Human-AI Interaction Paradigm for Evaluating Explainable Artificial Intelligence

Matija Franklin[1][0000-0003-1846-8907] and David Lagnado[1]

[1] University College London, London WC1E 6BT, UK
matija.franklin@ucl.ac.uk

Abstract. This article seeks to propose a framework and corresponding paradigm for evaluating explanations provided by explainable artificial intelligence (XAI). The article argues for the need for evaluation paradigms – different people performing different tasks in different contexts will react differently to different explanations. It reviews previous research evaluating XAI explanations while also identifying the main contribution of this work – a flexible paradigm researchers can use to evaluate XAI models, rather than a list of factors. The article then outlines a framework which offers causal relationships between five key factors – mental models, probability estimates, trust, knowledge, and performance. It then outlines a paradigm consisting of a training, testing and evaluation phase. The work is discussed in relation to predictive models, guidelines for XAI developers, and *adaptive explainable artificial intelligence* - a recommender system capable of predicting what the preferred explanations would be for a specific domain-expert on a particular task.

Keywords: Explainable AI, Human-AI Interaction, Evaluation.

1 Introduction

Research in Deep Learning (DL) – a large subfield of Machine Learning (ML) – has researched and developed Deep Neural Network (DNN) models which are capable of high-end performance on a range of complex tasks [1]. A significant number of deep neural network models are uninterpretable black-boxes, typically resulting in less trust from users [2]. This lack of interpretability and trust can result in negative outcomes – people might use an AI that errs, or not use an AI that could increase the likelihood of a desired outcome. To address these issues, Explainable Artificial Intelligence (XAI) research seeks to build method for explaining the behaviour of black-box models in human-understandable terms [3].

This raises the psychological question of what is a human-understandable explanation, and how to measure people's reactions to different explanations. In 2016, DARPA launched the Explainable AI (XAI) program [4]. It aimed to 1) produce more explainable models, 2) design better explanation interfaces, and 3) understand the psychological requirements for effective explanations. This paper aims to tackle DARPA's third aim by utilizing methods from cognitive and behavioural science in order to develop a framework for evaluating XAI methods from a user's perspective. It also describes a

corresponding human-AI interaction (HAII) paradigm with testable and operationalized measures for comparing human-XAI to human-AI (and no-AI) teams. Finally, it will outline how the paradigm can generate data that allows for optimizing the distribution of explanations to the right person for the right task.

1.1 Different explanations for different individuals in different contexts

A large number of XAI methods have been researched, developed and deployed to improve explainability in DNN models [3]. As it stands it is not clear which method is better for a given person at a specific time in a certain context. This is further complicated by the fact that there are different ways in which one can measure an explanations benefit to a person.

Feature importance methods (i.e., saliency methods) provide scores that show the importance of a feature (e.g., word vector or pixel) to the AI's decision [5]. Explanations from this group of methods can either be local or global [6]. *Local explanation* methods, such as LIME (Local Interpretable Model Agnostic Explanations), assign a numeric measure of importance to an input variable (i.e., the weighting it has in relation to the outcome variable) [7]. *Global explanations*, such as SHAP (SHapley Additive exPlanations) models, provide a numeric measure of importance to each of the input variables on the model's output [8]. Further, *Concept-based explanations* aim to explain a models output using pre-defined or auto-discovered sets of human-understandable concepts [9]. Finally, *counterfactual explanations outline* what the outcome of the model could have been had input to a model been changed in a certain way [10].

The current evidence does not provide a universal account of why different XAI methods would be better for a certain person performing a task at a given time. The way explanation methods are distributed from an AI to a human will thus not be optimal, reducing the quality of HAII. In human-to-human interaction, Theory of Mind is central to a person being able to provide an explanation to another person or to explain another person's behaviour [11]. Predictive models of how people react to explanations are thus needed.

Not optimising for what XAI methods are used will lead to at least three broader issues. First, explanations might obscure more information than it reveals. In psychology, this is referred to as *information overload* – receiving too much information or more specifically when the "amount of input to a system exceeds its processing capacity" [12]. This phenomena holds even when all of the information is task-specific. Recent research has found information overload effects in response to explanations provided by XAI methods [13]. The more detailed explanations was found to be less useful and trustworthy than the less information-rich explanation. Further, there are individual differences in how people react to more information rich explanations; specifically, people with backgrounds in AI preferred them to people who did not have backgrounds in AI [14]. Making normative predictions, such as that more information will result in better decisions, is thus not possible.

