 2019), also referred to as clarity (Ras, van Gerven, and Haselager 2018), understandability (Guidotti et al. 2018), and explicitness (Alvarez-Melis and Jaakkola 2018). It is explainee-dependent and best measured by subjective or behavioral measures such as understanding correctness and speed, although it is possible to proximate with some quantifiable latent properties (Doshi-Velez and Kim 2017). While comprehensibility and compactness can correlate, the latter is more about cognitive workload and does not guarantee easiness to understand.
- **Actionability**: *The explanation helps me determine follow-up actions to achieve my goal for the task* (Sokol and Flach 2020). This criterion is contingent on the explainee's goal, thus should be measured by goal-specific subjective responses or behavioral measurement on the success of user goals specific to the evaluation task.
- **Coherence**: *The explanation is consistent with what I already know about the domain* (Sokol and Flach 2020). Miller (2019) posits that a desirable property of explanation is coherence with the explainee's prior knowledge. This is a subjective criterion and should ideally be measured by subjective responses.
- **Novelty**: *The explanation provides new or surprising information that I otherwise would not know* (Sokol and Flach 2020; Carvalho, Pereira, and Cardoso 2019). In some contexts, such as scientific discovery, the utility of explanation may stem from providing novel information to guide users. Novelty is also highly subjective and should be measured by self-reported responses.
- **Personalization**: *The explanation is tailored to my needs and preferences, e.g. level of details, language style, etc.* (Sokol and Flach 2020). Lastly, we include a criterion that focuses on the *communication style* according to one's preferences. Satisfaction with personalization should ideally be measured by subjective responses.

These criteria are *not* all orthogonal. Some can be correlated (e.g., compactness and comprehensibility) or involve trade-offs (e.g., coherence and novelty, coherence and faithfulness, completeness and compactness, stability and faithfulness). Such relationships indeed underline our motivation to identify their priorities in different usage contexts.

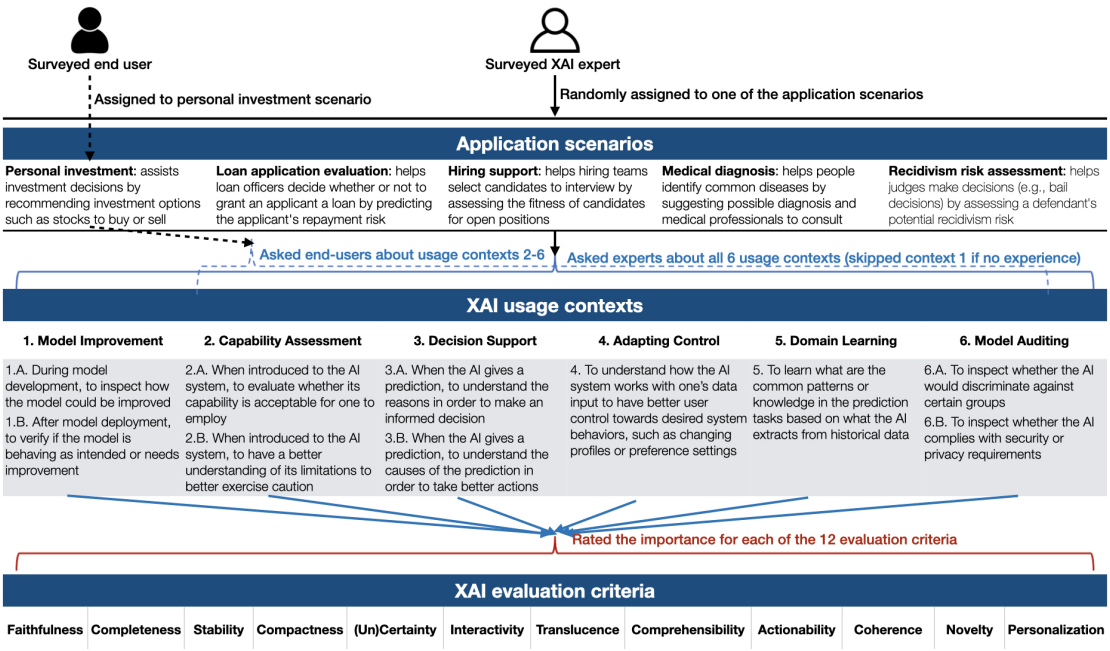

Figure 1: Experiment procedure and descriptions of AI applications and usage contexts. Each participant experienced one application, asked about all (5 for end-users) contexts, for each of which they rated the importance of all evaluation criteria

XAI Contexts

Our literature search on usage contexts of XAI discovered two bodies of work. First, many survey or perspective papers discussed conceptually the objectives for people to seek explanations (Doshi-Velez and Kim 2017; Adadi and Berrada 2018; Samek et al. 2019; Guidotti et al. 2018; Mohseni, Zarei, and Ragan 2018; Arrieta et al. 2020; Molnar 2020; Lipton 2018). For example, beside Suresh et al. (2021) reviewed earlier, Chen et al. (2021) emphasizes the need to make choices of XAI techniques according to use case goals and proposes a taxonomy of XAI use cases including model debugging, promoting trust in a model, assisting scientific discovery, assisting human decision-making, and providing actionable recourse. Another body of work summarizes common user objectives with XAI by empirically studying real-world use cases (Hong, Hullman, and Bertini 2020; Bhatt et al. 2020). We choose to mainly build on Liao, Gruen, and Miller (2020), which summarizes user objectives with XAI by studying 16 real-world AI systems, supplemented with additional definitions from others. Figure 1 shows our list of usage contexts: Model Improvement, Capability Assessment, Decision Support, Adapting Control, Domain Learning, and Model Auditing. We intentionally kept these definitions general to avoid priming participants' judgment by a specific form of explanation. Following Liao et al., under some categories we consider sub-categories that share a similar goal but have different focuses (e.g., assess AI capability versus limitations). These sub-categories were randomly selected to present to participants.

