Explanations of AI Are Greater than Explainable AI

Linus HOLMBERG [a,b,1] and Maria RIVEIRO [a]
[a] *School of Engineering, Jönköping University*
[b] *Department of Computer and Information Science, Linköping University*
ORCiD ID: Linus Holmberg https://orcid.org/0000-0002-0020-1756, Maria Riveiro https://orcid.org/0000-0003-2900-9335

Abstract. The eXplainable AI (XAI) community has traditionally focused on algorithm-centric methods to facilitate understanding. However, insights from related fields suggest alternative strategies for making AI systems understandable to users. In this positional paper, we argue that methods used in human-robot interaction (HRI) and human-computer interaction (HCI) can effectively support user understanding—often without relying on traditional XAI techniques, especially for end-users with limited AI expertise. Examples include user onboarding and communicating system states. While XAI methods can help generate such explanations, they are not always necessary, as users have diverse informational needs. We argue that explanations of AI extend beyond what is usually considered by the XAI community and that the XAI researchers would benefit from looking at how closely related fields make complex systems understandable.

Keywords. Explainable AI, Human-AI Interaction, Explanations, Human-Robot Interaction, Human-Computer Interaction

1. Introduction

The overarching aim of explainable AI (XAI) is to make AI systems understandable [1,2,3,4,5,6,7]. Historically, the XAI community has focused on algorithm-centric methods to elucidate the inner workings of AI systems [1,2,8,9], assuming that improving interpretability of algorithms inherently enhances users' understanding. While these methods are valuable for developers and researchers, they often fail to address the needs of end-users who interact with these systems in everyday contexts [1,10]. When interacting with any system, be that a microwave oven, a car, or an AI system, for most users, information on the system's internal mechanisms is neither necessary nor desired; instead, they just want the system to behave predictably and know what to do if it does not.

The new wave of XAI emerged in response to increasingly complex and opaque models [3,11]. As a result, much of the attention has thus been on creating methods that help interpret these models. However, this comes with an assumption that essential understanding requires opening the "black box" of such models. Still, this may not align with the information needs of all stakeholders [1,9].

[1] Corresponding Author: Linus Holmberg, linus.holmberg@ju.se

In this paper, we distinguish between explainability and explanations. Explainability is the degree to which something is explainable. In the context of AI, explainability methods are used to elucidate information that makes an AI model more readily understandable. Meanwhile, an explanation is the communication of a piece of information that makes something understandable. Indeed, traditional explainability methods can be beneficial in elucidating such information.

We argue that explanations of AI are more than what is usually considered by the XAI community, and facilitating understanding of AI systems can be done using insights from closely related disciplines such as Human-Computer Interaction (HCI) and Human-Robot Interaction (HRI). In this paper, we exemplify with a non-exhaustive number of methods, such as user onboarding and communicating the system's state or beliefs, how non-algorithm-centric methods can be sufficient to facilitate understanding of (some) AI systems and support successful interactions with their users.

As mentioned, one of the overarching aims of XAI is to make AI systems' output and behavior *understandable* to people. Like any aim, this could be accomplished through various methods. Yet, the methods used within the XAI community have predominantly been algorithm-centric [5,8,10,12,13], providing answers to very few questions the users of these systems normally ask. As there is no apparent reason why methods to facilitate understanding of AI systems should differ significantly from methods for understanding other complex systems, we call for the XAI community to look into how other closely related fields, such as HCI, HRI, and User Experience (UX), facilitate user understanding of systems.

2. Explainable AI

The efforts to explain AI behavior can be traced back to the era of expert systems. The field re-emerged in 2015 through a U.S. Defense Advanced Research Projects Agency (DARPA) project as a response to the increase in complex and uninterpretable models, such as Deep Neural Networks (DNNs) and Support Vector Machines (SVMs) [3,11].

