Helpful, Misleading or Confusing: How Humans Perceive Fundamental Building Blocks of Artificial Intelligence Explanations

Edward Small[†], **Yueqing Xuan**[†],
Danula Hettiachchi[†], and **Kacper Sokol**[†]

ARC Centre of Excellence for Automated Decision-Making and Society, School of Computing Technologies, RMIT University, Australia

edward.small@student.rmit.edu.au
yueqing.xuan@student.rmit.edu.au
danula.hettiachchi@rmit.edu.au
kacper.sokol@rmit.edu.au

Abstract

Explainable artificial intelligence techniques are developed at breakneck speed, but suitable evaluation approaches lag behind. With explainers becoming increasingly complex and a lack of consensus on how to assess their utility, it is challenging to judge the benefit and effectiveness of different explanations. To address this gap, we take a step back from sophisticated predictive algorithms and instead look into explainability of simple decision-making models. In this setting, we aim to assess how people perceive comprehensibility of their different representations such as mathematical formulation, graphical representation and textual summarisation (of varying complexity and scope). This allows us to capture how diverse stakeholders – engineers, researchers, consumers, regulators and the like – judge intelligibility of fundamental concepts that more elaborate artificial intelligence explanations are built from. This position paper charts our approach to establishing appropriate evaluation methodology as well as a conceptual and practical framework to facilitate setting up and executing relevant user studies.

Keywords

Evaluation; validation; human-centred; user study; artificial intelligence; machine learning; explainability; interpretability.

Motivation

It is often challenging to envisage how novel explainable artificial intelligence (XAI) and interpretable machine learning (IML) techniques will be perceived by explainees given their diverse skills, experiences and background knowledge [28]. The gap between design intention and actual reception of explanatory insights may sometimes be larger than expected – thus compromising their utility – especially when serving a lay audience [5, 12]. To alleviate such problems and enable better design and operationalisation of explainability systems, we should strive to understand the unique needs of diverse audiences as well as their specific interpretation of and satisfaction with different explanation types.

The availability of tools to peek inside opaque predictive models has ballooned in recent years [4], nevertheless it remains unclear what constitutes a "good" explanation [28]. To complicate matters further, a good explanation for one user may be unintelligible for another [11, 28]. Similarly, certain explanations may be easy to misinterpret or misunderstand [20], leading some users to overly rely on or misplace their trust in these insights [23] (akin to *negative comprehension* [11]). This can result in harmful explainers that appear faithful and trustworthy enough to convince a user of their truthfulness and utility while at the same time offering hollow or factually incorrect information [13].

Such scenarios motivate us to assess the intelligibility of *fundamental building blocks* found across XAI and IML tools, which tend to be complex and multifaceted sociotechnical systems tested predominantly as end-to-end explainers [30]. By explicitly accounting for how different demographics perceive distinct types and presentation modalities of explanations, our user study aspires to highlight the necessity of a comprehensive human-centred design and evaluation framework in this research space.

Objective

We aim to examine how users from diverse backgrounds perceive explanations of automated decision-making (ADM) systems. In particular, we intend to explore whether explanatory insights foster *warranted* (and justified) or *unwarranted* trust in predictive models [6, 10]. To this end, we propose three measures suitable for dedicated (online) user studies, hereafter referred to as the *3-C evaluation framework*.

Comprehension The gap between the information offered by an algorithmic explanation and the information understood by an explainee (i.e., new knowledge).

> *Positive* comprehension captures *known knowns* – the user knows what information an explanation offers – and *known unknowns* – the user knows what information it lacks.

> *Negative* comprehension encompasses *unknown knowns* – the user is unable to interpret the information communicated by an explanation – and *unknown unknowns* – the user is ignorant of the limitations of an explanation.

Confidence The degree of an explainee's certainty in their own understanding of an explanation.

Contentment The explainee's perception of how reliable and informative an explanation is.