Survey Studies

We solicited people's opinions on the relative importance of XAI evaluation criteria with two scenario-based survey studies. First, we surveyed XAI experts from both the AI and HCI communities for multiple types of AI applications. We then surveyed a selected group of crowd workers as target users of an XAI application assisting personal investment. We first describe the design and analysis of the two surveys separately, and then discuss results from both together.

Study 1: XAI Expert Survey

Survey Design The survey presents different contexts in which a stakeholder seeks explanations from an AI application. To explore whether results vary by domain, we created the same set of contexts for five decision-support AI applications. We started with an AI application scenario database created by Lubars and Tan (2019), then narrowed down our choices based on the following criteria: 1) It should be feasible for current AI so there is a shared expectation of how the AI works; 2) It should be commonly known so participants can easily imagine the scenario; 3) The decision is high-stakes so the need to understand AI likely presents. We selected the 5 AI application scenarios described in Figure 1: personal investment, loan application evaluation, hiring support, medical diagnosis, and recidivism risk assessment.

The experiment procedure is illustrated in Figure 1. Each expert participant was randomly assigned to questions regarding one of the 5 AI applications. The application was introduced in the instructions and presented at the top throughout the survey, including its functions, the data to train it, and example features of the model (See the survey page screen-

shot in Appendix A[1]). Descriptions of the 5 applications are also in Appendix A. On each survey page, participants were presented with one of the 6 XAI usage contexts in Figure 1 as a task that a stakeholder performs (see the Task screenshot in Appendix A). We customized the task descriptions for the given application. Participants finished survey pages for all 6 XAI contexts in Figure 1 (one of the sub-categories, if any, was randomly chosen), with the exception of Context 1: Model improvement. Participants were asked whether they were experienced in model development, and were only presented with Context 1 if their answer was yes. All but two expert participants answered yes. For each task, participants were asked to rate the importance of 12 evaluation criteria that explanations should satisfy for them to accomplish the task. All responses are five-point Likert-type from "not important" to "very important".

In summary, our survey introduces three independent variables for the analysis: each participant was assigned to questions regarding one of the five *AI applications*, performed judgment tasks with all *XAI contexts*, and rated all *evaluation criteria* for each context. The importance ratings for the evaluation criteria are used as the dependant variable.

Participants were given a consent form and instructions in the beginning, and asked to rate their expertise in XAI and confidence in addressing the AI application at the end. For each context, we asked an optional question to comment on the reasons of their ratings. 5 participants left at least one comment. We also asked participants to leave their contact information if they are willing to answer follow-up questions. We emailed those who did to solicit comments on their ratings and 5 responded. In the result section, we discuss some observations from these qualitative data.

Participants: XAI Experts We adopted a snowball sampling strategy by initially contacting 25 researchers in different organizations, who have published substantially on XAI. We strove to have a balance of AI and HCI researchers in our initial contacts to ensure both algorithmic and user perspectives were presented. We asked them to forward the recruiting message to other people working on the topic, emphasizing that this survey targets XAI experts. Participants' information based on the exit survey is summarized in Appendix C. Notably, 22 out of the 35 participants consider themselves to be "very knowledgeable" on the topic of XAI and none has "no knowledge" (another 11 knowledgeable, and 2 somewhat knowledgeable). 32 have the job title of researcher (the remaining 3 are data scientists).

Expert Survey Analysis We performed a mixed-effects regression analysis on participants' ratings, by including the AI application (between-subjects), XAI context (within-subjects) and evaluation criterion (within-subjects) as fixed-effects variables, and participant as a random-effects variable[2]. We found the main effect of XAI context ($F(5, 2106) = 3.351, p = 0.004$), evaluation criterion ($F(11, 2106) = 2.589, p = 0.003$), and the interactive effect between the two ($F(55, 2106) = 3.351, p = 0.038$) to be significant. The main effect of AI application is not significant ($F(4, 30) = 1.108, p = 0.371$), nor any of its interactive effects. These patterns indicate that experts' relative ratings of evaluation criteria were significantly varied by the XAI contexts, but not the types of AI application.

Given the *non-significance of AI application*, we merged data across the five applications. For each XAI context, we conducted an omnibus-ANOVA analysis on the ratings, with evaluation criterion as a within-subject variable. The main effect of evaluation criteria is significant ($p < 0.001$) for all six XAI contexts. We then conducted pairwise post-hoc t-tests for all contexts, using false discovery rate (FDR) adjustment to adjust p-values for multiple comparisons. The first part of the table in Figure 2 presents the mean ratings provided by expert participants. The last column presents the overall mean ratings by combining ratings for all 6 XAI contexts. We include all the post-hoc analyses statistics in Appendix D. We acknowledge that given the relatively small sample size, the individual post-hoc analysis may not have enough power to be conclusively insignificant. In Section 4.3 we discuss patterns, and highlight comparisons that are statistically significant even with the small sample.

Study 2: End-User Crowd Survey

Survey Design We chose to use only one scenario to conduct the end-user survey—personal investment. Compared to the other applications, its target user group are more general and thus easier to recruit crowd workers as suitable target users. Furthermore, the analysis of the XAI expert survey showed no significant effect of the application scenario. The "Model Improvement" context was removed since most workers do not have relevant experience.