One of the fundamental goals of XAI is to ensure that people can *understand* AI-systems' output or behavior [1,2,3,4,5,6,7]. The primary approach to promote understanding has been to create methods to elucidate the inner workings of AI systems [2,3].

Many attempts have been made to categorize or create taxonomies within the XAI field (e.g., [2,14,15,16,17]). AI models that are explainable without an external model interpreting them are considered transparent models [17]. Whereas AI models that cannot be directly interpretable are called black-box models [1,15]. XAI methods that are used to interpret (often black-box) AI models after training are considered post-hoc methods [7,14,17]. Examples of some highly influential post-hoc methods for explanations are LIME [18], SHAP [19], and GradCAM [20,21]. These methods aim to make model outputs understandable through feature importance or visualizing attention maps.

The recent surge in XAI research emerged as a response to the increase in complex and uninterpretable models [3]. This initially led to XAI methods designed mainly to assist model developers or researchers in examining the AI models [1,9,22]. As a reaction to the algorithm- or tech-centric focus, multiple scholars started arguing for a more human or user-centered perspective for XAI [1,6,8,12].

2.1. User-Centered Explainable AI

One of the main reasons that led to the user-centered reaction was the realization that different users need and want different kinds of explanations of AI systems depending on factors such as background, goals, and context [1,7,17,23,24]. Users are sometimes categorized into different groups and referred to as stakeholders [1,17,23,25]: model developers, business owners or administrators, decision-makers, impacted groups, and regulatory bodies are some of the commonly mentioned stakeholders [1,17,23]. Even though this categorization lacks granularity, it can be useful to identify the diversity of needs in explanations by users.

The growing interest in tailoring explanations to specific stakeholder needs has led to several reviews, some focusing on the HCI aspects of XAI [26,27,28,29,30]. Mueller et al. [29] review four decades of user-centered XAI, organizing explanation types by reasoning complexity and proposing empirically grounded design principles. Abdul et al. [26] highlight historical trends and the neglect of human-centered explanations, while Mohseni et al. [28] categorize design goals and evaluation methods, offering user-specific guidelines. Ali et al. [30] provide a broad overview of XAI techniques and tools, emphasizing user-focused approaches. Wang et al. [27] propose a road-map linking stakeholder needs to methods, underscoring the importance of clarifying intended stakeholders and addressing current gaps in XAI research.

The reviews indicate a growing community interest in addressing the actual needs of users interacting with AI systems. This paper focuses on end-users—those who directly interact with AI systems [3], such as physicians supported by AI in decision-making or warehouse workers operating in the same space as robots. The explanations that end-users need differ from the explanations, for instance, developers or regulatory bodies need. Most end-users probably do not care (or need to understand) about the underlying mechanisms of the system they are interacting with. In the same way, they do not need to understand the underlying mechanisms of their microwave oven. Instead, they need explanations that support successful interaction and a *relevant* understanding of the system. Most papers that have investigated how XAI methods influence end-users have done so in the context of AI-assisted decision-making, e.g., [31,32,33,34,35,36,37,38,39,40,41]. However, in this paper, we do not limit ourselves to AI-assisted decision-making. Instead, we address the fact that all AI systems could be subject to explanations.

3. Explanations of AI

As mentioned, one of the overarching goals of XAI is to make AI systems' behavior *understandable* to people. Likely due to how the field evolved, the XAI community predominantly focuses on one method to do so: open black boxes to elucidate internal mechanisms. Importantly, we note that understanding of AI-systems can be facilitated through methods usually overseen by the XAI community. Some of them are closely related to methods to support Situation Awareness (SA), i.e., "the perception of the elements in the environment within a volume of space and time, the comprehension of their meaning, and the projection of their status in the near future" (p.97) [42]. SA as described by Endsley [43,44,45], is often discussed in terms of three levels: *perception* (level 1), *comprehension* (level 2), and *projection* (level 3). In the context of interacting with a system,

level one is the perceptual cues the user gets from the system and the environment. Level two is the sense-making of the combination of perceived signals. Lastly, level three is to be able to predict likely future states. The higher levels, comprehension and projection, build directly on the user's mental model of the system [44].