Methodology

To ensure the simplicity of our study as well as generalisability of our findings, we investigate a comprehensive and representative subset of predictive models, explanation types and modalities thereof. We focus exclusively on explanations that are derived directly from the tested models to guarantee their full fidelity and truthfulness [22]. Specifically, we draw prototypical data-driven predictive models from across their distinct categories: linear, quadratic and the like from

$$f(\mathbf{x}) = 0.7x_1 - 0.3x_2$$
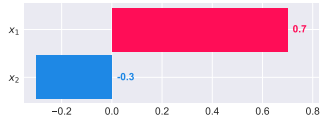
Figure 1: Mathematical representation of a model.

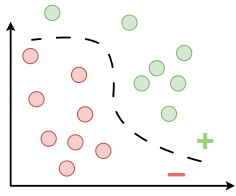**Figure 2:** Visualisation of feature influence/importance derived from a model (e.g., its parameters).

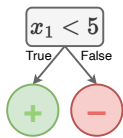**Figure 3:** Visualisation of a decision boundary used by a predictive model.

Figure 4: Visualisation of a decision tree structure.

Had your income been €10,000 more, you would have been awarded the loan.

Figure 5: Textual counterfactual explanation.

the *geometric* family; logistic regression from the *probabilistic* group; and (shallow) decision tree from the *logical* collection [7]. We look into *algorithmic explanations* that explicitly communicate the model's (mathematical) operation as well as *observational explanations* that capture diverse aspects of its predictive behaviour through different categories – *associations between antecedent and consequent*, *contrasts and differences*, and *causal mechanisms* – and types – *model-based*, *feature-based*, and *instance-based* – of explanatory insights [25]. Notably, this broad range of explanations accounts for their varied scope (local, cohort or global) and limitations (whether implicit or explicit). We present them through different media – (statistical or numerical) summarisation, visualisation, textualisation, formal argumentation, and a mixture thereof – testing both *static* and *interactive* modalities [27, 29].

For example, a geometrical model can be explained via its mathematical formulation (Figure 1), its parameter values (Figure 2), the visualisation of its decision surface within the desired space (Figure 3), exemplars sourced from both sides of its decision boundary, or what-if statements. Similarly, a logical model, e.g., a decision tree, can be visualised as a diagram (Figure 4) and explained with decision space partition, logical rules (root-to-leaf paths), feature importance, exemplars extracted from leaves as well as what-ifs and counterfactuals (Figure 5) derived directly from the model [26].

To better understand how the perception of different explanations varies across the participants, we envisage collecting basic demographic information such as age, gender, educational attainment and profession. Literacy is another important dimension that affects explanation comprehensibility [5]; while it is difficult to capture in a generic setting, we will adapt pre-existing scales [32] to assess explainees' *digital literacy* (ability to interact with technology and familiarity with artificial

intelligence techniques), *English proficiency* and *numeracy* (ability to interpret mathematical concepts).

The explainees' needs as well as the function, depth and scope of explanations must also be considered since XAI and IML tools are used by diverse stakeholders to different ends, e.g., engineers (technical expertise), consumers (no expertise) and auditors (limited technical knowledge) [1, 14]. Among these audiences, the general public may care only about the logic leading to a particular decision; regulators may need to access and assess the end-to-end functioning of a predictive system; and practitioners may require the same insight but with more technical depth. Evaluating the utility of explanations in view of varied needs, expectations and expertise can yield important findings that will benefit the designers and consumers of XAI and IML systems [18].

The within-subject online user study will be based on *linear*, *polynomial*, *logistic* and *decision tree* models, and a diverse set of explanations spanning model presentation (Figures 1 & 4), model summarisation (Figure 2), decision boundary visualisation (Figure 3) and textual counterfactuals (Figure 5). We will rely on familiar and unfamiliar domains (see the next section) to simulate background knowledge and lack thereof, e.g., loan application [31] and medical diagnosis [2]. For every predictive model and data domain, the participants will complete three tasks, each based on two explanations; in the process, they will answer targeted questions and provide feedback on the explanatory artefacts. The study workflow for each task is envisaged as follows:

1. Present an automated decision.
2. Offer its explanation using method A (Explanation 1).
3. Measure comprehension (Questions 1 & 3) and confidence (Question 2).
4. Measure contentment (Question 4) with user rating of the quantity and quality, communication, and reliability

Explanation 1: Loan application.