The survey design was largely similar to the expert survey, except that we also designed a *training task* to introduce crowd workers who are not familiar with XAI to the definitions of the 12 evaluation criteria. As presented in Appendix B, using a movie recommender scenario, each training task page introduces the definition of one evaluation criterion, shows two explanations, and asks participants to choose which better exhibits the given criterion. After submitting the response, the page indicates the right choice and highlights the corresponding part of the explanation that exhibits the criterion considered. In the survey, participants could hover over a criterion and a pop-up window would show the training task to help them recall the definition. Participants were told that they would only be qualified for the survey if they pass at least half of the 12 training tasks. If they fail, they would be compensated with a base payment of $2 USD. If they finish the survey they would receive an additional $4 USD in compensation. Most participants took 20-40 minutes to complete the survey.

Participants: End-Users We added a filter "Financial Asset Owned" on Mechanical Turk to target workers with experience in financial investment. In total, we gathered completed survey responses from 37 people, and excluded data

[1] Appendixes can be found at tinyurl.com/xai-eval

[2] We acknowledge the potential loss of information by modeling Likert-type responses using parametric tests (Owuor 2001). We also modeled the data with an ordinal logistic regression, and the patterns largely hold. However, the results are more difficult to interpret. We use parametric tests for presentation clarity.

points from 5 workers who took less than 10 minutes (less time than any experts). The background information of the 32 participants is shown in Appendix C. In contrast to participants of the expert survey, none consider themselves to be "very knowledgeable" on XAI, and 27 to be "no knowledge" or "somewhat knowledgeable". The two groups are comparable in other dimensions such as confidence in addressing the AI application and demographic attributes. Possibly due to the filter used, these participants appear to be relatively tech-savvy compared to the general population, which we consider as consistent with a user population that would use an AI assisted investment application.

End-User Survey Analysis We started by examining whether the experts' and end-users' ratings are significantly different. We performed a mixed-effect regression on ratings by including XAI context, evaluation criterion and participants group (experts/end users) as fixed-effects variables, and participant as random effects. We found the main effect of group ($F(1, 65) = 0.436, p = 0.511$) as well as the interactive effect of XAI context, evaluation criterion, and group ($F(44, 3835) = 0.970, p = 0.529$) to be non-significant; but the two-way interaction between evaluation criterion and group is significant ($F(11, 3835) = 2.689, p = 0.002$). These patterns suggest that there is no significant difference between how XAI experts and end-users considered the relevant importance of XAI evaluation criteria in different XAI contexts, but there might be differences in their overall ratings of evaluation criteria, which we discuss later.

An ANOVA analysis was conducted on the ratings for each XAI context from end-users. We found the main effect of evaluation criterion to be significant for every XAI context ($p < 0.001$), then conducted post-hoc pairwise comparisons. Detailed statistics are in Appendix D. The second part of the table in Figure 2 presents the mean ratings of evaluation criteria for each context provided by end-user participants. The last column presents the overall mean ratings by combining all five XAI contexts (no "Model Improvement" in the end-user survey). We also merged data from experts and end-user surveys and conducted the same ANOVA and pair-wise analyses, with the combined mean ratings presented in the third part of Figure 2 and the detailed statistics are in Appendix D.

Discussions of Survey Results

We refer to the table in Figure 2 to discuss the results from both surveys. Each column presents participants' mean ratings for evaluation criteria for one usage context, in decreasing order. The last column presents the overall mean ratings by combining data across all contexts. The first part of the table shows ratings from the expert survey; the second part is from the end-user survey; and the last part combines data from the two surveys. As mentioned, the three-way interaction between usage context, evaluation criterion and participant group is non-significant, meaning the experts and end-users did not have a statistically significant difference in how they consider the relative importance of evaluation criteria for different contexts. Therefore, below we start with patterns in the combined results, and then discuss the nuanced differences between experts' and end-users' ratings.

Fonts and color coding are used to visually highlight a few patterns in Figure 2. First, we bold the "top group" for each context. While our sample size might have lacked enough power to conclude all insignificant pairwise comparisons, we adopt the following heuristic: for each XAI context, we look at the pairwise comparisons with the top criterion, and consider criteria that had less than marginal significance ($p > 0.100$) to be closer to the top criterion than others, as in the top group. Second, we color code the relative importance of evaluation criteria that *vary in different XAI contexts*. For each set of results (experts/end-users/combined), we use the criteria positions in the last column of overall ratings as a reference and highlight prominent differences in other columns. We visually highlight the differences with the following heuristics: 1) we code criteria that increase more than 2 ranks in *red* and double-underlines; 2) we code criteria that increase 1-2 in ranks and also enter the "top group" in *orange* and underline; 3) we code criteria that decrease more than 2 ranks in *blue* and *italic* . In short, the red and orange colors indicate a relative increase in position for a usage context compared to others, while the blue color indicates a relative decrease. Based on Figure 2, in the following, we first enumerate the patterns shown for each usage context, then summarize some take-aways.

Patterns by Usage Contexts *Across all contexts*, faithfulness was considered the most important criterion, followed by translucence, uncertainty and stability. Both experts and end-users also considered interactivity and comprehensibility to be relatively important. Novelty, compactness and coherence were rated consistently at the bottom. While experts also rated personalization to be of little importance, end-users considered it slightly more important. Below we elaborate on the results for each usage context.

