# Evaluating human understanding in XAI systems

**Davide Gentile**
University of Toronto
Toronto, ON M5S, CA
dgentile@mie.utoronto.ca

**Greg A. Jamieson**
University of Toronto
Toronto, ON M5S, CA
jamieson@mie.utoronto.ca

**Birsen Donmez**
University of Toronto
Toronto, ON M5S, CA
donmez@mie.utoronto.ca

*ACM CHI HCXAI Workshop, May 8–9, 2021, Online virtual conference*
© 2021 Copyright is held by the owner/author(s).
ACM ISBN 978-1-4503-6819-3/20/04.
DOI: https://doi.org/10.1145/3334480.XXXXXXX
*update the above block & DOI per your rights review confirmation (provided after acceptance)

## Abstract

Explanation in human-AI systems can provide the foundations for supporting joint decision-making and effective reliance. However, this desideratum depends on the relevant stakeholders understanding the AI's capabilities and the reason behind its outputs. In this position paper, we compare two approaches used to measure human understanding in XAI: proxy tasks, i.e., artificial tasks evaluating user ability to simulate the AI decision, and mental models, i.e., user internal representation of the structure and function of a given system. We argue that, although widely used, proxy tasks (i) can fail at evaluating the effectiveness of explanations in human-AI systems due to the absence of a realistic end-goal, and (ii) may not translate to system performance in actual decision-making tasks. We further propose that existing research in human factors and the social sciences can guide mental model-based evaluations of human understanding with realistic decision-making tasks. Given the objective of providing explanations that facilitate decision-making tasks, we conclude by arguing that a rigorous evaluation of explainable systems needs to integrate a quantitative assessment of users' prior knowledge of AI systems.

## Author Keywords

Explainable AI; mental models; human-AI systems; evaluation methods.

## CSS Concepts

- **Human-centered computing~Human computer interaction (HCI)**

## Introduction

Quantitatively evaluating human understanding of XAI systems is a current challenge in XAI research [9]. Despite the proliferation of quantitative evaluation metrics proposed in the literature, there is still no clear consensus on how to measure the impact of explanations on human understanding. This becomes particularly challenging if the target human subjects have no technical expertise in AI systems. After an examination of recent relevant studies with non-expert users (i.e., end-users of AI systems with no technical expertise), we observed that quantitative evaluation of human understanding of AI systems tend to fall into two main approaches. The first evaluation approach involves the use of proxy tasks, i.e., simplified tasks where human understanding is evaluated on the basis of successful prediction of the decision or the decision boundaries of the AI through the assistance of explanations [2][3][8]. The second evaluation approach involves the elicitation and measurement of users' mental models of an AI system, which can be defined as humans' internal representations of the structure and function of a given system [11] and can be elicited and analyzed in the context of evaluations of human-XAI systems [6]. After describing these two approaches in the following section, we identify limitations for proxy task-based evaluations and argue for the appropriateness of mental model-based evaluations. We further note that, although necessary, this distinction is not sufficient to measure the impact of explanations on human understanding. Humans' prior general knowledge of AI systems can affect how they leverage explanations during the decision-making process. Thus, we argue for the need of including an assessment of this prior knowledge (e.g., through an AI knowledge scale) into the evaluation design of XAI systems.

## Evaluation metrics

### *Proxy tasks*

First proposed by Doshi-Velez and Kim [3], proxy tasks are based on the idea of human understanding as the ability of users to simulate AI behavior. In their review, they describe two human-subject tasks designed to assess what is defined as simulatability of an AI/ML model. The first task is termed *forward simulation* and assesses human ability to predict an AI's output given input data and an explanation of the AI's behavior. The second task is termed *counterfactual simulation* and assesses human ability to use input data, an explanation, and an AI's output to predict the AI's output given a perturbation of the original input data. Proxy task-based evaluations have been adopted particularly in computer science research, where other types of proxy tasks have emerged, including evaluation of the impact of individual input attributes on the output [2] and verification of consistency between explanation and AI output [8]. However, forward and counterfactual simulations have been the most widely used proxy tasks across studies [2][3][5][8].

