Report from Dagstuhl Seminar 25142

Explainability in Focus: Advancing Evaluation through Reusable Experiment Design

Simone Stumpf*[1], Stefano Teso*[2], and Elizabeth M. Daly*[3]

1 **University of Glasgow, UK.** simone.stumpf@glasgow.ac.uk
2 **University of Trento, IT.** stefano.teso@unitn.it
3 **IBM Research, IE.** elizabeth.daly@ie.ibm.com

Abstract

This report summarizes the outcomes of Dagstuhl Seminar 25142, which convened leading researchers and practitioners to address the pressing challenges in evaluating explainable artificial intelligence (XAI). The seminar focused on developing reusable experimental designs and robust evaluation frameworks that balance technical rigor with human-centered considerations. Key themes included the need for standardized metrics, the contextual relevance of evaluation criteria, and the integration of human understanding, trust, and reliance into assessment methodologies. Through a series of talks, collaborative discussions, and case studies across domains such as healthcare, hiring, and decision support, the seminar identified critical gaps in current XAI evaluation practices and proposed actionable strategies to bridge them. The report presents a refined taxonomy of evaluation criteria, practical guidance for experimental design, and a roadmap for future interdisciplinary collaboration in responsible and transparent AI development.

Seminar March 30 – April 2, 2025 – https://www.dagstuhl.de/25142
2012 ACM Subject Classification Computing methodologies → Artificial intelligence; Human-centered computing → Human computer interaction (HCI); Computing methodologies → Machine learning
Keywords and phrases Explainability, Mental Models, interactive machine learning, Experiment Design, Human-centered AI Dagstuhl Seminar
Digital Object Identifier 10.4230/DagRep.15.3.201

1 Executive Summary

Simone Stumpf (University of Glasgow, simone.stumpf@glasgow.ac.uk)
Stefano Teso (University of Trento, Italy, stefano.teso@unitn.it)
Elizabeth M. Daly (IBM Research, Ireland, elizabeth.daly@ie.ibm.com)

This summary outlines the key outcomes of Dagstuhl Seminar 25142, which focused on the role of explanations in advancing Responsible and Ethical AI. The discussion emphasized the importance of explainability in AI systems to:

- **Demystify AI systems:** Helping users understand the rationale behind AI-generated outcomes.
- **Promote accountability:** Enabling users to verify that decisions are based on valid, unbiased data.

* Editor / Organizer

Explainability in Focus: Advancing Evaluation through Reusable Experiment Design, *Dagstuhl Reports*, Vol. 15, Issue 3, pp. 201–224
Editors: Elizabeth M. Daly, Simone Stumpf, and Stefano Teso
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

- **Encourage transparency:** Reinforcing trust and confidence in AI technologies through clear, interpretable outputs.
- **Support debugging and decision-making:** Assisting users in evaluating whether to trust a prediction or recommendation.

This seminar brought together researchers, practitioners, and experts in the field of explainable AI to collaboratively develop reusable resources aimed at standardizing the evaluation of explainability methods. The goal was to ensure that evaluation practices are robust, consistent, and adaptable across diverse contexts and applications.

A major outcome of the seminar was the identification of three key challenges:

1. Balancing technical rigor with human-centric considerations when determining which aspects of explanations should be assessed;
2. Developing consistent and reliable metrics for evaluating the selected criteria; and
3. Ensuring that both criteria and measurements are appropriately tailored to specific use cases where explainability is critical.

To illustrate practical applications of the discussed frameworks, we presented case studies showcasing end-to-end evaluation examples.

2 Table of Contents

3 Overview of Talks

3.1 Why are so many studies measuring XAI wrong?

Simone Stumpf (University of Glasgow – UK, simone.stumpf@glasgow.ac.uk)

Reflecting on the original vision of XAI in 2016, three points were important:
1. providing an *explanation* of a "decision" and the "reasoning" behind it;
2. to increase *understanding* or knowledge of the AI;
3. it should be useful to a *"user"* who doesn't really have AI knowledge to result in *appropriate trust*

There are nowadays lots of *"technical"* ways to measure XAI explanations (e.g. complexity, fidelity, consistency, etc) but these are described in different terms and measured in different ways, making their consistent application problematic.

Most importantly, many XAI studies are *never evaluated with humans*. We lack consistent human-centered XAI measures, possibly both subjective and objective measurements revolving around:
- Understandability and preferences
- Understanding
- satisfaction
- trust and reliance
- other effects of explanations (e.g. actionability, model improvements, etc)

3.2 Replication in explainable AI: a case study in group recommender systems

Nava Tintarev (Maastricht University, NL, n.tintarev@maastrichtuniversity.nl)

Joint work of Cedric Waterschoot, Raciel Yera Toledo, Francesco Barile, Nava Tintarev

We have few instances of reproduction or replication studies in XAI. I discuss a series of replication studies using group recommender systems as an application area [1]. I highlight several design considerations including the choice of baseline, experimental procedure (within or between subjects; internal vs external evaluator), and task complexity[1]. I conclude with a brief introduction of a recent study evaluating objective (task performance) and subjective (perceived) understanding of explanations in group recommender systems.[1]

This work was led by Francesco Barile.

References

1 F. Barile, T. Draws, and O. et al. Inel. Evaluating explainable social choice-based aggregation strategies for group recommendation. *User Model User-Adap Inter*, 34:1–58, 2024.

[1] To appear UMAP'25 [2].

2 Cedric Waterschoot, Raciel Yera Toledo, Francesco Barile, and Nava Tintarev. With friends like these, who needs explanations? evaluating user understanding of group recommendations. In *UMAP (to appear)*, 2025.

3.3 A Review of Taxonomies of XAI Evaluation Methods

Timo Speith (University of Bayreuth – Bayreuth, Germany, timo.speith@uni-bayreuth.de)

The evaluation of explainable AI (XAI) systems remains a fragmented field, with diverse metrics and taxonomies across the literature. In this talk, I present preliminary insights from a systematic literature review of XAI evaluation methods taxonomies. Across 160 publications, I found nearly 250 properties that were proposed to evaluate XAI systems. Taxonomic efforts often center around the type of evaluation used (human-based vs. mathematical), yet newer approaches emphasize process-oriented perspectives. I highlight the challenges posed by terminological inconsistencies–such as synonymous or overlapping terms and conceptual ambiguities–and propose that evaluating explainability should better attend to the objects of measurement (e.g., understanding of explanation vs. understanding of output vs. understanding of model). This talk aims to contribute to the development of reusable experimental designs by advocating for more coherent evaluation frameworks.