Model improvement. Only experts provided ratings for this context. Faithfulness and stability were rated as the top criteria. Stability was especially deemed more important than for other contexts, while interactivity was considered less of a requirement than for other contexts. These patterns suggest that to support model improvement, such as debugging, it is important to *provide faithful and stable explanations for users to inspect the true causes of model issues*. Since users are often ML experts to perform this task, they may be more receptive to complex or information-rich explanations (Narkar et al. 2021; Hohman et al. 2019) without the need for interactive inquiries, as evidenced in P20's response in explaining their ratings:"*It is vital to know how much can be relied upon the explanation. Hence, the need for fidelity, stability and translucence... I deem interactivity as relatively less important due to the assumption that an expert attempts to improve it and thus has the skills and competencies to get to the bottom of some model behavior.* "

Capability assessment. For the combined results, comprehensibility, uncertainty, translucence, faithfulness, stability and personalization were rated as top criteria. Comprehensibility, uncertainty and personalization were rated more important for capability assessment than for other contexts, while faithfulness was comparatively lower. Interactivity

Model Improve.		Capability Assess.		Decision Support		Adapt Control		Domain		Model Auditing		Overall	
Expert Survey													
1. **Faithfulness**	**4.61**	**Translucence**	**4.40**	**Interactivity**	**4.29**	**(Un)Certainty**	**4.31**	**Translucence**	**4.23**	**Faithfulness**	**4.54**	**Faithfulness**	**4.30**
2. **Stability**	**4.27**	**(Un)Certainty**	**4.29**	**(Un)Certainty**	**4.26**	**Interactivity**	**4.31**	**Faithfulness**	**4.20**	**Translucence**	**4.43**	**Translucence**	**4.25**
3. (Un)certainty	4.24	**Comprehend.**	**4.20**	**Faithfulness**	**4.14**	**Translucence**	**4.29**	**Stability**	**4.17**	**Completeness**	**4.37**	(Un)Certainty	4.15
4. Translucence	4.06	***Faithfulness***	**4.09**	**Translucence**	**4.11**	***Faithfulness***	**4.26**	Comprehen.	**4.14**	(Un)Certainty	4.00	Stability	4.01
5. Comoprehen.	3.94	**Stability**	**4.00**	Comprehend.	**3.97**	**Stability**	**4.00**	**Completeness**	**4.00**	Interactivity	3.89	Interactivity	3.92
6. Completeness	3.67	**Compactness**	**4.00**	Actionability	**3.91**	Actionability	**3.86**	**Interactivity**	**3.86**	*Stability*	3.83	Comprehend.	3.89
7. Actionability	3.48	Interactivity	3.80	*Stability*	3.83	Completeness	3.80	*(Un)Certainty*	3.83	Comprehend.	3.49	Completeness	3.70
8. Compactness	3.42	Actionability	3.69	Personalization	3.43	Comprehen.	3.63	Compactness	3.60	Actionability	3.06	Actionability	3.55
9. *Interactivity*	3.33	Personalization	3.63	Compactness	3.31	Coherence	3.46	Personalization	3.57	Compactness	3.03	Compactness	3.45
10.Coherence	3.18	Coherence	3.54	Novelty	3.20	Compactness	3.31	Actionability	3.29	Coherence	2.97	Coherence	3.23
11.Novelty	2.88	*Completeness*	3.23	Coherence	3.17	Personalization	2.94	Novelty	3.17	Personalization	2.57	Personalization	3.14
12.Personalizatio	2.67	Novelty	2.51	*Completeness*	3.11	Novelty	2.91	Coherence	3.03	Novelty	2.37	Novelty	2.82
End-User Survey													
		Comprehen.	**4.28**	**(Un)Certainty**	4.38	**Translucence**	**4.44**	**Stability**	**4.38**	**Faithfulness**	**4.78**	**Faithfulness**	**4.38**
		Personalization	**4.28**	**Translucence**	4.38	**Faithfulness**	**4.38**	**Faithfulness**	**4.28**	Translucence	4.28	**(Un)Certainty**	**4.26**
		(Un)Certainty	**4.19**	**Faithfulness**	4.31	**(Un)Certainty**	**4.38**	**(Un)Certainty**	**4.25**	Completeness	4.19	**Translucence**	**4.25**
		Faithfulness	**4.12**	Comprehen.	4.19	**Stability**	**4.19**	Comprehen.	**4.25**	(Un)Certainty	4.12	Stability	4.15
		Stability	**4.12**	Actionability	4.03	**Personalization**	**4.09**	**Translucence**	**4.09**	Stability	4.12	Comprehen.	4.05
		Translucence	**4.06**	Stability	3.94	Actionability	4.03	**Completeness**	**4.09**	Interactivity	3.97	Interactivity	3.91
		Actionability	3.88	Interactivity	3.81	Interactivity	3.97	Personalization	**4.06**	Comprehen.	3.72	Actionability	3.89
		Coherence	3.84	Personalization	3.75	*Comprehen.*	3.81	**Interactivity**	**4.03**	Actionability	3.66	Personalization	3.88
		Interactivity	3.75	Coherence	3.72	Completeness	3.72	Actionability	3.84	Coherence	3.59	Completeness	3.86
		Completeness	3.69	Completeness	3.59	Coherence	3.66	Coherence	3.72	Personalization	3.19	Coherence	3.71
		Compactness	3.12	Novelty	3.16	Novelty	2.88	Compactness	3.59	Novelty	3.12	Novelty	3.04
		Novelty	3.00	Compactness	2.84	Compactness	2.78	Novelty	3.06	Compactness	2.62	Compactness	2.98
Combined Results													
		Comprehen.	**4.24**	**(Un)Certainty**	**4.31**	**Translucence**	**4.46**	**Stability**	**4.27**	**Faithfulness**	**4.66**	**Faithfulness**	**4.31**
		(Un)Certainty	**4.24**	**Translucence**	**4.24**	**(Un)Certainty**	**4.34**	**Faithfulness**	**4.24**	Translucence	4.36	**Translucence**	**4.27**
		Translucence	**4.24**	**Faithfulness**	**4.22**	**Faithfulness**	**4.31**	**Comprehen.**	**4.19**	Completeness	4.28	(Un)Certainty	4.20
		Faithfulness	**4.10**	Comprehen.	**4.07**	Interactivity	**4.15**	**Translucence**	**4.04**	(Un)Certainty	4.06	Stability	4.05
		Stability	**4.06**	**Interactivity**	**4.06**	**Stability**	**4.09**	**Completeness**	**4.03**	Stability	3.97	Interactivity	3.97
		Personalization	**3.94**	**Actionability**	**3.97**	Actionability	3.94	***(Un)Certainty***	**3.94**	Interactivity	3.93	Comprehen.	3.96
		Actionability	3.78	*Stability*	3.88	Completeness	3.76	**Interactivity**	**3.81**	Comprehen.	3.60	Completeness	3.78
		Interactivity	3.78	Personalization	3.58	Comprehen.	3.72	Personalization	3.55	Actionability	3.34	Actionability	3.72
		Coherence	3.69	Coherence	3.43	Coherence	3.55	Actionability	3.55	Coherence	3.27	Personalization	3.54
		Compactness	3.58	*Completeness*	3.34	Personalization	3.49	Compactness	3.36	Personalization	2.87	Coherence	3.46
		Completeness	3.45	Novelty	3.18	Compactness	3.06	Coherence	3.12	Compactness	2.84	Compactness	3.22
		Novelty	2.75	Compactness	3.09	Novelty	2.90	Novelty		Novelty	2.73	Novelty	2.93

