Measuring Explanation Quality – a Path Forward

Prof. Nava Tintarev

Department of Advanced Computing Sciences, Maastricht University, the Netherlands
ORCID (Prof. Nava Tintarev): https://orcid.org/0000-0003-1663-1627

Abstract. In this paper, I describe lessons learned in nearly 20 years of evaluating explanation interfaces with people. In my work, I have observed that tailoring the explanation to the user (e.g., a domain expert or a layperson) and task (e.g., decision support or model improvement), or context (e.g., under time pressure) is necessary for meaningful assessment of explanation quality (e.g., correct decisions or better understanding). Learning from trends in empirical research methods (natural language processing, information retrieval, and machine learning), I further argue that this is an issue for both human-computer interaction and machine learning.

Real-world factors influencing the performance of systems are at best implicitly encoded when inferring probabilities from observations, and at worst, no longer applicable to the settings for which they are used. In seeking a balanced approach between this reality and a pragmatic, data-driven and (by necessity) reductionist approach, I make constructive suggestions for evaluating the quality of explainable artificial intelligence.

1 Introduction

Explanations in artificial intelligence focus on making AI systems' decision-making processes understandable to humans.[1] However, there is a lot of diversity the individuals receiving the explanations and the AI decision-making processes [36]. For this reason, the evaluation of explanations is different from many other machine learning tasks: the definition for the "quality" of explanations depends on the user, task, and context – see also Table 1 for examples [47].

More critically, I argue that experimental design, often implicitly, encodes decisions about which information to show people and how to present that information. These decisions lead to variability in experimental design. This variability could be seen as a lack of rigor. However, it may be more accurate to see it as an implicit encoding of factors necessary for explanations to be useful. In practice, these are often unclear or unintentionally vague.

I further argue that the importance of considering these sources of variability is not inherent to human-computer interaction. Rather, it is also an *inherent challenge in inductive learning* and endemic to machine learning (especially in the face of context and domain shifts). Restated more bluntly: if you want your explanations to be useful, you do not get rid of these evaluation challenges in functional (offline metric-based) evaluations of explanations.

To propose a constructive way forward, I first list some reasons why we have taken more reductionist approaches to XAI in the past. Motivated by examples from previous studies, I argue for more careful reporting of not only experimental designs but also of intended context, users, and tasks. In understanding the sources and magnitudes of variance in our measurements of explanation quality, we can move closer to understanding whether the "improvements" we see are meaningful (and *when* they are so). By identifying these gaps, we *can focus our studies of explanation quality on the aspects that truly move our understanding further.*

2 Why do we need explanations?

Before we can gauge whether an explanation is useful, we have to have a sense of its purpose. "Interpretability" has by some been qualified as the degree to which a human can understand the cause of a decision [36].[2]

However, understanding can happen on multiple levels and is rarely an end-goal in itself. Some papers on explainable artificial intelligence (XAI) are motivated by the claim that explanations can improve user trust. While this has been successfully demonstrated in some cases (e.g., [24]), it is certainly not a given criterion for all tasks and contexts [7]. At a minimum, a myopic focus on trust is misleading. In some works, end-users over-rely on system outputs and explanations [14]; while in other systems, people use the systems despite relatively low trust [33].

In fact, there is no universal consensus on what constitutes a good explanation or explanation task, or even an *agreed-upon* taxonomy of explanatory tasks. This is despite XAI being a highly active research area. 'Transparency', 'Explainability', and 'Interpretability' are active topics of research at the main conference of both ECAI and IJCAI. To name a few further notable examples, there is also a dedicated international conference on XAI,[3] as well as Dagstuhl seminars,[4] and several survey papers focused specifically on XAI evaluation [26, 38, 58]. This lack of consensus (at least in part) reflects the diversity of explanatory goals, e.g., system debugging versus supporting decisions, and different user characteristics (strong visual working memory or less so) or even roles (lay-users versus experts). Furthermore, some of my early work found that the same explanations can do well on one explanatory goal, but poorly on another [51]. In several experimental setups, I repeatedly found that personalized explanations of recommended items could be very satisfying to participants, while not aiding or occasionally harming decision support. In other work, we studied a very different notion of quality. We measured the impact that explanations can have in mitigating the confirmation bias that people have when making decisions such as which movies to see next [53], or which articles to read on a disputed topic [11]. So

[1] Typically, the term explainable AI is used. However, the methods are more often based on inductive reasoning on large amounts of data, that is, supervised machine learning.

[2] A further discussion of notions of understandability and measurements thereof can be found in Section 4.2

[3] https://xaiworldconference.com/2025/

[4] https://www.dagstuhl.de/25142; https://www.dagstuhl.de/25272

Factor	Description	Examples
User	Who you are explaining to	Expertise, personality
Task	What the explanation is doing	Debugging, decision support
Context	The setting in which they are using the explanation	Time pressure, cost, reliability of advisor

Table 1. Factors that influence what counts as a good explanation and therefore also evaluation methodology.

while the notion of quality is a familiar one from other contexts (decision support, or accuracy in decisions), the notion of what counts as accurate (different from what they usually choose), and even what counts as a suitable explanation (contextualizing their behavior relative to other users rather than promoting a specific decision), is different.