3.4 Doing multiclass Shapley values properly

Peter Flach (University of Bristol, UK)

Joint work of Paul-Gauthier Noé, Miquel Perelló-Nieto, Jean-François Bonastre, Peter A. Flach
Main reference Paul-Gauthier Noé, Miquel Perelló-Nieto, Jean-François Bonastre, Peter A. Flach: "Explaining a Probabilistic Prediction on the Simplex with Shapley Compositions", in Proc. of the ECAI 2024 – 27th European Conference on Artificial Intelligence, 19-24 October 2024, Santiago de Compostela, Spain – Including 13th Conference on Prestigious Applications of Intelligent Systems (PAIS 2024), Frontiers in Artificial Intelligence and Applications, Vol. 392, pp. 1124–1131, IOS Press, 2024.
URL https://doi.org/10.3233/FAIA240605

Originating in game theory, Shapley values are widely used for explaining a machine learning model's prediction by quantifying the contribution of each feature's value to the prediction. This requires a scalar prediction as in binary classification, whereas a multiclass probabilistic prediction is a discrete probability distribution, living on a multidimensional simplex. In such a multiclass setting the Shapley values are typically computed separately on each class in a one-vs-rest manner, ignoring the compositional nature of the output distribution. I gave a brief introduction to *Shapley compositions*, a well-founded way to properly explain a multiclass probabilistic prediction, using the Aitchison geometry from compositional data analysis. In particular, the norm of Shapley decompositions can be used to quantify feature compositions over all classes.

3.5 Explanations as Constructed Arguments

Peter Clark (Allen Institute for AI – Seattle, US, peterc@allenai.org)

In this talk I'll offer some perspectives about explanations and their role. I use the following definition: *Explanations are constructions to convey a well-founded argument about why a conclusion is valid.* While a human or machine may arrive at a decision via some opaque method, e.g., with an LLM, we may then *explain* those decisions in a symbolic way, showing how the conclusion systematically follows from facts which the model believes or is provided with. Note that the explanation does not necessarily reflect what the model *did*, but rather why the conclusion is valid. The explanation can be viewed as an orthogonal, but equally valid, way of showing why the model's output is rational given its inputs. In the work my group has been doing, we have been using textual entailment as the formalism for building such chains of arguments, in which a LM first generates an explanation then validates that it itself believes (via self-querying) both the facts and inferences in that explanation – hence it is a "faithful" explanation. If the system's conclusion turns out to be incorrect, we thus now have a way of debugging where the error was (a fact, or an inference) in the system's argument, and potentially correcting that error by updating the model [1]. Four evaluation criteria are useful for such explanations: (a) are the basic facts correct? (b) is the reasoning accurate? (c) Can the user comprehend it? and more generality (d) does the explanation also help the user predict answers to other questions, i.e., has the explanation conveyed a broader "mental model" of the machine? [2, 3, 4]. Argument-based explanations like these are particularly useful for model improvement, as users finally have an interpretable view of what the model "knows" and how that knowledge justifies its conclusions.

Providing a documentation for a Dagstuhl Seminar is mandatory. We focus on talk abstracts and show that a talk abstract can be tagged with co-authors appearing in the joint-work-of-field. Furthermore, a talk abstract can state one main reference on which the talk is based.

References

1 Bhavana Dalvi, Oyvind Tafjord, and Peter Clark. Towards teachable reasoning systems: Using a dynamic memory of user feedback for continual system improvement. *ArXiv*, abs/2204.13074, 2022.
2 Harsh Jhamtani and Peter Clark. Learning to explain: Datasets and models for identifying valid reasoning chains in multihop question-answering. In *EMNLP*, 2020.
3 Peter Clark, Bhavana Dalvi, and Oyvind Tafjord. Barda: A belief and reasoning dataset that separates factual accuracy and reasoning ability. *ArXiv*, abs/2312.07527, 2023.
4 Bhavana Dalvi, Peter Alexander Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. Explaining answers with entailment trees. In *EMNLP*, 2021.

4 Introduction

Explanations have garnered escalating interest within the AI and Machine Learning (ML) communities. Yet, at times, a crucial aspect that tends to be overlooked is the recognition that explanations can be leveraged for different objects and when evaluating the utility of

these methods the objective needs to be taken into account. Explanations can enhance transparency, help users for a cognitive model of a trained ML system, aid in debugging, or assist users in determining whether to place trust in a prediction or recommendation. While many explanatory mechanisms have been proposed in the community, comparing these solutions remains challenging without the adoption of more standardized practices in terms of evaluation. Compounding this issue is the versatile nature of explanations, meaning algorithm designers in reality should tailor their evaluation strategies to specific tasks. To address this, we build upon the taxonomy presented [15] to identify the different objectives and tasks for explainability methods create guidelines for adaptable tasks and experiments for the community.

Several other taxonomies and frameworks to enable practitioners to develop and evaluate explanations about AI system behaviour [15, 20]. Fundamental questions should be considered when evaluating the impact of an explainability method, what are the **goals or objectives** of the XAI method, what **tasks or usage context** with the method be used for and importantly who are the **stakeholders** of the system [14].

5 Evaluation Criteria for Explanations

To frame our discussion, we began by defining and refining our working definition of evaluation criteria for explainable AI (XAI). Our review of the taxonomy provided by [15] highlighted that, while it serves as a useful starting point, it presents several limitations:

- The criteria are heavily focused on computational or technical aspects, with limited attention to human-centered metrics.
- The taxonomy does not comprehensively map the space of possible evaluation methods.
- Several important criteria are either missing or inadequately defined, including:
  - Trust, Calibrated Trust, and Reliance
  - Human-AI Team Performance
  - Situational Awareness
  - Cognitive Load
  - Explanation Satisfaction
  - Fluency in Human-Autonomy Teaming
  - User Satisfaction
  - Efficiency (e.g., number of explanations/interactions required)
  - Accessibility and Modifiability
  - Distinctions between model and explanation evaluation
  - Calibration and Human-AI Alignment

These gaps reflect a broader issue: current evaluation metrics tend to focus solely on the XAI method itself. Moreover, the criteria are often not well-defined or easily interpretable for HCI researchers and practitioners. Given that usage contexts vary, evaluation criteria must be carefully selected and adapted accordingly. It is rarely feasible to optimize for all criteria simultaneously, making it essential to prioritize based on context. However, the lack of clarity around evaluation criteria makes it difficult to reason about or prioritize trade-offs.