### *Mental models*

Studies on mental models constitute a large body of research in human factors and cognitive psychology, where they have been leveraged to study how people interact with intelligent systems and decision aids [6]. In human-computer interaction studies, human subjects' mental models have been elicited both qualitatively, for example through structured interviews [15] and quantitatively, for example through comprehension scores [7]. Nonetheless, non-expert humans' mental models are likely to deviate from designers' conceptual model of a system, and thus they could be incomplete or imprecise [4]. As noted by [13], research in mathematics education offers an example of how humans' mental models of an abstract concept can easily differ from the mathematical and logical properties of the same concept. Similarly, in the context of XAI, expecting non-expert users to build an accurate representation of the technical and logical properties of an AI system is unrealistic. In this paper, we adopt Greca and Moreira's perspective that mental models do not have to be accurate to serve their function [4], as long as they enable a level of understanding that allows humans to interact safely with an AI system (e.g., with the awareness that it can fail).

## Limitations of proxy task-based evaluations

The adoption of proxy task-based evaluations of human understanding in XAI systems is motivated by the fact that they can be conducted on lay humans, thus allowing for testing a potentially large subject pool without the need to compensate trained humans who are experts in AI or the domain under investigation [3]. We identified two main limitations associated with proxy task-based methods. The first limitation concerns the absence of a realistic end-goal in the interaction between humans and the XAI system, which makes the decision-making process not reflective of real-world scenarios. In [1], Buçinca et al. noted that proxy task-based evaluations require human-subjects to allocate their cognitive focus on the AI and the explanation provided, whereas in real decision-making tasks their focus is on their intended end-goal and they can choose "whether and how much to attend the AI." (p. 454). This limitation suggests the inefficacy of proxy task-based evaluations to measure whether AI explanations satisfy the goal of *effectiveness* as defined in [12] and [14] (i.e., explanations are effective if they help users make good decisions). The second limitation is potentially a consequence of the first: Measures of performance in proxy tasks may not be predictive of performance in more realistic decision-making tasks. Buçinca et al. [1] compared two explanation designs (i.e., inductive vs. deductive) in both a proxy task and an actual decision-making task and found that participant trust and satisfaction were higher for inductive explanations in the proxy task, while they were higher for deductive explanations in the actual decision-making task. Similarly, limitations of these methods were also noted in [2], where the authors called for replication in more realistic settings.

## Advantages of mental model-based evaluations

The limitations above do not apply to mental model-based evaluation of human understanding in XAI systems. While mental model-based evaluations can also be conducted on lay humans, they are rather motivated by the idea that "users of any computer system are not just passive consumers of information, but rather active partners who will form their mental model of the system as they make sense of it" [16]. We identified two advantages offered by mental model-based evaluations of human understanding in the context of XAI. The first advantage is that evaluating understanding through mental models allows for the parallel assessment of human understanding and accuracy in a decision-making task. This enables XAI researchers to not only observe the *effectiveness* of explanations as defined above, but also to measure any correlation between the *quality* of one's mental model and the respective *performance* on the decision-making task. Whether good mental models correspond to good performance in XAI remains an open question which gives the opportunity to address it empirically in actual decision-making tasks [6]. The second advantage is that mental models can reveal specific fallacies in one's internal representation of the mechanisms and behavior of an AI system, that cannot be necessarily captured by proxy tasks. A human-centered strategy to support understanding should begin with the humans' perspective. Incomplete or incorrect mental models offer the opportunity to follow changes in mental models throughout the explanation process [13] and guide the types of explanations needed to support understanding.