Finally, while explanations can be useful, it is important at the same time to acknowledge that they may not *always* be needed. If user and system decisions align, this may cause unnecessary overhead. However, if the system has information the user does not have, explanations can help the user identify and understand these gaps in information. In turn, this can support better performance in line with a vision of ultra-strong AI as defined by Michie [35] as early as 1988: The AI is required to teach a (learned) symbolic hypothesis to a human, whose performance is consequently increased to a level beyond that of the human studying the training data alone (c.f., the approaches of [10, 12]).

3 Who are you explaining to (User)

> The quality of an explanation is determined by the fit to the recipient(s). Ensure as much as possible that the measurement of quality is done by the intended target group. Report on the efforts to align these, even when they are not successful.

In the previous section, I indicated two possible reasons for explaining: debugging and decision support. Debugging a system assumes the user has a certain level of technical expertise and places a greater emphasis on transparency and fidelity. In contrast, a decision-maker can assume that the minimal requirement for these is met, but needs to be able to distinguish between correct and incorrect decisions effectively. Illustratively, Liao et al. [34] provides more examples of XAI usage contexts: model improvement, capability assessment, adapting control, domain learning, and model auditing.

In the same way, every (implicit or explicit) decision we make in the user-centered evaluation of explanations influences the conclusions we can draw about explanation quality. In this section, I discuss how the implicit user influences participant selection and assumed individual characteristics.

3.1 Participant recruitment.

The selection of study participants can influence the outcome of the evaluation, and the intended and actual users should be reported to indicate the reliability or generalizability of findings [5]. A similar recommendation to report participants more precisely was also made in a meta-analysis of the evaluation of explanations in recommender systems [55]. Ideally, participants are representative of the target group (e.g., a system tester who is doing the debugging), but these are not always accessible to the experimenter. Or there are only a small number of available representative participants – as often is the case with experts.

Crowdsourcing. Pragmatically, many researchers turn to platforms that can be used to reach a larger number of participants, such as Prolific and Amazon Mechanical Turk. These platforms sometimes support stratification of participants (e.g., by location or stance on a given topic), but these are largely reported by the platform users themselves and rarely verified. For example, in a study where we balanced participant selection on their previously reported position on the topic of legalization of abortion, our resulting sample (in subsequent questions) still contained many more participants who supported the legalization of abortion [17]. Crowdsourcing platforms typically also have a limited demographic distribution, e.g., under-representation in the global south.

One user or more. In our recent work on group recommender systems, we have not only explained to a single user, but to several. In these cases, both the composition of the group regarding preferences, and who in the group is being asked influences the assessment of decision quality (albeit not directly the explanations) [4]. To generalize further, explanations might be presented to pairs of users or even entire teams, e.g., a range of medical specialists.

We are currently working with colleagues to investigate the explanatory needs of different stakeholder types in human resources (recruiter and job seeker) [43] and advertising (viewer and advertiser) [59].

Differences between sets of users. One way to study the influence of the participants on results is to repeat the same study (replication or A/A test) but with different participants. Recently, attempts of this kind of experimental replication have been conducted in the field of natural language generation – a subfield of computational linguistics and artificial intelligence that focuses on creating written or spoken language from structured or unstructured data. In this spirit, a recent joint task, ReproGen,[5] focused on the repeatability of human evaluations of automatically generated text (not per se explanations). While small-scale, preliminary results indicate an effect of different participant cohorts. Participants evaluated the same text on the same criteria (e.g., Fluency), and on the same scales, but these still had a large difference in normalized scores between the original study and its respective replication [6].

3.2 Individual characteristics

The work with my team over the years has highlighted the importance of considering individual user characteristics (e.g., personality, expertise, cognitive orientation, working memory, prior beliefs, or privacy concern), especially when

[5] Part of the ReproHum project: https://reprohum.github.io/

they are known to be relevant or influential in the task or domain. For example, we studied the effects of personalizing an interactive graphical interface for music recommendations [29]. We found a benefit for personalizing both (explanatory) visualization and interactive elements to a user characteristic related to expertise in music (musical sophistication). In another project we studied which visual presentational choice works best for presenting complex plans, and found that visual working memory was a crucial moderating factor for decision-support in a dual-task paradigm [52].

Most relevant perhaps are the findings that suggest that individual characteristics such as propensity for trusting (automated) technology, decision-making styles (e.g., intuitive vs rational) and cognitive orientations (such as need for cognition), may influence which format of explanation is most effective [23]. In our team, we have also seen that prior beliefs strongly influence confirmation bias [11], which in turn affects how people interpret explanations.

There is a wide range of ways of representing expertise in the literature. For example, one may distinguish between *expertise*, which reflects the stakeholder's level of familiarity with AI or the domain-specific knowledge, and *role*, which describes their functional involvement with the AI system [50, 18]. Some studies have found effects of individual factors on comprehension, including experience with XAI studies and level or type of education [57] .

More generally, a recent survey summarizes different user characteristics (demographics, personality, expertise) for 164 studies in explanations of recommender systems [55]. It also summarizes for which explanatory goal these explanations were evaluated. While the survey was focused on explanations in recommender systems specifically, this still supports the argument for a need to consistently report and measure the effect of individual characteristics on the efficacy of XAI.