5.1 Self-Reported and Observed Understanding

The idea of *ultra-strong machine learning* suggests that ML systems should not only be able to learn hypotheses in symbolic form but also teach humans about what they have learned, enabling stronger human-AI team performance overall [17]. Explainable AI can help to achieve this end by supporting humans-in-the-loop to develop strong, accurate, and aligned mental models about AI system behavior which enable them to flexibly interact with and apply these systems across all necessary operational contexts. In this sense, one of the main objectives of XAI is to build robust human mental models that facilitate user perception, comprehension, and prediction of AI behavior. In order to know whether explanations have achieved this end, we need a way to assess a human user's comprehension of a system before and after receiving explanations about its functioning. Thus, one key criteria to consider in assessing the overall efficacy of any given explanation is user understanding. We suggest evaluating understanding through a suite of objective and subjective assessments, which we break into two primary categories: self-reported understanding and observed understanding. The value of assessing understanding both subjectively and objectively is that this allows us to compare a user's perceived understanding versus their true understanding and whether these two are aligned – in other words, whether understanding is well-calibrated. This is critical as over- or under-confidence could lead to over-use or under-use of this system, hampering team performance overall [18].

5.1.1 Self-Reported Understanding

Criterion: Perceived Understanding

Definition: The extent to which users believe they understand the model, its outputs, and the explanations.

Source: Human

Type: Subjective (e.g., self-reported via questionnaire)

In order to assess self-reported understanding, in line with other subjective assessments from the human factors literature [22], we suggest developing a suite of Likert scale-based questions, which probe users about their individual perceptions of how well they comprehend an AI system given any explanations that they have been provided. While Hoffman et al. have proposed explanation satisfaction and trust scales [10], scales that focus on self-reported human understanding have been underexplored to date.

Previous examples of such questions include [3, 24] (1) I understand how the model works to predict whether a defendant will reoffend [whether the primary tree species in an area is spruce/fir; "I understand the admission algorithm"]; (2) I can predict how the model will behave.

In addition to asking for Likert-based responses to questions from the categories above, it would be additionally useful to ask users to self-report their confidence for each item. We note that items may vary in terms of understanding complexity, and show different patterns in performance across a set of participants [25]. Therefore, we also recommend analysing performance on individual or specific questions rather than computing them on aggregate (e.g., sum of accurate responses).

5.1.2 Observed Understanding

Criterion: Actual Understanding

Definition: The accuracy of a user's understanding of the model, its inputs, outputs, and explanations.

Source: Human

Type: Objective (e.g., mental model elicitation)

The fundamental challenge in assessing user understanding in an objective manner is selecting assessments and metrics that faithfully reflect the user's underlying mental model. Mental models can be shallow, covering only a functional understanding of a system (e.g. a driver knows how to operate a car), or deep, achieving a more structural understanding of how a system functions (e.g. a mechanic understands the inner-workings of a car and how to fix broken cars or extend their capabilities) [13]. Importantly, the appropriateness of such mental models depends on a user's context, including their attributes and expertise, tasks, use cases, and goals. Previous work has proposed a mental model soundness scare which addresses these factors and incorporates domain-specific comprehension questions [13].

One structured approach to determine a user's goal-informed informational needs within a given context is based on the situation awareness (SA) framework from the human factors literature. Endsley defines SA as the perception of elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status into the future [7]. SA requirements for a given context define a person's informational needs which can be met in part through the application of XAI when a human is working with an AI system within their context. Sanneman and Shah apply the SA framework to XAI, and define the following three levels of XAI [20]:

- Level 1: XAI for Perception—explanations of what an AI system did or is doing and the decisions made by the system
- Level 2: XAI for Comprehension—explanations of why an AI system acted in a certain way or made a particular decision and what this means in terms of the system's goals
- Level 3: XAI for Projection—explanations of what an AI system will do next, what it would do in a similar scenario, or what would be required for an alternate outcome

A process such as goal-directed task analysis (GDTA) can be applied to elicit a user's informational needs at these three levels [8], applicable explanations techniques can be selected to meet these needs, and their efficacy at supporting a user's mental model can be assessed by applying a technique such as the situation awareness global assessment technique (SAGAT), a validated SA assessment from the human factors literature [7, 19].

There are various other models. Speith et al. [21], for example, generalize Sannemann and Shah's model by incorporating findings from various disciplines outside of human factors (e.g., cognitive psychology, philosophy). Their model comprises six levels of skills that can be tested in studies and are said to correlate with varying degrees of understanding. Their aim is to make studies more comparable.

5.2 Explanation Fidelity and Stability

After generating an explanation, the first step is to assess its correctness. To this end, we identified two key properties that allow for a technical evaluation of explanations, namely explanation fidelity and stability. According to the state of the art, these criteria are essential

for assessing the reliability and trustworthiness of explanations, although their formalization and practical application are still challenging [2, 9]. In particular, the literature offers many different definitions of fidelity, faithfulness and correctness, each leading to distinct technical implementations, depending on the context, kind of data, ML model and explanation. Given the complexity of this setting, our first priority is to verify the technical validity of the explanation. Our rationale is that, before evaluating the quality of an explanation through user studies, we must first ensure that it is mathematically sound.

To this end, we consider an ML model $b(\cdot)$ that has already been trained on a dataset D_{train} and exhibits good predictive performance, with generalization capabilities and no signs of overfitting. To explain our model $b(\cdot)$, we consider an explanation method $g(\cdot)$ that outputs explanations e. At this stage, we intentionally keep the definitions at a high level to allow for the definition of criteria that are general and not tailored to a specific ML task or explanation type. As such, the explanation in this setting may take various forms, including feature importance, decision rule, or even a chain of thoughts. The same applies to the ML task.

5.2.1 Explanation fidelity

Criterion: Fidelity

Definition: The degree to which the explanation accurately reflects the model's decision-making process.