Figure 2: Average importance ratings of evaluation criteria by XAI experts, end-users and combined. The later digits were used for ties. Bold fonts are "top groups". Double-underlined red ones increase more than 2 ranks compared to the overall ranking; Single-underlined orange ones increase 1-2 ranks and also enter the top group; Italic blue ones decrease more than 2 ranks.

and completeness were considered less important. End-users especially rated comprehensibility and personalization high, while experts rated compactness higher than for other contexts. These patterns suggest that to support capability assessment, such as during users' onboarding stage, it is important to provide *easy-to-understand explanations to help users make an efficient assessment*, even at some cost of faithfulness; *being transparent about model uncertainty and explanation limitations* are also important; but *complete and interactive explanations may not be necessary as users prioritize intuitiveness and efficiency*. A nuanced difference is that end-users considered personalization, while experts considered compactness, as important to make an efficient assessment. P33 explained their willingness to compromise faithfulness and completeness for comprehensibility: "*Fidelity is still important, just didn't seem to be as important... because some omission on details may be beneficial for first-comers. Same goes for completeness, as users may just want to assess the model within the context of their own use.*"

Decision support. Uncertainty, translucence, faithfulness, comprehensibility, interactivity and actionability were rated closely high in importance. Uncertainty, comprehensibility and actionability were deemed more important than for other contexts, while stability and completeness were less important. Interestingly, experts and end-users disagreed on the importance of interactivity, with experts rating it as a top criterion. These patterns suggest that for decision support, it is important that explanations should *communicate uncertainty of model predictions and the limitations of explanations*; it is also important to ensure the explanations are *easy to understand and actionable for the decision goal*, which may also require *being interactive* to help users reach the necessary understanding. Since decision support often focuses on explaining individual predictions, stability (whether it applies to similar cases) and completeness (whether it generalizes or covers all decision components) are less important. As P19 commented: "*Completeness is more important when I need an understanding of the whole system... here this is mostly not the case. Instead it is very important to know the certainty, so I can take the prediction with 'a grain of salt' if necessary and be extra careful. Interactivity is very important when I have the feeling I need to 'dig deeper' ... Comprehensibility is so that I learn to integrate the explanation into my work routine without losing efficiency*".

Adapting control. Translucence, uncertainty, faithfulness, interactivity and stability were rated as top criteria. It suggests that to support a better understanding of how the system works with one's data to make adjustment, explanations should *be transparent about the model uncertainty and explanation limitations*, and also *be faithful and stable to allow users to discover and change the true causes of sub-optimal system behaviors*. Interestingly, comprehensibility was rated lower than for other contexts, suggesting a willingness to invest time for this task that may only be performed occasionally but has a lasting impact, as P10 commented: "*I think that comprehensibility can be lower here because the user has*

time to stop and think for a while." Experts rated interactivity higher for this context, but not end-users; instead end-users rated personalization higher. They suggest the two groups may have conceptualized the requirements differently. Since adapting control requires users to locate precise causes they can control, such as changeable preference settings, experts may have envisioned it as an interactive process to narrow down the causes, while end-users may have preferred the system to proactively adapt to their desired level of details. P19 (expert)'s comment provides support: "*Having low certainty tells me where I need to tweak it further. Translucence is important because I do not want to make any changes based on information that was not even relevant. Stability is important because if I instruct the system to behave differently I expect it to generalize this behavior to other similar cases. Interactivity empowers me to dig deeper in cases where I am puzzled why a certain behavior occurred and may test some what-if hypothesis.* "

Domain learning. Stability, faithfulness, comprehensibility, translucence, completeness, uncertainty and interactivity were rated as top criteria. Stability, comprehensibility and completeness were rated higher than for other contexts, while uncertainty was rated less important, mainly due to experts' opinions. These patterns suggest that to support learning, explanations should *be stable, faithful and relatively complete to help people discover true patterns of the domain*; it should also be *easy-to-understand*, as users may not be ML experts or interested in the model details. P20 commented on the importance of stability, faithfulness, comprehensibility and translucence: "*These four properties are in my opinion needed to enable a user to ascertain actionable knowledge about a domain. This matches with the literature on education and training—humans require reliable explanations they can understand to acquire new knowledge.*"

Model auditing. Only faithfulness is in the top group, as a clear winner for this usage context, followed by translucence. Meanwhile, completeness was rated higher for this context than for others. These patterns suggest that in order to support model auditing, explanations should be first and foremost *faithful to the underlying model*, *transparent about the explanation limitations*, and *cover the decision components completely, to allow accurate and comprehensive auditing*, as supported by P2's reasoning: "*faithfulness and translucence are important because I need to ensure that I am diligent when checking for legal-compliance. Hence I need explanations that are truthful to the model (rather than basing my assessment on potentially wrong information) and cover all components of the system being assessed.*"

Implications for Designing and Evaluating XAI We reflect on the results above and discuss some implications.