The loan request was rejected because of your income (€42,000).

Question 1: Comprehension.

Income was the strongest factor in this specific loan application.

a) <mark>True.</mark> b) False.
c) Can't say. d) Don't know.

Question 2: Confidence.

How confident are you in your previous answer?

1	2	3	4	5	6	7

Question 3: Comprehension.

Increasing the monthly loan repayment to €400 will change the automated decision.

a) True. b) False.
c) <mark>Can't say.</mark> d) Don't know.

Question 4: Contentment.

The explanation was easy to interpret.

1	2	3	4	5	6	7

of the information provided by the explanation.

5. Offer another explanation based on method B.
6. Re-evaluate the 3-Cs.

Case Studies

To ensure the *ecological validity* of our study, i.e., the generalisability of its findings, we will consider *real* and *fictitious* scenarios spanning *low-* and *high-stakes* domains. Because the individual perception of each scenario's impact can be subjective [8], we will ensure that one case study is always recognised as having less at stake than the other, albeit low-stakes cases may not necessarily be of low impact to all the users across the board [3]. When designing specific scenarios, we will draw inspiration from relevant application domains listed in the General Data Protection Regulation (GDPR) to ensure that they are relatable and align with real-life ADM systems [33]. Such a comprehensive approach will allow us to better understand the 3-Cs of XAI and IML tools in a variety of settings.

When designing the low- and high-stakes ADM case studies, we will either embed them in reality [15, 21] or narrate them in a fictitious context that is clearly disconnected from the real life [19]. An example of a real-world, high-stakes domain is a financial trading assistant powered by an ADM system; in this case the explainees risk incurring monetary loss if they misplace their trust in or misunderstand an insight output by the agent. Alternative high-stakes scenarios can be situated in a justice (e.g., pretrial bail), healthcare (e.g., administration of a life-altering medication) or job screening contexts. A real-world, low-stakes domain, on the other hand, may be based on using ADM tools to identify bird species or assist in playing board games. Similar case studies can be embedded in unfamiliar contexts by asking explainees to: follow recommendations of an ADM medical decision support tool to decide on a treatment of a sick extraterrestrial (high-

stakes), and vet outfits recommended to an extraterrestrial by an ADM personal stylist based on attire preferences and body type (low-stakes).

Both fictitious and real-world use cases come with pros and cons. By employing the former we can prevent the explainee's background knowledge and preferences from influencing their interaction with XAI and IML systems [19]. For example, user study participants cannot use their pre-existing medical expertise when dealing with a fictitious extraterrestrial healthcare problem, which forces them to rely exclusively on the information provided by an explanation. However, such a setting can compromise the ecological validity of our study since fictitious domains may be perceived as abstract and unrelatable, decreasing the participants' motivation and engagement. Despite the extraterrestrial healthcare setting being inherently high-stakes, user study participants may be unable to judge the severity and consequences of misunderstanding ADM systems given the unrealistic narrative. This phenomenon may be difficult to capture and implicitly impact our 3-Cs when comparing them between low- and high-stakes domains since the latter setting may not be perceived as such by some participants.

Real-world use cases are inevitably affected by the background knowledge of user study participants, which varies from person to person, is difficult to capture or account for, and may bias the measurement of our 3-Cs. Such a setting, however, ensures ecological validity of the study findings and maximises (active) engagement of the participants (especially for high-stakes domains) since they are on the receiving end of ADM systems [15]. Additionally, the influence of background knowledge can be reduced by substituting relevant terminology with made-up nomenclature, enabling our 3-C framework to capture the desired properties. While we will explore both real and fictitious settings in our preliminary

study, as it stands we lean towards real-world scenarios.