Source: Model (may require human-verified ground truth)

Type: Objective

We first address the definition of explanation fidelity (or faithfulness). The objective in this case is to evaluate whether an explanation accurately reflects the internal reasoning of the model. We consider this an objective property, since it concerns the alignment between the explanation and the model's actual behavior, independently of any human interpretation. The definition of explanation fidelity is not straightforward. In the literature, we can find many terms, including faithfulness and correctness. However, faithfulness may be misleading due to its connotation of belief or trust, which does not align with the technical nature of the concept. In addition, its definition is highly task-dependent and varies significantly depending on the type of explanation considered. For instance, when considering feature importance-based explanations, faithfulness can be evaluated by masking the most important features at prediction time, but this only applies for this specific kind of explanations. Rather than relying on a single notion of fidelity, a useful perspective may come from related concepts such as *comprehensiveness* and *sufficiency* [6]. A comprehensive explanation contains all the critical components that influence the prediction, while a sufficient explanation identifies the minimal set of elements necessary to reach the same outcome. In particular, an explanation makes claims about the factors that cause or influence the model's prediction. Therefore, the main objective of *explanation fidelity* is to verify whether these factors are truly influential. In practice, if a change in the explanation leads to a change in the prediction, this supports the causal relevance of the explanation's components. As already mentioned, the technical implementation of this criteria can be challenging given the different ML tasks available, as well as the various kind of explanations. However, the evaluation can be approached by altering the components identified in the explanation, such as masking an important feature in a classifier or removing a rule in an expert system, and observing whether the model's output changes accordingly.

5.2.2 Explanation stability

Criterion: Stability

Definition: The consistency of explanations for similar inputs or outputs.

Source: Model / Human

Type: Objective / Subjective

Another important aspect of the explanation is its stability. Also in this case we can find different terms, such as consistency or robustness, and many definitions. After our discussion, we claim that stability refers to the consistency of a model's predictions and explanations when provided with similar inputs. A stable explanation method should produce similar outputs and explanations for inputs that are close according to a similarity metric. To set the stage, we can consider two similar records, x and x^1 that give the same output, $b(x) = b(x^1)$. In this case, we expect two similar explanations, e and e^1. Therefore, the evaluation requires two components: a metric for measuring similarity between input instances x and x^1, and a metric for comparing the corresponding explanations e and e^1. This property can be assessed through objective criteria, but we prefer to consider it as an *objective and subjective* criteria. In fact, when defining what constitutes *similar* inputs or explanations, it may also involve human intervention to define or validate similarity measures. The specific approach to measuring stability can vary depending on the task at hand and the type of explanation being considered.

5.3 Trust, Reliance and Performance

It is necessary to clearly distinguish between "trust" and "reliance" in a system. Although these concepts are interconnected, they are fundamentally separate [1]. Reliance is a behavior when adopting the recommendations of the system or delegating (sub)tasks to it, while trust is considered a relational process between a trustor and a trustee in a specific context with a trust goal [1]. Reliance can indicate trust, but does not entail it. While reliance on a system can be observed and measured in different ways, there is no distinct way to measure users' trust in a system.

5.3.1 Performance

Criterion: AI-Human Team Performance

Definition: The overall performance of decisions made by the AI-human team.

Source: Human

Type: Objective (e.g., task performance)

Although the performance of a Human-AI team is often the primary goal in most applications and heavily affected by users' trust and reliance, the measurement of performance is left here vague on purpose, as it is very task specific.

You would use the same metric that you would use to evaluate the model in isolation (or the user in isolation). For instance, in model debugging performance means the quality of the model one obtains, whereas in a classification task it is rather the accuracy the joint performance of human and system.

5.3.2 Self-reported Trust

Criterion: Self-Reported Trust

Definition: How much the user reports trusting the AI system.

Source: Human

Type: Subjective (e.g., questionnaire)

Self-reported trust is the extent to which a user believes they trust in the AI model. There is a multitude of questionnaires to measure self-reported trust (see [12] for an overview). Some of the frequently used questionnaires may not be appropriate in every situation. For example, the questionnaire by Hoffman et al. [10] tends to include concepts different from trust.

Additionally, self-reported trust can be measured by means of betting markets, i.e., how much the participants are willing to bet on the model giving a correct output in situations without having personal stake.

Additional thoughts on self-reported trust in a specific instance:

- possibly only for experts, as laypeople might not distinguish between the system and its individual explanations
- maybe the very same metrics apply (e.g. trust in automation), with some customization

5.3.3 Observed Reliance

Criterion: (Appropriate) Reliance

Definition: The extent to which users appropriately follow correct or

Source: Human

Type: Objective (e.g., behavioral observation)

Methods to measure observed reliance include:

1. Investment games in which the participant initially has several points that can be used to bet on the AI system. The number of points they are willing to spend indicates the reliance on the system. Two interesting resources for this kind of task are: [16] for general for general XAI models and [11] for explainable reinforcement learning.
2. Stakes scenarios that investigate the behavior in different settings, e.g., using the system to recommend a movie vs. to diagnose a life-threatening illness
3. Delegation tasks that examine in which cases the users entirely delegate a (sub)task to the system. These can be augmented with betting.
4. Measurement of "switch rates" [26]: How often and when do users switch to the system's recommendation in case it deviates from their own initial assumption or outcome? This can be designed as a multi-step approach where the user makes an initial assumption and then is presented the system's recommendation. In case the user and system results are different, the participants in a final step have to decide whether they keep their own result or adopt the system's recommendation.
5. Willingness to follow advice: how much do people deviate towards the algorithmic estimate based on their own estimate. This approach is similar to the switch rate task but applied to numeric decisions.

6 Relationship between Usage Context and Criteria

As highlighted by [15] different criteria become more relevant in assessing explanations depending on the target usage context. In the following sections we begin to reason about and prioritise the most relevant criteria and measurements for a subset of the usage contexts.

6.1 Metrics for Model Improvement

Besides helping the user, one important objective of explanation is to help experts and/or system builders **improve the problem-solving system** ("model", henceforth) itself. Such model improvements can occur:

- During model development, to inspect how the model could be improved and make such improvements
- after model deployment, to verify if the model is behaving as intended or needs further improvement

There are numerous mechanisms that can be employed to modify model behavior, e.g., adding extra training data and retraining; modifying a rule or concept in a rule-based component; changing weights on features; masking out elements known to be irrelevant to a result. The role of explanation is to help the expert/system designer understand why a model produced a wrong answer, and what kinds of interventions might correct the error both for a specific case and future cases. This whole endeavor is not just about producing good explanations to help in this process, but also designing a model architecture in the first place that supports such explanations and allows easy model improvements – the technology of interactive explainable AI (XAI) [23]. Note that model improvement may occur both

6.1.1 Relevant Metrics

How can we measure whether explanations help experts/designers improve models? We consider two types of measurements:

Primary Measures.
We identify several **primary measures** that can be used to directly measure model improvement:

- Performance, e.g., accuracy, performance of the human AI team
- Quality of the explanations (that aid in the end goal of improving performance)). Measures of quality (defined and described in more detail elsewhere in this document) include:
  - Stability
  - Faithfulness
  - Understanding
  - Contextless (knowing the bounds or limitations of the explanations e.g. where it would not hold)
  - Action-ability

Note that simply measuring the change in such metrics before/after a model update is not sufficient to show that explanations help. Rather, the experiment should compare improvement without/with explanations helping the person improving the model.