4 What is the explanation doing? (Task)

- The task is designed to support certain kinds of quality measurements. This directly influences decisions about what stimuli are chosen and presented.
- To understand the influence of order on results, study designs should specify how each category of stimuli has been allocated to participants.
- Of particular interest are explanations for predictions of low confidence or instances for which we cannot generate explanations.

The task of the explanation, from a human-centered perspective, is to support a given task such as debugging or decision support. Beyond describing how models map inputs to outputs, explanations should be designed and evaluated with that specific end in mind.

Explanation quality is then measured in terms of (joint human-AI) performance on that task. Each study and task design encodes several decisions about what a useful explanation is. In the following sections, I make these design decisions and their relationship to the notion(s) of usefulness more explicit. This is not meant as an exhaustive guide for evaluation design, but rather to illustrate the relationship between study design and the conclusions that can be drawn about the quality of an explanation.

4.1 Study Design

In the ideal experiment, these decisions are shaped by the setting for which the explanations are intended (see also Task Design in Section 4.2).

Baselines. The selected baseline(s) should perform well for the task, and helps assess the added value of a new explanation method. If the baselines are poor, this sets the evaluated explanation at an unfair advantage [3] – it may be better than a terrible explanation, but it may still not be good. A common user-centered baseline is the AI alone, compared against the performance of the AI (with explanations) and the human together. This answers the question whether an explanation can be helpful, but does not address the question if another explanation might have been more helpful. For user studies, an additional motivation for having a relevant baseline is to control Social Desirability Bias effects (e.g., telling the experimenter that a system (or its explanations) is good to please them), and another to help the experimenter to make sense of quantitative data (e.g., is a score of 3.5/5 "good"?).

How to present versions. As soon as there are more types of explanations, an experimenter also needs to decide which versions of explanations each participant sees. These choices indirectly influence the conclusions that can be drawn about explanation quality. For example, will these explanations be shown between-subjects (different people see different versions), within-subjects (same people see the versions), or mixed designs (some versions are between- and others are within-subjects). Are these sequentially shown (first version 1 and then version 2) or concurrently (versions 1 and 2 are shown together, and the user indicates a preference)? Among other considerations, a within-subjects design requires fewer participants, but may cause issues related to fatigue effects, as well as learning/order effects (see also Ordering below).

Within or between subjects. Whether an experiment is set up as within- or between-subjects design can more crucially influence the evaluation of explanation quality as well. We conducted a replication study where each evaluator only scored individual explanations rather than multiple (between-subjects). In contrast to earlier work (within-subjects), we found no added value of explanations [4]. While other factors (such as the participant cohort or other small changes to our experimental design) could have influenced our results, the difference in experimental design is likely a contributing factor to the difference in outcome. By showing multiple explanations to participants, they are likely to have compared them to each other.

In some contexts, a comparison by human participants is desirable. Not only because it requires fewer participants, but also because it may be easier for a person to compare two options than to rate a single one. However, a comparison in evaluating explanations is rarely explicitly stated in relation to the explanability task. In my work, I eventually learned that experiments that collect repeated measures (e.g., rating the explanation multiple times) for each participant allow us to run tests that statistically control for differences between individuals (including ones we did not model explicitly).

Ordering. With more stimuli, the order of explanations influences the score(s) each explanation receives. Explanations that participants see earlier influence explanations that participants see later. An exploratory analysis of earlier works also indicates that people learn different strategies of using AI assistance depending on what AI assistance they saw previously, indicating that we may need to take this into account when designing adaptive AI assistance [49]. Even though explanations are not shown side-by-side, participants are likely to make comparisons.

A common solution is randomization of participants to conditions, or of task order in repeated measures. Exceptional are tasks where explanations are designed to improve learning, and might increase in complexity over time. Or if the end task for other reasons follows steps in a certain order (e.g., following a process for a diagnosis).

4.2 Task Design

Recent work on human-centered XAI explicitly distinguishes between one-stage (AI and human decision concurrent), two-stage (user decides and updates after receiving more information), and on-demand (where the user can explicitly request the explanations) [12].

Others have introduced the notion of Frictional AI and pro-hoc explanations, intended as cognitive interventions or to boost users' motivations to engage analytically with AI assistance [10]. Miller et al also make a case for presenting possible explanations for competing decision outcomes, rather than promoting a single decision or outcome [1]. This *hypothesis-driven decision support using evaluative AI* also makes the distinction between XAI as a decision-making tool rather than a persuasive explanation mechanism. These interventions all apply two-stage evaluation methodologies.

Longer interactions can be found in the recommender systems domain, where users can see the connection between inputs, parameters, and outputs (e.g., [29]). With this, we also see a methodological development for evaluation over repeated interactions, c.f., [31, 54].

The majority of user studies in XAI are set up as a one-shot exposure to explanations. In some cases, these are a series of decisions (in a within-subjects design)– but typically, each decision is associated with precisely one explanation, and no revision is supported. In a scenario where multiple decisions are taken by an individual (e.g., detecting fractures in X-rays for different patients) repeated evaluations are congruent with the intended task (high ecological validity).

Finally, the user may have new information that could change the prediction [46]. In that case, the interaction may allow the user to supply this new information.

Metrics. Task design and metrics are closely linked. The task should support certain kinds of measurements. Studies measuring the same constructs can use different questions or even scales. I survey here some possible measurements for decision-making and understandability.