Second, explanations are persuasive and will thus influence behaviour and preference in currently unpredictable ways. Current research suggests that an XAI compared to an AI tool can produce a greater behaviour change [15, 16]. Researchers have also

developed XAI methods which can generate misleading explanations that are capable of both increasing trust towards the AI whilst misleading domain experts [17]. Understanding the direction of the behaviour change, whether it is desirable, and how this will differ from one XAI method to another is thus essential. This is especially relevant given the bidirectional causal relationship between preference and behaviour [18], allowing influential AI systems to change preference by changing behaviour [19, 20].

Finally, explanations can produce different outcomes to specific tasks and user groups [21]. A systematic review of 137 articles on XAI in different application domains and tasks reveals certain patterns [22]. First, most current research has been directed towards safety-critical domains. Second, visual explanations are on average more acceptable to end-users. Finally, studies have mostly been directed at experts users, with more research needed for how general users react to explanations. Altogether, there is a need for a universal account capable of predicting how and why different users react to different XAI methods in different application domains. This raises the question of how to evaluate people's reactions towards different XAI explanations.

1.2 XAI Evaluation

Current research aiming to evaluate a person's reaction to an explanation provided by an XAI method reveals some initial patterns that can serve as a foundation and inspiration for a broader HAII paradigm for evaluating XAI. A systematics literature review of 241 papers explored how the validity and utility of explanations have been evaluated by the authors of those XAI methods [23]. Most studies lacked evaluations or conducted user studies in simple scenarios. The results show that in 32% of the research papers there was no attempt at any type of evaluation. Furthermore, 59% of the research papers conducted a user study evaluating the usefulness (i.e., the increase to performance) of the explanation (with a small minority also evaluating user trust towards the AI system). A final 9% used an algorithmic evaluation, not involving any empirical user research.

Studies directed at evaluating XAI have been conducted but they tend to use simplistic proxy ("toy tasks"), third-person vignettes, or do not consider any specific tasks at all. These studies nevertheless provide useful information on the variables one needs to consider when evaluating people's reactions the explanations. Users want to understand how and why an AI system makes predictions [24]. Different explanations will result in variations of people's performance on a task [25, 26]. Further, explanations influence people's trust in and understanding of an AI, but have no influence on people's perceptions of fairness and their general attitudes towards the AI [13]. However, this is not always the case for all explanations. For example, feature importance explanations do not always increase trust, understanding and performance [27].

Other researchers have proposed frameworks and taxonomies which outline the metrics and variables researchers should consider when evaluating explanations provided by XAI [28, 29]. For instance, [30] propose key measurements such as user satisfaction, trust, and performance. A crucial contribution to the present paper is that it offers a framework which corresponds to a HAII paradigm which can been deployed towards different people and application domains.

2 Framework

Based on how explanations have influenced people in previous research, our framework posits that from the perspective of a receiver, an effective explanation needs to: 1) provide knowledge, 2) be trustworthy, 3) be useful, 4) update the receiver's estimation about the probability of events occurring, and 5) change the receiver's mental model (See Fig. 1). The causal relationship between each factor is proposed in Fig. 1, with an arrow pointing from one factor to another suggesting that the first factor causally influenced the other, and bidirectional arrows suggesting a bidirectional causal relationship.

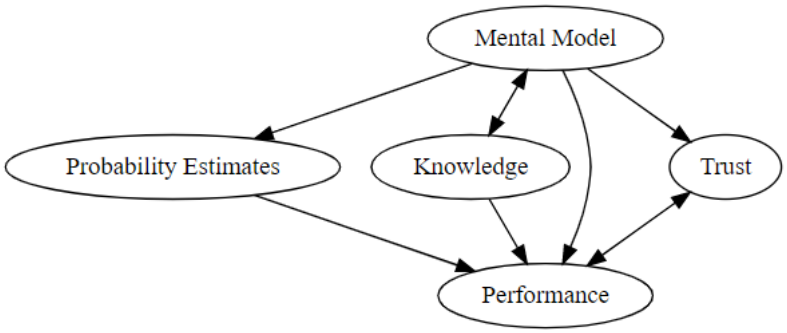

Fig. 1. Framework of key metrics for an effective XAI explanation

Knowledge. An effective explanation provides knowledge to its users. An AI without an explanation allows people to make inferences from the AI's decisions and detect patterns, as shown in [31]. The learning from XAI is more direct. This learning could be *procedural*, measured as a change in a certain ability, or *semantic*, measured as an increase in factual knowledge [32].