Secondary Measures.
In addition, there are some **secondary measures** that do not directly measure performance improvement, but are likely correlated with it and desirable to also improve (and at least observe):

- Change in trust
- Faithfulness

Finally we note that other measures, e.g., end-user satisfaction, are less relevant for the specific goal of model improvement (though clearly critical for the end system).

6.1.2 Trust and Model Improvement

There is a delicate relationship between trust and model improvement. If a domain expert is involved, then exposing them to model errors may reduce their trust in the system. Conversely, if they have been involved in improving the model, then this may help increase their trust, and in fact involving domain experts may ultimately help them appropriately calibrate the right level of trust they should have in the system's behavior. This is an important aspect to track and study in model improvement experiments, even though it is not the primary objective.

6.1.3 Sources of Model Misalignment

Training data misalignment Data misalignment between training data vs test (debugging) data Fundamentally missing from the training data (vs complete ground truth of all the possible data) Improving alignment between the AI model and expert user's knowledge on the task Example based correction: by explaining the missing examples, one can point out which part of data is missing.

Why might a system be making mistakes in the first place? It is useful to consider two dimensions of misalignment (between the actual model and the ideal/perfect target model):

Training data limitations.
In a machine learning context, models are only exposed to a sample (namely, the training data) taken over the distribution of problem-solving tasks, and thus the learned model may be somewhat misaligned with the actual ideal model. While we do not have access to that ideal model (ideal ground truth), we approximate this by using an independent, hidden test set (a sample of that ideal ground truth), to measure model performance.

Domain expertise.
There may also be additional knowledge beyond that captured in examples that is relevant to the task, again contributing to a misalignment between the actual model and an ideal oracle model. The model improvement process provides an opportunity for experts to inject that extra knowledge into the system, e.g., by providing additional examples for areas of the problem space that the model is either ignorant of or unsure about.

6.2 Metrics for Capability Assessment/Auditing

An important usage context of explanations is the assessment of a system's capabilities (e.g., fairness, safety, performance), also known as *auditing*. In general, a capability is part of the system; and there are specific criteria that the model has to satisfy to have a certain

capability. Auditing, then, is the way to find out whether the system satisfies specific criteria and, thus, has the corresponding capabilities. As explanations can help to find out whether specific criteria are satisfied, they are a means to auditing.

Since most of these criteria are those that concern the whole system, global explanations are most helpful for auditing. Nevertheless, the usefulness of local explanations should not be overlooked, as they can indicate that something is not correct and can be the base for further deliberations on a capability of interest. If explanations reveal that some criteria for a system capability are not met (e.g., for fairness that no protected attributes unduly influence decision-making), then the system is flawed in this respect and mitigation strategies must be initiated.

In general, it is quite open who can be the auditor. Obvious possibilities are external (accredited) watchdog organizations (such as the TüV in Germany), but also regulators and interested individuals. The only requirement is that they are as objective as possible with regard to the capability they are auditing. An example of an interested individual who does audits is the developer who aims at a high accuracy of the system.

Audits can take place at various points in time. They can be conducted regularly, for example when expiring certificates need to be renewed or when required by regulations, or only when it is discovered that something has gone wrong. In the latter case in particular, the affected party is usually left out of the loop because they either did not realize that they had been negatively affected (which is often the case with discrimination, for example) or do not have the means to defend themselves against a false assessment. This raises the question of how affected parties can be better involved in the auditing process. Related to this question, but also going beyond it, is the question of which explanations are most useful at what part of the process. Accordingly, the measures for evaluating these explanations are also diverse.

In the case of auditing, metrics such as (human-AI) performance and reliance are not relevant. The most important metrics are:

- (actual) understanding: the auditor should understand the system
- fidelity and stability: the explanations should reliably and truthfully track the system's decision-making processes.
- coverage (i.e., the distribution of explanations): the explanations should cover as many cases as possible.
- self-reported trust: the auditor should believe that they trust the system.

However, auditing using XAI can be difficult in many cases if the important criteria can only be checked when the system is in use or are even outside the (usual) realm of XAI. An example of this is security, where the provenance of the training data is also important. Another example is fairness [4, 5]. Especially when it comes to fairness, a mere consideration of outcome fairness is often not enough [5]. One reason for this is that there is no ground truth in certain areas. In the case of loan applications, for example, there is only partial ground truth data, as it is never known whether someone who has been refused a loan would not have repaid it after all.

Furthermore, there are various types of fairness that can also be important. Informational fairness (which is not part of the model), for example, deals with the question of what information a particular party has received about a process and whether this information is sufficient, faithful, and adequately prepared. Procedural fairness, on the other hand, asks whether the decision-making process itself is designed in such a way that it leads to fair outcomes. Both are types of fairness whose fulfillment cannot be determined by traditional XAI explanations. Accordingly, auditing requires explanations that XAI does not yet provide.

7 Case Study

In this section, we discuss an end-to-end pipeline for designing an XAI experiment. The participants broke up into groups and collaboratively developed concrete evaluation scenarios across different application domains. Each group explored how to operationalise key explainability criteria within their context, identified appropriate stakeholders and metrics, and proposed experimental designs to assess the impact of explanations on human-AI interaction.

7.1 Model Improvement (Medical Domain)

We focus on image-based classification task in the medical domain. Specifically, we discussed classification of tumors or skin lesions. The goal is to classify an image based on the tumor type detected in the image. There are five tumor types considered in this example, with values $1 - 5$. The label indicates the progression of the tumor, with 1 being the least and 5 the most dangerous. The challenge of classification problem is in distinguishing between similar classes, specifically discerning between types 2 and 3 of tumors.