Decision-making Multi-stage experimental designs are particularly well-suited for evaluating the notion of super-strong AI, i.e., whether human decision-making is enhanced with the addition of AI (in comparison with human-only).

Previous experiments measured the correctness of decisions. They also measured change in preference for recommended items [8] [51]. Participants saw the explanation for the recommended item, to which they gave a score. Then they learned more about the item (to simulate the actual decision or consumption of the item), and gave a new score for the item. The idea was that a good explanation would result in a small change in score – that is, the user should be able to estimate the actual preference or score for the item well with only the explanation.

This evaluation could measure not only the final outcome, but the process leading to the outcome. A solid framework for understanding how users make choices and how to best support them is further supplied in Jameson et al. [28].

Understandability. Understandability, on the other hand, is more relevant for other usage contexts such as model improvement and model auditing. These usage contexts, in turn, define the type of tasks for which performance is measured. Understandability is also a notoriously difficult in XAI: even data scientists have been found to misinterpret the explanations given by off-the-shelf tools [32]. This measurement can be both task-based (more objective) or based on user perceptions (more subjective) [56]. The former allows for a quantitative approach that employs a questionnaire consisting of a collection of curated questions about a predictive system for a given task that are aligned with a selected definition of comprehension. E.g., whether users can derive or simulate the model's output and identify feature influence [42, 56]. Earlier work often made use of the subjective understanding, employing post-survey questionnaires which ask users whether an explanation is understandable and fulfills their needs, a.o., [21, 40]. In measuring understandability, it is crucial to calibrate task difficulty to avoid floor and ceiling effects (all explanations are highly understandable or poorly understandable). In addition, it is valuable to report accuracy per task type rather than on aggregate.

Stimuli selection. The choice of task also indirectly informs stimuli or instance selection (e.g., which images, text, or items to recommend and explain). This is particularly relevant for local instance-based explanations. For example, experiments that include difficult instances are suitable for situations where low confidence is likely to occur [13]. Studying both correct and incorrect instances is necessary for studying over-reliance [14].

For global explanations, different considerations arise regarding *coverage, confidence, and reliability*. For example, in previous work, we found that not all instances could be recommended and therefore also could not be explained [4]. Explanation (and prediction) coverage is helpful information for indicating the boundaries of experiment content validity (we may be measuring the right construct, but some aspects are missing from the measurement).

There is also a growing body of research that studies explanations that are (intentionally or accidentally) misleading. Recent work has studied whether participants draw conclusions that are not specified or possible to deduce in the explanations [57]. An additional possible complication is that certain predictions have low confidence. In such cases, it is important to convey this low confidence or choose not to explain at all [14]. NIST[6] proposed a Knowledge limits principle, which indicates that a system only operates under the conditions it was designed for and when it reaches a sufficient confidence in its output or actions [39]. While there is understandable resistance to conveying uncertainty, recent work in visualizations argues that uncertainty communication necessarily reduces degrees of freedom in viewers' statistical inferences [27]. The need to convey confidence is only likely to grow with the increasing use of generative AI.

4.3 In which setting are they using the explanation? (Context)

> The context influences how explanations are processed and which users and tasks are (most) relevant to study. It is valuable to study and report on context systematically.

Furthermore, a user with the same characteristics and the same task, may still need a different explanation in new contexts. From a data or offline evaluation perspective, the analogy here is one of domain shift, and specifically, covariate shift. In machine learning, we also see examples of label shift and conditional shift. Said differently, machine learning models make assumptions on the distribution of input, output, and key context variables. By making these assumptions explicit, we are better informed when making interpretations or identifying when these assumptions are violated.

For example, certain contextual factors influence over- and under-reliance. E.g., Sutherland et al. [48] studied effects of: the cost of receiving advice (time required for the adviser to give advice), the reliability of the adviser (% of time correct), and the predictability of the environment. This means that the relationship between explanation quality is moderated by these contextual factors. We typically pick a specific context for our experiments, which may or may not be relevant for future scenarios.

In addition, user, task, and context are often interdependent. E.g., task difficulty can be considered a contextual variable, which is also mediated by the domain. Domain also influences the perceived stakes, where some domains are higher stakes than others. The way cost is defined in the experiment will likely influence the explanation quality. A third example can be found in explanations for multiple users. This raises contextual issues, examples include intergroup dynamics and privacy concerns [37].

5 Metric-based functional evaluation

In the following sections I discuss why this problem is not unique to human-computer interaction approaches to XAI, and is (also) in fact endemic to inductive (data-driven) machine learning. I move on to identify other issues that are inherent to the evaluation of explanations, regardless of the choice to conduct a user-centered or offline metric-based evaluation.

> Switching to offline metric-based evaluation may appear to resolve some issues, but these are latent rather than absent. If changing the user, task, or context 'changes' explanation quality by 10%, it may not be meaningful to report a 2-3% performance improvement that does not control for these variables.

Users are hard. Let's do offline evaluation. A reader may, at this point, conclude that they are grateful to avoid the messy complexity of user-centered evaluation and return to offline metric-based evaluations. Aside from the fact that these methods are not fit for measuring the impact on people, *analogies for the complexities listed for user evaluation also exist in offline evaluation*. The user, task, and context are assumed to be (implicitly) represented in the data. More often than not, it is not (yet) explicitly encoded.