Trust. An effective explanation is trustworthy. [33] propose that trust is a multidimensional concept, with one being able to trust another's performance (i.e., the extent to which someone is reliable and/or capable) or morals (i.e., the extent to which someone is sincere and ethical). Trust can also be operationalized in multiple ways [30]. Foremost, trust is predictive of whether people choose to use an AI tool at all, and thus can be measured through adoption. Trust can also be measured in terms of the alignment between an AI's suggestions and a user's decisions. Finally, trust can be measured with self-reported measures, of which there are many (see [34]).

Performance. An effective explanation positively impacts a user's performance i.e., increases accuracy. Measuring performance in a task is highly context-specific [30]. One can also identify whether explanations increase performance at the upper end, or increase performance through reducing the frequency of mistakes.

Probability Estimates. An effective explanation updates the receiver's estimation about the probability of events occurring. Specifically, this can be operationalized as a measure of confidence or certainty in the users' own predictions [35]. It thus serves as a measure of people's estimates of uncertainty.

Mental Model. Mental models are representations of how a person understands some system. An effective explanation changes the receiver's mental models about the task, the broader application domain, and the AI. People are able to infer causal structures from explanations [36]. Explanations establish the presence and change the direction of intuited cause and effect relationships between different factors. This in turn influences other aspects of the user's cognition (See Fig. 1).

3 Paradigm

The outlined paradigm can serve as a foundation for research seeking to evaluate XAI explanations. It is intentionally flexible, allowing researchers to pick particular measures as they see fit. The HAII paradigm consists of at least three experimental groups (See Fig. 2). The XAI group consists of participants performing the task with the help of an AI that provides them with explanations. There can be multiple groups if multiple XAI methods are being evaluated. Participants in the AI group perform the task with the help of an AI. The control group contains participants performing the same task without receiving help. In the paradigm participants are given a task, and in the XAI and AI group an AI advisor, which they too to do to the best of their ability.

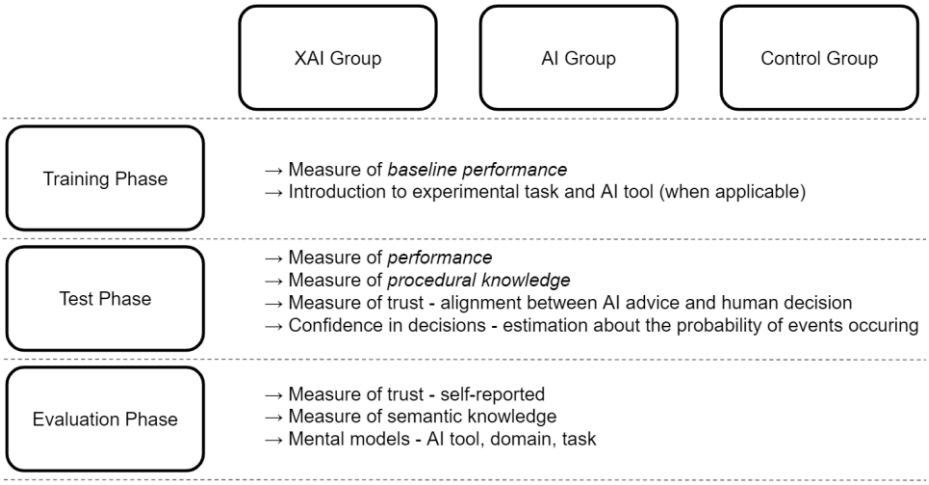

Fig. 2. Experimental groups and phases of the HAII paradigm for evaluating XAI.

Performance is measured as changes in people's task-related accuracy, and is context dependent. Knowledge is measured as people's performance in the absence of the AI tool (i.e., procedural knowledge) or with questions related to the task and domain (i.e.,

semantic knowledge). Trust is measured as the alignment between the AI's advice and the human's decision (i.e., more alignment equals more trust) and with a selected self-report measure (see [34]). Probability estimates are measured as the probabilities users attach to how confident they are in their own decisions for future task performance (on a 0-100 scale). Finally, mental models are measured using a nearest neighbor task, where participants select the explanation or diagram that best fits their beliefs, or with concept mapping, in which users create a diagram which outlines their knowledge [see 30]. Finally, researchers can also measure participants' socio-demographic background or domain-knowledge to identify individual differences in participants' responses.