7.1.1 Stakeholders

We distinguish between stakeholders and their roles in pre and post deployment scenarios. We identify the main roles from the perspective of model improvement task as following:

1. Pre-deployment:
   a. Model developers: the goal is to improve the model for deployment
   b. Domain experts: might be consulted during pre-deployment to add expert knowledge to the model. Domain experts could spot model errors and gaps, identify corrections, provide labels, annotations, and data in general
2. Post-deployment:
   a. System user: with the goal of performing the end task
   b. Model auditor:

7.1.2 xAI Pipeline

AI Model. To implement image-based classification, we focus on the following two classification models:

1. CNN – representing a black-box approach to image classification task.
2. Concept bottleneck model – capable of providing more high-level concept explanations.

xAI Methods. To generate explanations for the AI model's decisions we discussed the following explanation methods:

1. Saliency maps could identify the parts of the images that led the AI to make a specific decision. Could be especially interesting to uncover spurious correlations in data.
2. Concept level explanation: these could be a result on the concept bottleneck models.
3. Example based:to explain a current decision, a previous example where the same decision was reached could be informative.
4. Prototypical: explains decisions by offering "prototypes" of different classes (e.g. the input is classified as y because it looks like the representative of this class).
5. Countrastive/Conterfactual (near miss example): could highlight the nuances between similar classes for critical decisions on the decision boundary.

Data Collection and Preparation. To train the AI model a dataset of labeled images from this medical domain is required. The features and labels in the dataset can be verified by domain experts during pre-deployment. The data should be split into three sections – train, validation and test. A (potentially flawed) model is then trained on the train dataset. The validation dataset is presented to the user or domain expert who can advise on potential corrections. Then the retrained model (taking into account user/domain expert's feedback) is evaluated on the test split.

Additionally, at this point it might be advisable to identify the type of errors on which the model improvement task focuses. These can be a consequence of known spurious correlations in the data, missing or incorrectly labeled data.

xAI Study Data Collection and Preparation. A dataset of 20 images can then be selected to be used in the xAI user study. As the goal of the study is to investigate model improvement, the model should likely make mistakes on some of the presented input images. However, a large number of mistakes could quickly lead to deterioration of user trust and impact the measurements of other xAI metrics. The advised number of mistakes on a dataset of 20 images is 3.

xAI Study Design. The study should begin by informing the participants of the domain problem, the AI model, explanations (depending on the study condition) and their task in correcting the model outputs. It is advisable to inform the participants that the AI system is well trained, however, it can also make mistakes. Otherwise, users might quickly lose trust in the system after observing errors. However, it is not clear should the reported accuracy equal the real accuracy or be set to a predefined value (e.g. 95%).

To investigate if the expert's corrections are a consequence of the presented explanations, it is likely that a baseline condition is needed where experts are asked to correct the model behavior without access to explanations. Furthermore each explanation type is going to elicit another study condition.

xAI Metrics. The following metrics were discussed to evaluate the quality of explanations for the model improvement task:

1. Human-AI performance: measures the accuracy of the expert who can receive recommendations on the decision. Does the expert follow the model's outputs or makes its own decisions? In the ideal scenario, the expert agrees with the model when the model is correct but corrects it when its makes an error.
2. Model performance after corrections: after receiving expert's corrected labels and retraining the model, does the performance (on some metric like accuracy) improve?

7.2 Capability Assessment (Hiring)

We considered capability assessment – and specifically fairness assessment – of an AI system for hiring. The system is intended to pre-select or screen promising candidates out of a pool, so that successful applicants will later on undergo a hiring interview. This can be most naturally be cast as a binary classification task: "pass" vs "fail".

7.2.1 Stakeholders

In order to get a sense of what criteria and metrics would be useful for the specific use case, we consider different users that might be interested in assessing fairness of this AI, and specifically:

- The ***job applicant***: they are trying to get a job and they presumably receive a pre-screening output from the AI. Naturally, it is in their interest to verify that the AI is indeed fair.
- The ***hiring manager*** of the company offering the job.
- The "***watchdogs***", e.g., the ethics department of a company, the union, etc.
- (Optionally) The ***job centre personnel***, who are responsible for facilitating job applications.

7.2.2 Detailed Setup

We assume the model is a machine learning classifier trained to discriminate between promising candidates and the rest on historical data. The training examples and the input instance would consist of text records (e.g. CV, education, prior positions) either pre-processed into or paired with tabular data (e.g., personal information). We do not focus on a specific classifier architecture, for two reasons. First, we assume the classifier has been trained – presumably by either the company or the job centre – for good performance out of the many option available. Second, many high-quality explainability techniques are model agnostic and can provide competitive explanations for a variety of classifier architectures. Of course, classification performance should be tracked (for instance, via model accuracy or F_1 score) for consistency with the primary goal of selecting promising candidates. It is a prerequisite that the classifier achieves non-trivial prediction accuracy.

We would expect four possible kinds of information would be of interest for assessing fairness for the four stakeholders we consider:

- Prediction confidence – useful across the board.
- Feature relevance – this is especially useful for hiring managers and watchdogs for understanding whether the model is, e.g., leveragin protected attributes for its decisions.
- Counterfactual explanations – these are especially useful for the applicant (and potentially the job centre facilitator) so as to gain actionable insights about the decision.
- Prototype-based explanation, e.g., distance from "ideal" applicants – these *might* be useful to get a sense what kind of profile(s) the classifier is expecting successful applicants to have, and whether these are in any way undesirable.

7.2.3 Evaluation Metrics

- ***Perceived prediction quality and fairness***: this is the basic capability being assessed.
- ***Satisfaction with explanation***: whether explanations are perceived as useful for assessing fairness.
- ***Actual understanding***: whether explanations have been in fact understood by stakeholders (rather than merely perceived as useful).
- ***Self-reported trust***: whether the stakeholder believes they can rely on the AI doing a good job at generating fair predictions.
- "Correctness" of explanations, and specifically:
  - ***Fidelty***: lack of fidelity means that it may be impossible to map poor justifications for the predictions (e.g., reliance on protected attributes) to the model's actual behavior and capabilities.
  - ***Stability*** & ***Coverage***: feature relevance explanations only provide a local, per-candidate view of the model's reasoning; this does not necessarily generalize to other instances unless the explanations are somehow stable, i.e., do not vary enormously for similar candidates (and potentially decisions); coverage refers to the fact that in order

to get a sense of the overall capabilities of the classifier as a whole, for all possible instances, it may be necessary to obtain local explanations for a sizeable number of individual cases, for statistical reasons.
All these metrics can be computed mechanically without users studies.