## Impact of prior knowledge on decision making

The advantages offered by mental models are not sufficient to conduct rigorous evaluations of explainable systems. One other challenge faced by the XAI research communities regards controlling for human subjects' knowledge of AI systems and familiarity with the application domain. Human subjects' prior knowledge can have an impact on their interpretation of the explanations and ultimately on their performance on the decision-making task. For example, participants' experience with machine learning (assessed through three levels of self-reported experience) has been found to have an impact on how they update their own mental models [10]. In addition, education level and technical literacy (i.e., familiarity with algorithms and programming experience) were also found to have an effect on accuracy in forward and counterfactual simulations [2]. These findings suggest that participants with different levels of prior knowledge are likely to have different explanation needs while interacting with a given XAI system. We propose that XAI researchers should integrate a quantitative assessment of users' existing knowledge of AI systems into the evaluation design, potentially in the form of a scale capturing participants' general knowledge of AI. Such a scale would need to be constructed and validated, as there are no such scales or tools to date.

## Conclusion

XAI researchers have taken different approaches to evaluate human understanding of XAI systems. While proxy tasks can be partly informative of human understanding, they can fail at reflecting how humans use explanations and make decisions in real world scenarios. Evaluating realistic decision-making tasks requires more time and resources, but ultimately explanations in human-AI systems should assist humans in making good decisions. A human-centered strategy should be formulated at the level of sociotechnical systems, and – although harder to translate into quantifiable measures - mental models offer a viable evaluation technique that accounts for the fallacies of human reasoning. However, we should also be mindful of the prior knowledge that human subjects bring with themselves into experimental settings, and keep in mind that this covariate can not only influence decision-making but also determine different explanatory needs across non-expert users operating with an AI system.

## References

[1] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (IUI '20). Association for Computing Machinery, New York, NY, USA, 454–464. DOI:https://doi.org/10.1145/3377325.3377498

[2] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. 2019. Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19).* Association for Computing Machinery, New York, NY, USA, Paper 559, 1–12. DOI:https://doi.org/10.1145/3290605.3300789

[3] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

[4] Ileana María Greca and Marco Antonio Moreira. 2000. Mental models, conceptual models, and modelling. International Journal of Science Education 22, 1 (2000), 1–11. https://doi.org/10.1080/095006900289976

[5] Peter Hase and Mohit Bansal. 2020. Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior? arXiv preprint arXiv:2005.01831.

[6] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. arXiv preprint arXiv:1812.04608.

[7] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on visual languages and human centric computing*, pp. 3-10. IEEE.

[8] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2019. "An evaluation of the human-interpretability of explanation." *arXiv preprint arXiv:1902.00006*.

[9] Zachary C. Lipton. 2018. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue 16, 3 (May-June 2018), 31–57. DOI:https://doi.org/10.1145/3236386.3241340

[10] Sean McGrath, Parth Mehta, Alexandra Zytek, Isaac Lage, and Himabindu Lakkaraju. 2020. When Does Uncertainty Matter?: Understanding the Impact of Predictive Uncertainty in ML Assisted Decision Making. arXiv preprint arXiv:2011.06167.

[11] Don A. Norman. 1983. Some observations on Mental Models. In *Mental Models*, Dedre Gentner and Albert L. Stevens (Eds.). Psychology Press, Chapter 1, 7–14.

[12] Ingrid Nunes and Dietmar Jannach. A systematic review and taxonomy of explanations in decision support and recommender systems. 2017. User Modeling and User-Adapted Interaction 27, no. 3 (2017): 393-444.

[13] Heleen Rutjes, Martijn Willemsen, and Wijnand IJsselsteijn. 2019. Considerations on explainable AI and users' mental models. In *CHI 2019 Workshop: Where is the Human? Bridging the Gap Between AI and HCI*. Association for Computing Machinery, Inc, 2019.

[14] Nava Tintarev and Judith Masthoff. 2011. Designing and evaluating explanations for recommender systems. In *Recommender systems handbook*, pp. 479-510. Springer, Boston, MA.

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

[16] Jennifer Wortman Vaughan and Hanna Wallach. 2020. A human-centered agenda for intelligible machine learning. *Machines We Trust: Getting Along with Artificial Intelligence*.