Big enough an improvement? One question is what magnitude of improvement should be considered meaningful. Below, I highlight findings in different flavors of machine learning and AI that indicate that performance improvements are not as meaningful as they seem, often incorrectly attributing the improvement to a new predictive model.

Already in 2009, a paper in information retrieval put into question the conclusions of a decade of results [3]. More recently, at ICML 2024, a position paper states that "a common but incomplete understanding of empirical research in machine learning that leads to non-replicable results, makes findings unreliable, and threatens to undermine progress in the field" [25]. Analogously, (in the relatively younger field of) recommender systems, a 2023 paper questioned the source of reported improvements in performance [44]. To highlight that this problem exists also for unsupervised learning, similar challenges have been identified for reinforcement learning (c.f., [22, 30]). *There is no reason why we would be immune to such a replication crisis within explainable artificial intelligence.*

Solving for the right task. Taken as a thought experiment, if changing the intended user or task changes the performance on offline metrics more than changing the predictive models, then we may be optimizing on factors that are not fit for purpose. For example, our recent results in video summarization (computer vision) suggest that metric-based evaluation fails in its default form to adhere to fundamental qualities of a good video summary – a sensitivity to order in the summary. In fact, summaries with perturbed temporal coherence still perform very well on offline metrics (F1-score alignment with a human-generated gold standard) [19]. Arguably, the human scores of scene relevance could be perfectly correct. However, they are unsuitable or limited when performing this task: summary generation. To the best of my knowledge, the suitability for user and task has not yet been systematically tested in this way. However, understanding how the change of user, task, or context influences the performance can give us a sense of which orders of magnitude are meaningful to aim for.

The field of Explainable Artificial Intelligence would do well to learn from these findings in the wider machine learning community, just as computer science has been learning from the reproducibility crises in medicine and psychology.

[6] National Institute of Standards and Technology

The calls to attention listed above should be taken positively. *They are not a damnation of the approaches or subdisciplines* I selected. Rather, these papers should be taken in the context of an evolving and positive development of reflection in machine learning. Even stronger, the increase in empirical studies and machine learning challenges is a *necessary condition for asking questions about how the field is evolving*. Other communities of machine learning, including LLMs, and evaluation of explanations, do not yet have the same richness of empirical and experimental data, but are will be facing this as a dormant issue.

5.1 Common issues

There are a number of issues that affect the evaluation of explanation more broadly (both when explanations are evaluated using offline metrics and in user studies).

Ground truth. Most XAI methods rely on a ground truth, and this includes most of the feature-based and saliency-based methods that are currently dominant in the literature. For user studies, this bottleneck appears as limited access to a sufficient number of representative and reliable users (to generate or label the ground truth). For most offline evaluations measuring some notion of fidelity (either completeness or soundness), the bottleneck is whether the data is sufficiently representative for the current situation. Implicit decisions encoded in a user study, can equally be encoded in a machine learning task. Possible information to encode could be: Who were the participants who wrote or assessed the ground truth explanations? If these were automatically generated, which indications do we have that they *still* represent the intended target users? Or, if explanations are given as a sequence or in combination, has this also been considered in the training of the predictive model?

There are however, some XAI evaluation methods that are not reliant on a ground truth [38]. For example, AXE measures how well a given explanation can help emulate model behavior [41]. Other exceptions use instance-based diversity metrics, e.g., how diverse are the examples supplied to justify this explanation?

Given that access to relevant ground truth is a common issue, some researchers are experimenting with large-language models as annotators (LLM-as-judge paradigm) [2, 15, 16]. These methods themselves have historical biases and are prone to context shifts. Furthermore, they suffer from issues of errors, ambiguity, and too much homogeneity (model collapse). However, they do have the advantage that they can be prompt-tuned to both users and tasks and tools are being designed to support joint annotations.[7] LLMs work better in resource-rich rather than scarce domains; however at the time of writing, no conclusive advice can be given about when these can be used reliably.

Generating explanations. Other challenges lie not in the evaluation protocol itself but in the process of generating the explanations. *Assumptions* (or errors) on the machine learning model can be *propagated* to the explanations themselves. For example, explanations based on inductive inference are, *by design*, vulnerable to data shifts that influence the quality of explanations that can be generated. This is also why it can be hard for a human assessor to assess explanations – they may be assessing the veracity of an explanation that builds on an erroneous or simply an outdated predictive model. The explanations then receive a poor score, even if the predictive model is (primarily) at fault. *To achieve better decision making, an ideal explanation could be (selectively) decoupled (or systematically varied) from the predictive model.*

6 Moving Forward

Having reviewed the many challenges for XAI evaluation, the reader may be forgiven for asking how to move forward constructively. In a sense, these problems are recognized in the broader fields of AI, machine learning, and human-computer interaction. In response, I offer concrete suggestions specifically for how XAI evaluation can move forward as a field.

> To improve the quality of explainable artificial intelligence, we would benefit from: **a)** systematic reporting of user, task, and context; **b)** an investment in reproducibility studies, and **c)** more meta-analysis of experiments.

Systematic reporting. There is a very good reason why we take a reductionist approach in our evaluations in machine learning. Collecting data is not only expensive, but every study is constrained in terms of the type and quantity that we can collect. This simplification is in fact necessary for empirical research. If we were to model for each specific user, task, and context, most data sets will have an insufficient number of data points with each intersection.