The paradigm consists of three phases - training, test and evaluation (See Figure 2). In the training phase, participants are introduced to the experimental task, and their baseline performance and probability estimates are measured for the particular task.

In the test phase, participants in the XAI and AI groups engage in the same task with the aid of their AI tool. The task behaviour that the participants display in this phase of the study serves as a measure of their performance. A higher alignment between the user's decision and AI's prediction implies more trust. At various random points within this phase of the paradigm participants will be asked to make probability statements of how confident they are in their own decisions. Participants in all groups will then be given the same task but without the help of an AI. The task related behaviour here serves as a measure of the participants procedural knowledge.

Finally, in the evaluation phase, participants will be asked a series of questions related to trust towards AI, semantic knowledge, as well as their mental models about the task, domain and AI. They will also be asked questions that can allow researchers to identify individual differences.

4 Discussion

This articles specified a framework and corresponding paradigm which can inform research seeking to evaluate XAI explanations. Apart from serving as a measure of an explanation's effectiveness, the paradigm can generate data that allows for the development of predictive models. Such models could predict which explanations are appropriate for different domains, tasks, and individuals (e.g., seniority or capacity), as well as the appropriate ordering of tasks. The generated insights can be useful for optimizing the deployment of XAI methods.

This predictive framework can provide guidelines for XAI developers. Researchers have previously drawn from psychology to propose frameworks for building XAI models [37]. Such frameworks have been successful when evaluated. A predictive framework could preempt people's reactions, thus guiding XAI model development.

By collecting data with this paradigm, research can work towards developing an *Adaptive Explainable Artificial Intelligence* - a recommender system capable of predicting what the preferred explanations would be for a specific domain-expert on a particular task. This would involve building a simple, interpretable model, which would be given data collected from the paradigm. The model would be capable of prioritization - recommending the right XAI method to the right person for the task at hand.

Achieving this with a simple model is possible given that the data contains high level, well-understood variables. Further, by distilling the data through the paradigm, a simple model can also produce salient explanations of its own process by default — a kind of *meta-interpretability*.

References

1. Shahroudnejad A. A survey on understanding, visualizations, and explanation of deep neural networks. arXiv preprint arXiv:2102.01792. (2021).
2. Miller T. " But why?" Understanding explainable artificial intelligence. XRDS: Crossroads, The ACM Magazine for Students, 25(3):20-5. (2019).
3. Molnar C. Interpretable machine learning. Lulu. Com. (2020).
4. Gunning D, Aha D. DARPA's explainable artificial intelligence (XAI) program. AI magazine, 40(2):44-58. (2019).
5. Bhatt U, Xiang A, Sharma S, Weller A, Taly A, Jia Y, Ghosh J, Puri R, Moura JM, Eckersley P. Explainable machine learning in deployment. InProceedings of the 2020 conference on fairness, accountability, and transparenc, pp. 648-657. (2020)
6. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee SI. From local explanations to global understanding with explainable AI for trees. Nature machine intelligence, 2(1):56-67. (2020).
7. Lee E, Braines D, Stiffler M, Hudler A, Harborne D. Developing the sensitivity of LIME for better machine learning explanation. InArtificial Intelligence and Machine Learning for Multi-Domain Operations Applications, Vol. 11006, p. 1100610. (2019).
8. Lubo-Robles D, Devegowda D, Jayaram V, Bedle H, Marfurt KJ, Pranter MJ. Machine learning model interpretability using SHAP values: Application to a seismic facies classification task. InSEG International Exposition and Annual Meeting. (2020).
9. Kazhdan D, Dimanov B, Jamnik M, Liò P, Weller A. Now you see me (CME): concept-based model extraction. arXiv preprint arXiv:2010.13233. (2020).
10. Verma S, Dickerson J, Hines K. Counterfactual explanations for machine learning: A review. arXiv preprint arXiv:2010.10596. (2020).
11. Shvo M, Klassen TQ, McIlraith SA. Towards the role of theory of mind in explanation. InInternational Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems, pp. 75-93. (2020).
12. Sutcliffe KM, Weick KE. Information overload revisited. In The Oxford handbook of organizational decision making. (2009).
13. Ssebandeke A, Franklin M, Lagnado D. Explanations that backfire: Explainable artificial intelligence can cause information overload. Unpublished Manuscript. (Submitted 2022).
14. Ehsan U, Passi S, Liao QV, Chan L, Lee I, Muller M, Riedl MO. The who in explainable AI: how AI background shapes perceptions of AI explanations. arXiv preprint arXiv:2107.13509. (2021).
15. Dragoni M, Donadello I, Eccher C. Explainable AI meets persuasiveness: Translating reasoning results into behavioral change advice. AI in Medicine, 105:101840. (2020).
16. Donadello I, Dragoni M, Eccher C. Explaining reasoning algorithms with persuasiveness: a case study for a behavioural change system. InProceedings of the 35th Annual ACM Symposium on Applied Computing, pp. 646-653. (2020).
17. Lakkaraju H, Bastani O. " How do I fool you?" Manipulating User Trust via Misleading Black Box Explanations. InProceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp. 79-85. (2020).