In the following sections, we cover a non-exhaustive number of ways to facilitate understanding of AI behavior (i.e., explanations of AI) that are not (necessarily) grounded in traditional algorithm-centric XAI methods.

3.1. Social Transparency

Ehsan and colleagues [10] argued to expand the scope of explainability from a mere algorithmic focus to include a sociotechnically informed perspective that incorporates the socio-organizational context of the system. Doing this, they introduce the term *Social Transparency* (ST). In their paper [10], they developed a scenario-based design that incorporated information about "*who* did *what* with the AI system, *when*, and *why* they did what they did (4W)"(p.16) in the explanations that accompanied the system's recommendations. By adding this contextual information, decision-makers can gain *crew knowledge*. That is, get actionable insights from how other users have interacted with the system in similar situations. Even though ST is not grounded in a traditional algorithmic-centric XAI method, it can act as a method to facilitate understanding of an AI system.

3.2. Communicating the State of a System

Communicating the state of a system to facilitate successful interaction and understanding is not a novel idea in HCI [46,47,48], or HRI [49,50,51]. Yet, methods to do so are often overseen by the XAI community. In Nielsen's ten usability heuristics [47], the number one is *Visibility of system status*. That is, "The system should always keep users informed about what is going on, through appropriate feedback within reasonable time." (p.1) [47]. Depending on the nature of the AI system, different strategies could be used to communicate with users. However, it can often be achieved via simple cues that the user has learned to associate with a state.

Communicating the state of a system can explain to a user why (not) the system behaves in a certain manner. A trivial example of this from HCI is a loading screen in a computer system (Figure 1), which explains to the user why the system is not proceeding.

Similarly, a problem in HRI is that lay people interacting with talking robots often struggle with turn-taking in conversations [52,53]. The reason for this is likely that people's initial mental model is often flawed, and they expect to be able to talk the same way with the robot as they do with a human. However, many voice-based AI systems cannot handle processing audio input simultaneously as they process previous input or generate a response [52,53]. A simple cue indicating that the system is processing, such as different coloring on the robot's ears or a loading screen, could potentially mitigate this issue. Indeed, cues could be more humanlike, like prosody or gaze [52].

Another trivial example of a simple cue that explains the (absence of) behavior of a system is the red blinking lightning on electrical products. No matter what kind of electrical-power-driven system, digital camera, laptop, or robot vacuum cleaner, most people understand that the system does not operate due to the battery being out.

States can be communicated through any modality, not just visual but also, for instance, audio. A simple audio signal can communicate anything the user has learned to

Figure 1. Left: Image of a computer loading screen. Middle: Visualizes what the car "sees" (its beliefs) and its planned action/intended maneuver (print screen from video [54]). Right: Illustration of how simple visual cues, such as light, can communicate future actions of an autonomous system (image from [55]).

associate with it. Moreover, if the system can produce natural language, this is a channel with which it can communicate its state.

There are other states that can readily be communicated via cues. For instance, a physical or virtual agent could communicate its uncertainty about the state of the world or the consequence of an action by looking around in the room, or if it has a face, make a facial expression that is similar to the ones humans make when confused.

3.3. Communicating Future Actions or Goals

By communicating "intended" actions or goals, future behavior can sometimes be explained even before the user has the question, "Why did the system do so?" Note that, we do not claim that AI systems *have* intentions [56]. However, people interacting with autonomous artifacts often attribute intentional mental states to these systems [57]. At the same time, "this does *not* mean that people *necessarily* really believe that the robots in question actually have such folk psychological, humanlike or animallike mental states." (p.352) [57]. A systematic review by Thellman and colleagues [58] found that mental state attribution to robots increases aspects like predictability, explainability, and trust.