In practice, however, most applications do not need to consider all or very fine-grained intersections of these dimensions. Much like transfer learning, smaller studies allow us to identify the dimensions of interest, which can later be expanded to larger or more expensive experiments. Qualitative and design-driven methodologies are excellent for these more exploratory stages, allowing the quantitative methods to focus on confirmatory analysis instead.

Careful reporting, in contrast, allows for a better matching between empirical findings and usefulness without causing unnecessary cost or evaluation paralysis. Similiarly, if more data collection of a specific type (e.g., for a given task) is needed, it is more cost-effective to know precisely where the data is sparse. A helpful reference, focused on what to report in human evaluation of text can be found in *The Human Evaluation Datasheet* [45].

Meta-analysis. Broadly speaking, a meta-analysis can be defined as a systematic literature review supported by statistical methods where the goal is to aggregate and contrast the findings from several related studies [20]. Meta-review allows for conceptual comparison of how factors such as task or even dependent variables (e.g., explanation quality) are defined across papers.

At the moment, we cannot do statistical meta-analyses in XAI since we are largely comparing widely disparate studies – *we are comparing apples and oranges*. For a successful meta-analysis, rigor in reporting is therefore a prerequisite.

[7] E.g., https://huggingface.co/spaces/EvalAssist/EvalAssist

The outcome measures of these analyses differ, but one common notion (represented by several metrics) is "effect size" – indicating how much a factor (e.g., an explanation type) influences the outcome variable (explanation quality). Statistical meta-analyses can also be used to answer other questions, including moderator analysis across multiple studies, e.g., whether the effectiveness of the explanations depends on the characteristics of the user or task. Statistical packages for meta-analysis exist, such as METAFOR[8], and META-ANALYSIS[9], but are not commonly used in computer science.

Replicability. Another barrier has been difficulty in replicating studies. Part of the solution lies in good replication practices and FAIR practices, e.g., making code and data reusable (including sufficient documentation) and findable (using rich metadata), as well as pre-registration studies[10] and registered reports [9]. Additional requirements arise for generative AI, such as reporting on version, date of access, exact prompts used, as well as multiple runs and benchmarking on open datasets. The recent General-Purpose AI Code of Conduct[11] as well as BenchmarkCards[12] provide guidance on model (and data) transparency. Finally, despite limitations, leaderboards and reproducibility tracks have also been fertile for standardization and benchmarking, and have helped us collect richer empirical evidence for specific machine learning tasks. Explainable AI can learn from these developments.

7 Conclusion

In this paper, I made the argument that the assessment of explanation quality is dependent on task, user, and context. These are issues that are common to both from human-centered and metric-based evaluations of explanation quality. This means that the lack of conscious tailoring in data-driven (inductive reasoning) approaches is particularly problematic and hinders progress in the field. Fortunately, we are starting to have enough empirical studies to collectively improve our reporting and meta-analysis. In understanding the sources and magnitudes of variance in our empirical findings, we can move closer to an understanding of whether the "improvements" we see are meaningful. Better reporting can then serve as a basis for charting which tasks and contexts are less understood, and to better target our research efforts.

Acknowledgements

Many experiments described herein are joint efforts with early-career researchers. My heartfelt thanks to venues that still support unstructured thinking time, such as the Schloss Dagstuhl and Bielefeld, which have provided fertile ground for many of the ideas presented here. Thank you also to the critical friends such as Oana Inel, Harmanpreet Kaur, Cynthia Liem, Tim Miller, and Ehud Reiter for valuable discussions on the topics covered in this paper. I take the responsibility, but not (all) the credit.

[8] https://www.jstatsoft.org/article/view/v036i03

[9] https://cran.r-project.org/web/views/MetaAnalysis.html

[10] E.g., https://osf.io/

[11] Model Documentation Form: https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai

[12] https://github.com/SokolAnn/BenchmarkCards

References

[1] A. Alshehri, H. Alotaibi, T. Miller, and M. Vered. A hypothesis-driven approach to explainable goal recognition. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems*, pages 105–114, 2025.

[2] N. Arabzadeh and C. L. Clarke. Benchmarking LLM-based relevance judgment methods. *arXiv preprint arXiv:2504.12558*, 2025.

[3] T. G. Armstrong, A. Moffat, W. Webber, and J. Zobel. Improvements that don't add up: ad-hoc retrieval results since 1998. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 601–610, 2009.

[4] F. Barile, T. Draws, O. Inel, A. Rieger, S. Najafian, A. Ebrahimi Fard, R. Hada, and N. Tintarev. Evaluating explainable social choice-based aggregation strategies for group recommendation. *UMUAI*, pages 1–58, 2023.

[5] J. Beel, C. Breitinger, S. Langer, A. Lommatzsch, and B. Gipp. Towards reproducibility in recommender-systems research. *UMUAI*, 26(1):69–101, 2016.

[6] A. Belz, A. Shimorina, S. Agarwal, and E. Reiter. The reprogen shared task on reproducibility of human evaluations in nlg: Overview and results. In *INLG*, pages 249–258, 2021.

[7] M. Benk, S. Kerstan, F. von Wangenheim, and A. Ferrario. Twenty-four years of empirical research on trust in AI: a bibliometric review of trends, overlooked issues, and future directions. *AI & society*, pages 1–24, 2024.