Faithfulness is critical, but can be compromised in some contexts. Both experts and end-users recognize that faithfulness is the most important criterion across the board, especially for model improvement and auditing to enable diagnosing the true model issues, validating the field's current focus on quantifying faithfulness (Alvarez-Melis and Jaakkola 2018; Yeh et al. 2019). However, participants considered it somewhat compromisable in other contexts such as capability assessment. Importantly, these observations imply that whether post-hoc explanations or explanations with less faithful details are desirable depends on the usage contexts.

Translucence and uncertainty are important requirements but not well-supported in current XAI research. A striking result is that participants consistently rated translucence and uncertainty communication high for all contexts, despite currently limited XAI approaches that support these criteria (though there has been a long line of research on directly quantifying uncertainty (Bhatt et al. 2021; Ghosh et al. 2021)). We highlight such user needs for the field to focus on, including defining evaluation tasks that capture whether an XAI technique can support users to make correct judgments of model uncertainty and explanation limitations.

Requirement for comprehensibility is prominent when efficiency and reducing cognitive load matter. Comprehensibility was generally rated important but less so for adapting control, model auditing and model improvement, all of which require more engagement with inspecting and changing the model. In participants' qualitative responses, its high priority was frequently mentioned together with the importance of efficiency, or users' unwillingness to spare time or cognitive resources, such as when gauging system capability or prediction reliability in everyday use.

Stability and completeness are important when having a generalizable understanding matters. Stability is a criterion that reflects consistency and robustness of explanations for similar cases. Completeness reflects how well the explanation covers all components (e.g., features, decision paths) so users can apply the knowledge to understand not one but many predictions. We found the relative ranks of the two often have correlated changes: both were rated more important for domain learning, but less important for decision support. We postulate that both criteria are desired when users aim to identify generalizable and robust patterns. Curiously, completeness was rated low for capability assessment. We suspect the reason is that completeness is seen as requiring more time and cognitive effort, as evidenced by the often negative correlation in the rank movement between completeness and comprehensibility.

Requirements for interactivity and personalization are perceived differently by XAI experts and end-users. Experts rated interactivity high for decision support and adapting control, but end-users did not consider it as important; while experts consistently rated personalization at the bottom, end-users considered it a top criteria for capability assessment, adapting control and domain learning. While caution should be exercised interpreting the results as the two groups may have different conceptualizations, including what is technically feasible, of these criteria, these differences are worth reflecting on and being further explored. Interactivity is increasingly discussed in XAI literature as an important direction for the field (Miller 2019; Weld and Bansal 2019), which would allow user inquiries to guide the provision of desired explanations to achieve their goal. This trend may have influenced experts' positive ratings on interactivity. It does not necessarily mean well-designed interactivity is undesired by end-users, but at a conceptual level, they might have preferred the system to take the initiative to provide de-

sired explanations directly, such as through personalization. The difference between the two groups' ratings on personalization may reflect a gap between what experts of the field prioritize and what end-users actually need.

Requirement for actionability is neutral. The reason could be that the success of supporting users' end goals is better captured by other more specific criteria (e.g., comprehensibility for capability assessment) rather than the generic statement about "supporting follow-up actions". Given that actionability is also hard to operationalize without running an application-grounded evaluation, we postulate that when designing simplified and proxy evaluation tasks, it may be more productive to focus on the other top criteria.

The importance of explanation novelty, coherence and compactness is unclear. Despite being discussed in the literature as important usability requirements (Sokol and Flach 2020), these criteria were consistently rated at the bottom. It is possible that some other top criteria correlate with them, but better capture people's desired properties. For example, comprehensibility could be a more informative criterion than compactness for the field to focus on. As shown in the example used in our training task, an example-based explanation could be more intuitive, but not more compact compared to a feature-based explanation. As for novelty and coherence, it is possible that personalization could better capture the desired relation between explanations and ones' prior knowledge. We also believe more research is required to explore the roles of these criteria in actual user interactions beyond the retrospective approach we took.

Discussions and Future Work

Our central thesis is that to close the gaps between algorithmic research and their effectiveness in deployment necessitates explicitly considering different requirements in different downstream usage contexts. Towards this goal, we introduce a perspective of contextualized evaluation. We empirically show that the relative importance of explanation "goodness" criteria varies across prototypical usage contexts of XAI. We suggest ways for future work to use our results.

Articulating the appropriate usage contexts of XAI algorithms. This is one area where contextualized evaluation is critically needed for responsible research and enabling others, including practitioners, to make appropriate use of research outputs. It is possible to use our results to perform a weighted analysis for that purpose. Assuming measurement scores of a given XAI technique can be obtained for all the evaluation criteria, one can use the importance ratings in Figure 2 to inform a set of weights for each usage context, then generate a weighted sum of evaluation score for each context respectively. Contexts with low weighted scores should be cautioned against applying the given technique. For example, an example-based explanation may score relatively low in stability and completeness, but high in uncertainty communication and comprehensibility. Its weighted scores would be high for decision-support but low for model improvement and auditing, which is consistent with observations in HCI studies (Cai, Jongejan, and Holbrook 2019; Buçinca et al. 2020; Dodge et al. 2019). Analyses can also be performed with a set of existing XAI techniques to benchmark desired scores for different usage contexts.