7.2.4 Evaluation

The idea is to use a between-participant experimental design to assess the relative performance of different UI designs. We suggest to focus on a total of 6 designs, one for each combination of explanation type (3 total) and (2 total).

An online study seems to be sufficient for evaluating the AI and its explanations given the chosen metrics and setup. The task could be briefly summarized as "look at UIs and then complete a questionnaire: how well do you think it does its job, given your specific role?"

The question is how to evaluate the UIs from different user perspectives, and specifically how to carry out the recruiting. Ideally, we would need a reasonably large sample of applicants, job centre personnel, hiring manager, and watchdogs. This immediately poses the issue of how to get access to such a sample. Naturally, power analysis can help to identify a sufficient sample size. For certain users – e.g., applicants – one could implement a role playing setup in which Prolific participants are asked to act as applicants and evaluate the AI system under the aforementioned metrics, obtaining feedback via questionnaires. This is more problematic for specialized users like hiring managers, who might be more difficult to simulate or role play properly.

7.2.5 Hic Sunt Leones

Mental model /situational awareness extraction framework Questionnaire about perceived fairness / trust Satisfaction with explanations Open text response: any other info that would have helped you assess? Demographics

7.3 Decision Support (ICU Triage)

In this section we describe an example evaluation methodology for decision support systems.

To ground the discussion, we chose an example use-case instead of describing a general decision support problem.

7.3.1 Chosen use case for the evaluation

For the use-case, we decided on a medical diagnosis task. Specifically, a situation where a healthcare professional has to make a single decision. An example could be ICU triage – one single patient where the decision has to be made whether the person needs to wait or needs emergency treatment now. So, we can think of a single nurse that is working during a particular shift and one ICU bed has opened up, who of the top 5 patients at risk do they admit.

7.3.2 Things to think about that you might need for your evaluation

For our use-case we mainly need to things. Mecial Use Case/Data: We need a medical use case given by patient data. This data normally consists of both self-reported (interview) data and objective measurements (vitals). The modality of this data is often very multi-modal including text, images and tabular time series (EHR). Ideally, we would like to get this data from use case studies for medical training.

Model: A predictive AI Model that outputs urgency rankings for each patient. Based on this, the model gives a ranking where, e.g., Person A is more urgent than Person B.

7.3.3 Chosen sample explanation method

To ground the discussion some more, we imagined two example explanations we want to compare. Because the task is comparative (which of the participants do you admit to the ICU), we chose two contrastive explanations. First, counterfactual explanations that change the input such that the model rates Person B as more urgent than Person A. Second, feature attribution that show how relevant each input was for the AI's decision to assigning more urgency to Person A than Person B. In general, our evaluation example focuses on local explanations.

7.3.4 Participants

Because medical background knowledge is crucial to properly interpret the explanations we would aim to recruit healthcare professionals for our user study. In particular, for our use case, we would try to recruit nurses.

7.3.5 Task Setup

At first, the participants are given a description of the task and of the AI and Explanations methods that they will see during the study. [optional:] They will also get a pre-questionnaire about their demographics, medical expertise and previous experience with AI and XAI.

Afterwards, the participants will be given X decision tasks where they are told that one ICU bed opened up and they have to decide which of 10 patients they admit. First, they have to decide by themself, without AI support, which patient they admit. After that, they see an AI Recommendation for the urgency ranking of the particpants. Depending on the condition, they will also see an explanation here.

After X decisions, they will [optional: move to the understanding task and then] get a pos-questionnaire.

7.3.6 Conditions

We envision a between subject design.

- Baseline Condition: These participants only see the recommendation of the AI.
- Explanation Condition 1: These participants see the AI recommendation together with the first explanation method – in our case counterfactual explanations.
- Explanation Condition 2: These participants see the AI recommendation together with the second explanation method – in our case contrastive feature attribution.

7.3.7 Metrics

We will measure two tothree observed (or objective) metrics:

- **Performance**: Did nurses/doctors select the correct patient? And also interesting: how far is the distance from the 1st rank is the admitted person (relative performance). To incoporate the possibility of the AI making mistakes we will assume a somewhat realistic accuracy of 80%. That means that in 80% of the example decision, the AI will be correct. We do not choose a lower rate, since it might unrealistically bias the participants against the AI. However, in addition to the general performance, we will also reported individual performance for the group of correct and incorrect AI predictions.

- **Appropriate AI Reliance:** To what extent did people deviate towards the AI advice, given that they gave their initial estimate, then received the AI's advice, and based on the latter, decided to comply (or not – or to what extent) with the AI's estimate ("weight on advice")
  Appropriate reliance = AI was recommending patient #1 ranking, and the user was following it (final user estimate = AI estimate = best) Underreliance = AI was recommending patient #1 ranking but the user did not follow it (final user estimate ≠ AI estimate & AI estimate = best)
  Overreliance = was NOT recommending patient #1 ranking and the user followed it (final user estimate AI estimate & AI estimate ≠ best)
  Special case = doctor's initial estimate = best = AI estimate → "reinforced" appropriate reliance
- **Observed Understanding [Optional, not as crucial for this task]**: After the X decisions, we could add a prediction task as proposed by [10]. In this task, participants will see Y additional examples. Here, they will only see the input and the explanation (or no explanation for the baseline). Based on this they have to predict the urgency ranking of the AI. Depending on how good they are at predicting the AI will be used to judge their understanding of the AI models reasoning.

Additonally, we will measure several self-reported (subjective) measures in the post-questionnaire:

- Self-reported Trust: Trust in Automation (e.g., Perceived Competence, → Madsen & Gregor, 2000)
- Perceived Understanding
- Perceived Helpfulness
- Satisfaction with Explanation
- User experience questionnaire? (Maybe)
- Task Load / Perceived Accomplishment

8 Conclusion

This seminar marked an important step in bridging the gap between human-computer interaction (HCI) and artificial intelligence (AI) research communities. By bringing together experts from both fields, we created a multidisciplinary forum that encouraged critical reflection on the goals, assumptions, and evaluation methods of explainable AI (XAI). Discussions focused not only on technical soundness but also on human-centered evaluation and usability in real-world contexts. Through collaborative case studies, taxonomy refinement, and shared methodological frameworks, we initiated a dialogue that will continue to shape the development of robust, transparent, and user-aligned AI systems. This seminar laid the groundwork for ongoing collaboration between HCI and AI researchers, emphasizing the importance of inclusive, reproducible, and context-aware evaluation practices in the evolving landscape of responsible AI.