[8] M. Bilgic and R. J. Mooney. Explaining recommendations: Satisfaction vs. promotion. In *Proceedings of the Wokshop Beyond Personalization, in conjunction with the International Conference on Intelligent User Interfaces*, pages 13–18, 2005.

[9] N. C. Brown, D. Weintrop, V. Ojha, and K. Isenegger. Registered reports and preregistration: A new way to conduct research. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 2*, pages 1252–1252, 2022.

[10] F. Cabitza, C. Natali, L. Famiglini, A. Campagner, V. Caccavella, and E. Gallazzi. Never tell me the odds: Investigating pro-hoc explanations in medical decision making. *Artificial intelligence in medicine*, 150:102819, 2024.

[11] F. Cau and N. Tintarev. Navigating the thin line: Examining user behavior in search to detect engagement and backfire effects. In *ECIR*, 2024.

[12] F. M. Cau and L. D. Spano. The influence of curiosity traits and on-demand explanations in ai-assisted decision-making. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pages 1440–1457, 2025.

[13] F. M. Cau, H. Hauptmann, L. D. Spano, and N. Tintarev. Supporting high-uncertainty decisions through AI and logic-style explanations. In *IUI*, 2023.

[14] F. M. Cau, H. Hauptmann, L. D. Spano, and N. Tintarev. Effects of ai and logic-style explanations on users' decisions under different levels of uncertainty. *ACM Trans. on Interactive Intelligent Systems*, 13(4):1–42, 2023.

[15] G. Chen, S. Chen, Z. Liu, F. Jiang, and B. Wang. Humans or llms as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327, 2024.

[16] C. L. Clarke and L. Dietz. LLM-based relevance assessment still can't replace human relevance assessment. *arXiv preprint arXiv:2412.17156*, 2024.

[17] T. Draws, J. Liu, and N. Tintarev. Helping users discover perspectives: Enhancing opinion mining with joint topic models. In *2020 International Conference on Data Mining Workshops (ICDMW)*, pages 23–30. IEEE, 2020.

[18] U. Ehsan, S. Passi, Q. V. Liao, L. Chan, I.-H. Lee, M. Muller, and M. O. Riedl. The who in XAI: how ai background shapes perceptions of ai explanations. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–32, 2024.

[19] A. Ganesh, M. Popa, D. Odijk, and N. Tintarev. Does spatio-temporal information benefit the video summarization task? In *AEQUITAS 2024: Workshop on Fairness and Bias in AI| co-located with ECAI 2024*, pages 10–25, 2024.

[20] G. V. Glass. Primary, secondary, and meta-analysis of research. *Educational researcher*, 5(10):3–8, 1976.

[21] S. Hadash, M. C. Willemsen, C. Snijders, and W. A. IJsselsteijn. Improving understandability of feature contributions in model-agnostic explainable ai tools. In *Proceedings of the*

2022 CHI Conference on Human Factors in Computing Systems, pages 1–9, 2022.

[22] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger. Deep reinforcement learning that matters. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[23] D. C. Hernandez-Bocanegra and J. Ziegler. Effects of interactivity and presentation on review-based explanations for recommendations. In *IFIP Conference on Human-Computer Interaction*, pages 597–618. Springer, 2021.

[24] D. C. Hernandez-Bocanegra and J. Ziegler. Explaining review-based recommendations: Effects of profile transparency, presentation style and user characteristics. *i-com*, 19(3):181–200, 2021.

[25] M. Herrmann, F. J. D. Lange, K. Eggensperger, G. Casalicchio, M. Wever, M. Feurer, D. Rügamer, E. Hüllermeier, A.-L. Boulesteix, and B. Bischl. Position: Why we must rethink empirical research in machine learning. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

[26] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman. Measures for explainable ai: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-ai performance. *Frontiers in Computer Science*, 5:1096257, 2023.

[27] J. Hullman. Why authors don't visualize uncertainty. *IEEE transactions on visualization and computer graphics*, 26(1): 130–139, 2019.

[28] A. Jameson, B. Berendt, S. Gabrielli, F. Cena, C. Gena, F. Vernero, K. Reinecke, et al. Choice architecture for human-computer interaction. *Foundations and Trends® in Human–Computer Interaction*, 7(1–2):1–235, 2014.

[29] Y. Jin, N. Tintarev, N. N. Htun, and K. Verbert. Effects of personal characteristics in control-oriented user interfaces for music recommender systems. *User Modeling and User-Adapted Interaction*, 30(2):199–249, 2020.

[30] S. M. Jordan, A. White, B. C. Da Silva, M. White, and P. S. Thomas. Position: benchmarking is limited in reinforcement learning research. In *Proceedings of the 41st International Conference on Machine Learning*, pages 22551–22569, 2024.

[31] P. K. Kahr, G. Rooks, M. C. Willemsen, and C. C. Snijders. Understanding trust and reliance development in AI advice: Assessing model accuracy, model explanations, and experiences from previous interactions. *ACM Transactions on Interactive Intelligent Systems*, 14(4):1–30, 2024.

[32] H. Kaur, H. Nori, S. Jenkins, R. Caruana, H. Wallach, and J. Wortman Vaughan. 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*, pages 1–14, 2020.

[33] S. Kreps, J. George, P. Lushenko, and A. Rao. Exploring the artificial intelligence "trust paradox": Evidence from a survey experiment in the united states. *Plos one*, 18(7):e0288109, 2023.