18. Ariely D, Norton MI. How actions create–not just reveal–preferences. Trends in cognitive sciences, 12(1):13-6. (2008).
19. Ashton H, Franklin M. The problem of behaviour and preference manipulation in AI systems. InThe AAAI-22 Workshop on Artificial Intelligence Safety (SafeAI 2022). (2022).
20. Franklin M, Ashton H, Gorman R, Armstrong S. Recognising the importance of preference change: A call for a coordinated multidisciplinary research effort in the age of AI. AAAI-22 Workshop on AI For Behavior Change. (2022).
21. Tomsett R, Braines D, Harborne D, Preece A, Chakraborty S. Interpretable to whom? A role-based model for analyzing interpretable machine learning systems. arXiv preprint arXiv:1806.07552. (2018).
22. Islam, M. R., Ahmed, M. U., Barua, S., & Begum, S. (2022). A systematic review of explainable artificial intelligence in terms of different application domains and tasks. *Applied Sciences*, *12*(3), 1353.
23. Anjomshoae S, Najjar A, Calvaresi D, Främling K. Explainable agents and robots: Results from a systematic literature review. In18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), pp. 1078-1088. (2019).
24. Liao QV, Gruen D, Miller S. Questioning the AI: informing design practices for explainable AI user experiences. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1-15. (2020).
25. Lage I, Chen E, He J, Narayanan M, Kim B, Gershman S, Doshi-Velez F. An evaluation of the human-interpretability of explanation. arXiv preprint arXiv:1902.00006. (2019).
26. Narayanan M, Chen E, He J, Kim B, Gershman S, Doshi-Velez F. How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation. arXiv preprint arXiv:1802.00682. (2018).
27. Kindermans PJ, Hooker S, Adebayo J, Alber M, Schütt KT, Dähne S, Erhan D, Kim B. The (un) reliability of saliency methods. InExplainable AI: Interpreting, Explaining and Visualizing Deep Learning, pp. 267-280. (2019).
28. Chromik M, Schuessler M. A Taxonomy for Human Subject Evaluation of Black-Box Explanations in XAI. ExSS-ATEC@ IUI. (2020).
29. Sperrle F, El-Assady M, Guo G, Chau DH, Endert A, Keim D. Should we trust (x) AI? Design dimensions for structured experimental evaluations. arXiv preprint arXiv:2009.06433. (2020).
30. Hoffman RR, Mueller ST, Klein G, Litman J. Metrics for explainable AI: Challenges and prospects. arXiv preprint arXiv:1812.04608. (2018).
31. Peltola T, Celikok MM, Daee P, Kaski S. Modelling user's theory of ai's mind in interactive intelligent systems. arXiv preprint arXiv:1809.02869. (2018).
32. Berliner DC, Calfee RC. Handbook of educational psychology. Routledge. (2013).
33. Malle BF, Ullman D. A multidimensional conception and measure of human-robot trust. In Trust in Human-Robot Interaction, pp. 3-25. (2021).
34. Kaur D, Uslu S, Rittichier KJ, Durresi A. Trustworthy Artificial Intelligence: A Review. ACM Computing Surveys (CSUR), 55(2):1-38. (2022).
35. Tversky A, Kahneman D. Causal schemas in judgments under uncertainty. Progress in social psychology, 1:49-72. (2015).
36. Kirfel L, Icard T, Gerstenberg T. Inference from explanation. Journal of Experimental Psychology: General. (2021).
37. Wang D, Yang Q, Abdul A, Lim BY. Designing theory-driven user-centric explainable AI. In Proceedings of the 2019 CHI conference on human factors in computing systems, pp. 1-15. (2019).