Discussion

The target audience of XAI and IML systems is at the heart of effective, human-understandable explanations of automated decisions. Each group of people has different requirements, preferences and expectations with respect to the explanations, fulfilling which secures their trust and confidence in the delivered information. In view of this, we envision to capture how the demographic characteristics of explainees may affect explanation comprehensibility and elucidate the anticipated variability through the lens of our 3-C framework. This will help us to analyse what types of explanations are preferred by laypersons and which modalities impose usability barriers.

In particular, we are interested in the manifestation of information overload – a situation in which explainees fail to process incoming information as it overwhelms their cognitive abilities. The trade-off between richness and utility of information varies to different degrees by explanation type and modality; for example, providing exhaustive textual explanations is likely to fail in communicating high-impact information, which in such scenarios can be easily overlooked. Exploring the acceptance threshold is therefore critical in advancing XAI and IML human-centred evaluation frameworks.

Explaining the overall operation of an automated decision-making system is just as important as doing so for an individual output. While the outcome of an "opaque" predictive model can be justified, such an explanation is fundamentally different from an inherently transparent model that is open to scrutiny [16, 22]. Within this purview, is there a point where transparency – allowing an agent to access the raw, unadulterated model – does not improve understanding [28]? We anticipate to answer this question with our 3-C framework.

Ensuring that explanations are equally interpretable across distinct demographics and populations is a formative step in advancing fairness of data-driven systems [17]. By collecting basic demographic information about the user study participants, we can assess if individuals from different groups perceive certain explanations as equally comprehensible and trustworthy [31]. Algorithmic bias is another concern. Predictive models are generally developed and trained in Global North [24], which may result in implicit biases and prevent these systems from generalising to underrepresented demographics who differ in how they handle cognitive information. Common considerations include circumstances such as a discrepancy in literacy and ability to interpret textual or visual semantics [9, 31].

Conclusion and Future Work

In this position paper, we set out to assess how differences in explainees' literacy and education attainment, among many other traits, shape their interpretation of algorithmic explanations. We developed a quantitative framework to measure these aspects from three distinct perspectives: comprehension, confidence and contentment. Our 3-C evaluation paradigm offers comprehensive and nuanced insights into how fundamental building blocks of XAI and IML systems are interpreted. The proposed approach is a first step towards a grounded evaluation methodology based on user studies that accounts for individual differences between explainees.

In future work, we will focus on quantitative assessment of explainee literacy by designing structured knowledge tests. We also plan to use real-life explainers across different domains, such as loan approval (high-impact) and animal identification (low-impact), spanning tabular (e.g., generic numerical and categorical characteristics) and sensory (e.g., specialised medical imaging) data.

References

[1] Vaishak Belle and Ioannis Papantonis. 2021. Principles and practice of explainable machine learning. *Frontiers in Big Data* (2021), 39.

[2] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": Uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. *Proceedings of the ACM on Human-computer Interaction* 3, CSCW (2019), 1–24.

[3] Keeley Crockett, Matt Garratt, Annabel Latham, Edwin Colyer, and Sean Goltz. 2020. Risk and trust perceptions of the public of artificial intelligence applications. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[4] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. 2018. Explainable artificial intelligence: A survey. In *41st international convention on information and communication technology, electronics and microelectronics (MIPRO)*. IEEE, 0210–0215.

[5] Upol Ehsan, Samir Passi, Q Vera Liao, Larry Chan, I Lee, Michael Muller, Mark O Riedl, and others. 2021. The who in explainable AI: How AI background shapes perceptions of AI explanations. *arXiv preprint arXiv:2107.13509* (2021).

[6] Andrea Ferrario and Michele Loi. 2022. How explainability contributes to trust in AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1457–1466.

[7] Peter Flach. 2012. *Machine learning: The art and science of algorithms that make sense of data*. Cambridge university press.

[8] Nicole Gillespie, Steven Lockey, Caitlin Curtis, Javad Pool, and Ali Akbari. 2023. Trust in artificial intelligence: A global study. (2023).

[9] Janet Holmes. 1997. Women, language and identity. *Journal of sociolinguistics* 1, 2 (1997), 195–223.