[34] Q. V. Liao, Y. Zhang, R. Luss, F. Doshi-Velez, and A. Dhurandhar. Connecting algorithmic research and usage contexts: a perspective of contextualized evaluation for explainable ai. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 10, pages 147–159, 2022.

[35] D. Michie. Machine learning in the next five years. In *Proceedings of the 3rd European conference on European working session on learning*, pages 107–122, 1988.

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

[37] S. Najafian, A. Delic, M. Tkalcic, and N. Tintarev. Factors influencing privacy concern for explanations of group recommendation. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, pages 14–23, 2021.

[38] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. Van Keulen, and C. Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. *ACM Computing Surveys*, 55(13s):1–42, 2023.

[39] P. J. Phillips, C. A. Hahn, P. C. Fontana, A. N. Yates, K. Greene, D. A. Broniatowski, and M. A. Przybocki. Four principles of explainable artificial intelligence. 2021.

[40] M. Radensky, D. Downey, K. Lo, Z. Popovic, and D. S. Weld. Exploring the role of local and global explanations in recommender systems. In *CHI extended abstracts*, pages 1–7, 2022.

[41] K. Rawal, Z. Fu, E. Delaney, and C. Russell. Evaluating model explanations without ground truth. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, page 3400–3411, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714825.

[42] Y. Rong, T. Leemann, T.-T. Nguyen, L. Fiedler, P. Qian, V. Unhelkar, T. Seidel, G. Kasneci, and E. Kasneci. Towards human-centered explainable AI: A survey of user studies for model explanations. *IEEE transactions on pattern analysis and machine intelligence*, 46(4):2104–2122, 2024.

[43] R. Schellingerhout, F. Barile, and N. Tintarev. A co-design study for multi-stakeholder job recommender system explanations. In *World Conference on Explainable Artificial Intelligence*, pages 597–620. Springer, 2023.

[44] F. Shehzad and D. Jannach. Everyone's a winner! on hyperparameter tuning of recommendation models. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 652–657, 2023.

[45] A. Shimorina and A. Belz. The human evaluation datasheet 1.0: A template for recording details of human evaluation experiments in NLP. *arXiv preprint arXiv:2103.09710*, 2021.

[46] A. Sivaprasad, E. Reiter, D. McLernon, N. Tintarev, S. Bhattacharya, and N. Oren. Patient-centred explainability in ivf outcome prediction. In *AIiH special session on AI for Maternity and Women's Health and Wellbeing*, 2025.

[47] K. Sokol and P. Flach. One explanation does not fit all: The promise of interactive explanations for machine learning transparency. *KI-Künstliche Intelligenz*, 34(2):235–250, 2020.

[48] S. C. Sutherland, C. Harteveld, and M. E. Young. Effects of the advisor and environment on requesting and complying with automated advice. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 6(4):1–36, 2016.

[49] S. Swaroop, Z. Buçinca, K. Z. Gajos, and F. Doshi-Velez. Personalising AI assistance based on overreliance rate in AI-assisted decision making. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pages 1107–1122, 2025.

[50] M. Szymanski, V. Vanden Abeele, and K. Verbert. Disentangling stakeholder role and expertise in user-centered explainable AI. In *ACM UMAP*, pages 32–39, 2025.

[51] N. Tintarev and J. Masthoff. Evaluating the effectiveness of explanations for recommender systems. *User Modeling and User-Adapted Interaction*, 22(4-5):399–439, 2012.

[52] N. Tintarev and J. Masthoff. Effects of individual differences in working memory on plan presentational choices. *Frontiers in psychology*, 7:1793, 2016.

[53] N. Tintarev, S. Rostami, and B. Smyth. Knowing the unknown: Visualising consumption blind-spots in recommender system. In *ACM Symposium On Applied Computing (SAC)*, 2018.

[54] N. Tintarev, B. P. Knijnenburg, and M. C. Willemsen. Measuring the benefit of increased transparency and control in news recommendation. *AI Magazine*, 45(2):212–226, 2024.

[55] K. Wardatzky, O. Inel, L. Rossetto, and A. Bernstein. Whom do explanations serve? a systematic literature survey of user characteristics in explainable recommender systems evaluation. *ACM Trans. on Recommender Systems*, 3(4):1–35, 2025.

[56] C. Waterschoot, R. Yera Toledo, N. Tintarev, and F. Barile. With friends like these, who needs explanations? evaluating user understanding of group recommendations. In *Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization (UMAP '25)*, New York, NY, USA, 2025. Association for Computing Machinery. doi: 10.1145/3699682.3728345.

[57] Y. Xuan, E. Small, K. Sokol, D. Hettiachchi, and M. Sanderson. Comprehension is a double-edged sword: Over-interpreting unspecified information in intelligible machine learning explanations. *International Journal of Human-Computer Studies*, 193:103376, 2025.

[58] J. Zhou, A. H. Gandomi, F. Chen, and A. Holzinger. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5):593, 2021.

[59] D. Zilbershtein, F. Barile, D. Odijk, and N. Tintarev. Bridging the transparency gap: Exploring multi-stakeholder preferences for targeted advertisement explanations. In *Joint Workshop on Interfaces and Human Decision Making for Recommender Systems*, 2024.