References

1 Michaela Benk, Sophie Kerstan, Florian von Wangenheim, and Andrea Ferrario. Twenty-four years of empirical research on trust in ai: a bibliometric review of trends, overlooked issues, and future directions. *AI & SOCIETY*, October 2024.

2 Francesco Bodria et al. Benchmarking and survey of explanation methods for black box models. *Data Mining and Knowledge Discovery*, 37(5):1719–1778, 2023.

3 Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. Explaining decision-making algorithms through ui: Strategies to help non-expert stakeholders. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–12, 2019.

4 Luca Deck, Jakob Schoeffer, Maria De-Arteaga, and Niklas Kühl. A critical survey on fairness benefits of explainable AI. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2024, Rio de Janeiro, Brazil, June 3-6, 2024*, pages 1579–1595. ACM, 2024.

5 Luca Deck, Astrid Schomäcker, Timo Speith, Jakob Schöffer, Lena Kästner, and Niklas Kühl. Mapping the potential of explainable artificial intelligence (xai) for fairness along the ai lifecycle. In Mattia Cerrato, Alesia Vallenas Coronel, Petra Ahrweiler, Michele Loi, Mykola Pechenizkiy, and Aurelia Tamò-Larrieux, editors, *Proceedings of the 3rd European Workshop on Algorithmic Fairness, Mainz, Germany, July 1st to 3rd, 2024*, volume 3908 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2024.

6 Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4443–4458. Association for Computational Linguistics, 2020.

7 Mica R Endsley. Toward a theory of situation awareness in dynamic systems. *Human factors*, 37(1):32–64, 1995.

8 Mica R Endsley. Direct measurement of situation awareness: Validity and use of sagat. In *Situational awareness*, pages 129–156. Routledge, 2017.

9 Peter Hase and Mohit Bansal. Evaluating explainable AI: which algorithmic explanations help users predict model behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, 2020.

10 Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. Measures for explainable ai: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-ai performance. *Frontiers in Computer Science*, 5:1096257, 2023.

11 Tobias Huber, Maximilian Demmler, Silvan Mertes, Matthew L. Olson, and Elisabeth André. Ganterfactual-rl: Understanding reinforcement learning agents' strategies through visual counterfactual explanations. 2023.

12 Spencer C. Kohn, Ewart J. de Visser, Eva Wiese, Yi-Ching Lee, and Tyler H. Shaw. Measurement of trust in automation: A narrative review and reference guide. *Frontiers in Psychology*, 12, 2021.

13 Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. Tell me more? the effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the sigchi conference on human factors in computing systems*, pages 1–10, 2012.

14 Q Vera Liao, Milena Pribić, Jaesik Han, Sarah Miller, and Daby Sow. Question-driven design process for explainable ai user experiences. *arXiv preprint arXiv:2104.03483*, 2021.

15 Q Vera Liao, Yunfeng Zhang, Ronny Luss, Finale Doshi-Velez, and Amit 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.

16 Tim Miller. Are we measuring trust correctly in explainability, interpretability, and transparency research? *arXiv preprint arXiv:2209.00651*, 2022.

17 Stephen H Muggleton, Ute Schmid, Christina Zeller, Alireza Tamaddoni-Nezhad, and Tarek Besold. Ultra-strong machine learning: comprehensibility of programs learned with ilp. *Machine Learning*, 107:1119–1140, 2018.

18 Raja Parasuraman and Victor Riley. Humans and automation: Use, misuse, disuse, abuse. *Human factors*, 39(2):230–253, 1997.

19 Raja Parasuraman, Thomas B Sheridan, and Christopher D Wickens. Situation awareness, mental workload, and trust in automation: Viable, empirically supported cognitive engineering constructs. *Journal of cognitive engineering and decision making*, 2(2):140–160, 2008.

20 Lindsay Sanneman and Julie A Shah. The situation awareness framework for explainable ai (safe-ai) and human factors considerations for xai systems. *International Journal of Human–Computer Interaction*, 38(18-20):1772–1788, 2022.

21 Timo Speith, Barnaby Crook, Sara Mann, Astrid Schomäcker, and Markus Langer. Conceptualizing understanding in explainable artificial intelligence (XAI): an abilities-based approach. *Ethics Inf. Technol.*, 26(2):40, 2024.

22 Neville A Stanton, Paul M Salmon, Laura A Rafferty, Guy H Walker, Chris Baber, and Daniel P Jenkins. *Human factors methods: a practical guide for engineering and design*. CRC Press, 2017.

23 Stefano Teso and Kristian Kersting. Explanatory interactive machine learning. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019.

24 Xinru Wang and Ming Yin. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*, pages 318–328, 2021.

25 Cedric Waterschoot, Raciel Yera Toledo, Francesco Barile, and Nava Tintarev. With friends like these, who needs explanations? evaluating user understanding of group recommendations. In *UMAP (to appear)*, 2025.

26 Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 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*, FAT* '20, page 295–305, New York, NY, USA, 2020. Association for Computing Machinery.

Participants

- Elisabeth André
Universität Augsburg, DE
- Jaesik Choi
KAIST – Daejeon, KR
- Peter Clark
Allen Institute for AI – Seattle, US
- Elizabeth M. Daly
IBM Research – Dublin, IE
- Peter Flach
University of Bristol, GB
- Jasmina Gajcin
IBM Research – Dublin, IE
- Tobias Huber
TH Ingolstadt, DE
- Eda Ismail-Tsaous
bidt – München, DE
- Patricia Kahr
TU Eindhoven, NL
- Francesca Naretto
University of Pisa, IT
- Talya Porat
Imperial College London, GB
- Daniele Quercia
Nokia Bell Labs – Cambridge, GB
- Lindsay Sanneman
Arizona State University – Tempe, US
- Ute Schmid
Universität Bamberg, DE
- Kacper Sokol
ETH Zürich, CH
- Timo Speith
Universität Bayreuth, DE
- Wolfgang Stammer
TU Darmstadt, DE
- Simone Stumpf
University of Glasgow, GB
- Stefano Teso
University of Trento, IT
- Nava Tintarev
Maastricht University, NL