Optimizing XAI for a given usage context. Another common scenario where our results can be applied is to optimize XAI features for a given usage context. A researcher or practitioner's task may be to select from available XAI techniques, or to evaluate a XAI technique and identify its shortcomings to make improvements. The results in Table 2 can help guide which evaluation criteria to focus on. Especially when performing a full application-grounded evaluation is beyond the resource allowed, one can focus on the top criteria identified for the given usage context using simplified or proxy metrics, or consider how to further optimize these top desired properties. For instance, when developing an XAI feature to support domain learning, one may want to prioritize using XAI techniques that score high in stability and faithfulness. One may also need to improve a current XAI feature by enhancing comprehensibility and translucence.

Reflecting on what evaluation criteria the field prioritizes as value-laden choices. Our work has another important goal—empirically examining the desired criteria in usage contexts and the gaps in what the field currently focuses on, as the choices of evaluation metrics can profoundly shape the outputs of a field. In most usage contexts, participants cared deeply about uncertainty communication and translucence, but these criteria about exposing the negative aspects of models and explanations are missing in current XAI work. If the field continues focusing solely on model intrinsic properties such as faithfulness, we may risk losing sight of what actually matters for users to achieve their objectives. We also contend that the negligence of diverse usage contexts of XAI can lead to blind spots in the field's focuses. For example, while it is true that faithfulness is paramount for model debugging, which is the use case that XAI researchers tend to fixate on, our results suggest that users are willing to compromise it to a degree for efficient comprehensibility for other usage contexts such as capability assessment. Furthermore, our results reveal nuanced differences in how end-users and experts in the field perceive some of the criteria, specifically pointing to blind spots in what users desire from personalization and interactivity.

Limitations and future directions The contribution of our study is made by soliciting people's opinions to illustrate patterns of varying relative importance of evaluation criteria. The quantification should be taken with caution. First, our sample size is relatively small. Second, the retrospective approach by soliciting opinions, while allowing exploration of the large problem space, may not reflect the effect size in actual perceptions and behaviors. Future work should explore complementary approaches, such as experimental studies, to further formalize the quantification of their relative importance. We further acknowledge that by synthesizing the list of evaluation criteria from current literature, our work does not explicate their underlying requirements, technical feasibility, relations between these criteria, or how they are perceived by end-users, which should be explored in future work. Our study also only begins to touch on how to measure these criteria, and we invite future work to explore measurements through simplified or proxy evaluation tasks.

Acknowledgements

We wish to thank all participants for their generous time and inputs. We also wish to thank all reviewers for their thoughtful feedback. The majority of the work was completed while the first and second authors were working at IBM Research. FDV also received support from NSF IIS-2107391.

References

Abdul, A.; von der Weth, C.; Kankanhalli, M.; and Lim, B. Y. 2020. COGAM: Measuring and Moderating Cognitive Load in Machine Learning Model Explanations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14.

Adadi, A.; and Berrada, M. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access*, 6: 52138–52160.

Alvarez-Melis, D.; and Jaakkola, T. S. 2018. Towards robust interpretability with self-explaining neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 7786–7795.

Arrieta, A. B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58: 82–115.

Arya, V.; Bellamy, R. K.; Chen, P.-Y.; Dhurandhar, A.; Hind, M.; Hoffman, S. C.; Houde, S.; Liao, Q. V.; Luss, R.; Mojsilović, A.; et al. 2019. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012*.

Bansal, G.; Wu, T.; Zhou, J.; Fok, R.; Nushi, B.; Kamar, E.; Ribeiro, M. T.; and Weld, D. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16.

Bhatt, U.; Antorán, J.; Zhang, Y.; Liao, Q. V.; Sattigeri, P.; Fogliato, R.; Melançon, G.; Krishnan, R.; Stanley, J.; Tickoo, O.; et al. 2021. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 401–413.

Bhatt, U.; Xiang, A.; Sharma, S.; Weller, A.; Taly, A.; Jia, Y.; Ghosh, J.; Puri, R.; Moura, J. M.; and Eckersley, P. 2020. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 648–657.

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 25th International Conference on Intelligent User Interfaces*, 454–464.

Cai, C. J.; Jongejan, J.; and Holbrook, J. 2019. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th international conference on intelligent user interfaces*, 258–262.

Carvalho, D. V.; Pereira, E. M.; and Cardoso, J. S. 2019. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8): 832.

Chen, V.; Li, J.; Kim, J. S.; Plumb, G.; and Talwalkar, A. 2021. Towards Connecting Use Cases and Methods in Interpretable Machine Learning. *arXiv preprint arXiv:2103.06254*.

Dodge, J.; Liao, Q. V.; Zhang, Y.; Bellamy, R. K.; and Dugan, C. 2019. Explaining models: an empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th international conference on intelligent user interfaces*, 275–285.

Doshi-Velez, F.; and Kim, B. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Ehsan, U.; and Riedl, M. O. 2021. Explainability Pitfalls: Beyond Dark Patterns in Explainable AI. *arXiv preprint arXiv:2109.12480*.

Ghosh, S.; Liao, Q. V.; Ramamurthy, K. N.; Navratil, J.; Sattigeri, P.; Varshney, K. R.; and Zhang, Y. 2021. Uncertainty Quantification 360: A Holistic Toolkit for Quantifying and Communicating the Uncertainty of AI. *arXiv preprint arXiv:2106.01410*.

Gilpin, L. H.; Bau, D.; Yuan, B. Z.; Bajwa, A.; Specter, M.; and Kagal, L. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, 80–89. IEEE.

Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; and Pedreschi, D. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5): 1–42.

Gunning, D.; Stefik, M.; Choi, J.; Miller, T.; Stumpf, S.; and Yang, G.-Z. 2019. XAI—Explainable artificial intelligence. *Science Robotics*, 4(37).

Hancox-Li, L. 2020. Robustness in machine learning explanations: does it matter? In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 640–647.

Hase, P.; and Bansal, M. 2020. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? *arXiv preprint arXiv:2005.01831*.