[10] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM conference on Fairness, Accountability, and Transparency*. 624–635.

[11] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–14.

[12] Bernard Keenan and Kacper Sokol. 2023. Mind the Gap! Bridging Explainable Artificial Intelligence and Human Understanding with Luhmann's Functional Theory of Communication. *arXiv preprint arXiv:2302.03460* (2023).

[13] Justin Kruger and David Dunning. 1999. Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology* 77, 6 (1999), 1121.

[14] Samuli Laato, Miika Tiainen, A.K.M. Najmul Islam, and Matti Mäntymäki. 2022. How to explain AI systems to end users: A systematic literature review and research agenda. *Internet Research* 32, 7 (2022), 1–31.

[15] Benedikt Leichtmann, Christina Humer, Andreas Hinterreiter, Marc Streit, and Martina Mara. 2023. Effects of explainable artificial intelligence on trust and human behavior in a high-risk decision task. *Computers in Human Behavior* 139 (2023), 107539.

[16] Zachary C. Lipton. 2018. The mythos of model interpretability. *Commun. ACM* 16, 3, Article 30 (Jun 2018), 30:31–30:57 pages.

[17] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.

[18] Christian Meske, Enrico Bunde, Johannes Schneider, and Martin Gersch. 2022. Explainable artificial intelligence: Objectives, stakeholders, and future research opportunities. *Information Systems Management* 39, 1 (2022), 53–63.

[19] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2018. How do humans understand explanations from machine learning systems? An evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1802.00682* (2018).

[20] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[21] Vincent Robbemond, Oana Inel, and Ujwal Gadiraju. 2022. Understanding the role of explanation modality in AI-assisted decision-making. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*. 223–233.

[22] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* 1, 5 (2019), 206–215.

[23] Tim Schrills and Thomas Franke. 2020. Color for characters-effects of visual explanations of AI on trust and observability. In *Artificial Intelligence in HCI: First International Conference, AI-HCI 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings 22*. Springer, 121–135.

[24] Divya Siddarth, Daron Acemoglu, Danielle Allen, Kate Crawford, James Evans, Michael Jordan, and E Weyl. 2021. How AI fails us. *Justice, Health, and Democracy Impact Initiative & Carr Center for Human Rights Policy* (2021). The Edmond & Lily Safra Center for Ethics, Harvard University.

[25] Kacper Sokol and Peter Flach. 2020a. Explainability fact sheets: A framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 56–67.

[26] Kacper Sokol and Peter Flach. 2020b. LIMEtree: Consistent and faithful surrogate explanations of multiple classes. *arXiv preprint arXiv:2005.01427* (2020).

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

[28] Kacper Sokol and Peter Flach. 2021. Explainability is in the mind of the beholder: Establishing the foundations of explainable artificial intelligence. *arXiv preprint arXiv:2112.14466* (2021).

[29] Kacper Sokol and Peter A Flach. 2018. Glass-Box: Explaining AI decisions with counterfactual statements through conversation with a voice-enabled virtual assistant. In *IJCAI*. 5868–5870.

[30] Kacper Sokol, Alexander Hepburn, Raul Santos-Rodriguez, and Peter Flach. 2022. What and how of machine learning transparency: Building bespoke explainability tools with interoperable algorithmic components. *Journal of Open Source Education* 5, 58 (2022), 175.

[31] Niels van Berkel, Jorge Goncalves, Daniel Russo, Simo Hosio, and Mikael B Skov. 2021. Effect of information presentation on fairness perceptions of machine learning predictors. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[32] Alexander J.A.M. van Deursen, Ellen J Helsper, and Rebecca Eynon. 2016. Development and validation of the Internet Skills Scale (ISS). *Information, communication & society* 19, 6 (2016), 804–823.

[33] Paul Voigt and Axel von dem Bussche. 2017. The EU general data protection regulation (GDPR). *A Practical Guide, 1st Ed., Cham: Springer International Publishing* 10, 3152676 (2017), 10–5555.