Hind, M. 2019. Explaining explainable AI. *XRDS: Crossroads, The ACM Magazine for Students*, 25(3): 16–19.

Hoffman, R. R.; Mueller, S. T.; Klein, G.; and Litman, J. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608*.

Hohman, F.; Head, A.; Caruana, R.; DeLine, R.; and Drucker, S. M. 2019. Gamut: A design probe to understand how data scientists understand machine learning models. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, 1–13.

Hong, S. R.; Hullman, J.; and Bertini, E. 2020. Human factors in model interpretability: Industry practices, challenges, and needs. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1): 1–26.

Hsieh, C.-Y.; Yeh, C.-K.; Liu, X.; Ravikumar, P.; Kim, S.; Kumar, S.; and Hsieh, C.-J. 2020. Evaluations and methods for explanation through robustness analysis. *arXiv preprint arXiv:2006.00442*.

Jesus, S.; Belém, C.; Balayan, V.; Bento, J.; Saleiro, P.; Bizarro, P.; and Gama, J. 2021. How can I choose an explainer? An Application-grounded Evaluation of Post-hoc Explanations. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 805–815.

Jeyakumar, J. V.; Noor, J.; Cheng, Y.-H.; Garcia, L.; and Srivastava, M. 2020. How can i explain this to you? an empirical study of deep neural network explanation methods. *Advances in Neural Information Processing Systems*.

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*, 1–14.

Keil, F. C. 2006. Explanation and understanding. *Annu. Rev. Psychol.*, 57: 227–254.

Kulesza, T.; Burnett, M.; Wong, W.-K.; and Stumpf, S. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*, 126–137.

Kulesza, T.; Stumpf, S.; Burnett, M.; Yang, S.; Kwan, I.; and Wong, W.-K. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on visual languages and human centric computing*, 3–10. IEEE.

Lakkaraju, H.; and Bastani, O. 2020. " How do I fool you?" Manipulating User Trust via Misleading Black Box Explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 79–85.

Liao, Q. V.; Gruen, D.; and Miller, S. 2020. Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–15.

Liao, Q. V.; and Varshney, K. R. 2021. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. *arXiv preprint arXiv:2110.10790*.

Lipton, Z. C. 2018. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3): 31–57.

Lombrozo, T. 2006. The structure and function of explanations. *Trends in cognitive sciences*, 10(10): 464–470.

Lombrozo, T. 2012. Explanation and Abductive Inference. *The Oxford Handbook of Thinking and Reasoning*, 260.

Lubars, B.; and Tan, C. 2019. Ask not what AI can do, but what AI should do: Towards a framework of task delegability. *arXiv preprint arXiv:1902.03245*.

Lucic, A.; Haned, H.; and de Rijke, M. 2020. Why does my model fail? contrastive local explanations for retail forecasting. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 90–98.

Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267: 1–38.

Miller, T.; Howe, P.; and Sonenberg, L. 2017. Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. *arXiv preprint arXiv:1712.00547*.

Mohseni, S.; Zarei, N.; and Ragan, E. D. 2018. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *arXiv preprint arXiv:1811.11839*.

Molnar, C. 2020. *Interpretable machine learning*. Lulu. com.

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. *arXiv preprint arXiv:1902.01876*.

Murdoch, W. J.; Singh, C.; Kumbier, K.; Abbasi-Asl, R.; and Yu, B. 2019. Interpretable machine learning: definitions, methods, and applications. *arXiv preprint arXiv:1901.04592*.

Narkar, S.; Zhang, Y.; Liao, Q. V.; Wang, D.; and Weisz, J. D. 2021. Model LineUpper: Supporting Interactive Model Comparison at Multiple Levels for AutoML. In *26th International Conference on Intelligent User Interfaces*, 170–174.

Owuor, C. O. 2001. *Implications of using Likert data in multiple regression analysis*. Ph.D. thesis, University of British Columbia.

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

Ras, G.; van Gerven, M.; and Haselager, P. 2018. Explanation methods in deep learning: Users, values, concerns and challenges. In *Explainable and interpretable models in computer vision and machine learning*, 19–36. Springer.

Samek, W.; Montavon, G.; Vedaldi, A.; Hansen, L. K.; and Müller, K.-R. 2019. *Explainable AI: interpreting, explaining and visualizing deep learning*, volume 11700. Springer Nature.

Schneider, J.; and Handali, J. 2019. Personalized explanation in machine learning: A conceptualization. *arXiv preprint arXiv:1901.00770*.

Sokol, K.; and Flach, P. 2020. Explainability fact sheets: a framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 56–67.

Suresh, H.; Gomez, S. R.; Nam, K. K.; and Satyanarayan, A. 2021. Beyond Expertise and Roles: A Framework to Characterize the Stakeholders of Interpretable Machine Learning and their Needs. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16.

Vasilyeva, N.; Wilkenfeld, D. A.; and Lombrozo, T. 2015. Goals Affect the Perceived Quality of Explanations. In *CogSci*.

Wang, X.; and Yin, M. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *26th International Conference on Intelligent User Interfaces*, 318–328.

Weld, D. S.; and Bansal, G. 2019. The challenge of crafting intelligible intelligence. *Communications of the ACM*, 62(6): 70–79.

Xie, Y.; Chen, M.; Kao, D.; Gao, G.; and Chen, X. 2020. CheXplain: Enabling Physicians to Explore and Understand Data-Driven, AI-Enabled Medical Imaging Analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13.

Yeh, C.-K.; Hsieh, C.-Y.; Suggala, A.; Inouye, D. I.; and Ravikumar, P. K. 2019. On the (in) fidelity and sensitivity of explanations. *Advances in Neural Information Processing Systems*, 32: 10967–10978.

Zhang, Y.; Liao, Q. V.; and Bellamy, R. K. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 295–305.