The idea to "play" on the human tendency to attribute mental states is not novel in HRI [49,55,59,60,61]. Pascher et al. [61] did a review on communicating robot motion intent. They identified three types of intent related to motion intent, which they refer to as attention, state, and instruction intent. Each of these types is further divided into intents related to the robot and ones related to the world. *Motion intent* directly communicates the intended action. One example is [55], which used a light-based cue to communicate the navigational intent of a robotic wheelchair to the passenger and surrounding pedestrians (Figure 1). *Attention intents* often come before motion intents to shift the human's attention to the robot. For instance, acoustic feedback can be used to gain humans' attention in order to avoid collision. By communicating the robot's *state*, the human can potentially deduce future actions. This connects to section 3.2. Lastly, *instruction intents* are cues that the robot provides to get the human to do something. For instance, hand over an object or get help to move an obstacle [61].

3.4. Communicating Beliefs

Similar to intentions, communicating beliefs can be useful to promote understanding of a system. When we use the term beliefs here, we solely mean that a system has a representation or model of something. For instance, in autonomous robots, such as highly

autonomous vehicles, the system has a representation of its surroundings and acts accordingly. Indeed, beliefs could sometimes be similar to states. However, states are about what *is*, and beliefs concern the *system's representations* of what is. Hence, these can, of course, be the same when the system has a correct representation. Communicating systems' beliefs can thus be useful to help users identify when the system and user have conflicting beliefs about the state of the world.

One way to communicate beliefs is by visualizing what the system "sees" to the user. One example is the screen in Tesla cars (Figure 1). By showing the user what the car "sees," the user can make sense of certain behaviors. For instance, if the car breaks for what to the human appears to be nothing. If the car visualizes that it "sees" a pedestrian close to the road, the human can understand why the car brakes. Note that this kind of explanation does not necessitate any traditional XAI method, as it solely visualizes the car's representations of the environment. However, worth noting is that this explanation does not explain to the user *why* the car misclassified the road post as a pedestrian. Nevertheless, it is unclear how most users would benefit from this information anyway. Indeed, similar strategies could be used for other autonomous systems, like robot vacuum cleaners, to communicate to the user what obstacles it sees or what areas it has visited.

Beliefs can, of course, be communicated through channels other than screens. Similar to how human drivers meet pedestrians' gaze and sometimes even nod to acknowledge that they have seen each other, an autonomous vehicle could do something similar. Likewise, a household robot could communicate that it knows that the owner is present by saying something or by "looking" at their face.

3.5. User Onboarding

The initial interaction phase is often not mentioned or completely overseen in user-centered XAI research. Even if successful interaction is preferably achieved via an easy-to-learn interface [62,63], HCI and UX research show that novice users of complex systems can benefit from a structured first-time experience [62,64]. This is likely partly due to initial mental models that differ too much from the actual system [62,65]. User onboarding can be defined as "...the sum of methods and elements helping a new user to become familiar with a digital product"(p.1) [64]. Onboarding methods can be guides or tutorials inside the system or online, or be done by expert teachers [62,64].

In user onboarding, central concepts or cues could be taught to the user. For instance, a novice user interacting with the social robot Furhat [66] may expect it to behave and have similar capabilities as a human [53,67,68]. Furhat can currently track a set number of people simultaneously. However, if a user is too far from Furhat, it cannot recognize them, and it will not remember them as they enter the trackable zone again [69]. A novice user could interpret this as if there is something wrong with Furhats' memory, as most healthy humans would not forget a person they just interacted with. However, if users were given this critical information to correct their mental model, they would understand that the reason Furhat does not recognize them is that they are out of the recognizable zone for a short while.

3.6. Understanding Over Time

User-centered XAI research has predominantly focused on how explanations influence users' understanding in the short term. However, most complex systems are not learned

through a handful of interactions. Instead, we often learn over time through repeated interaction with feedback. This feedback helps users create (and update) their mental model of how they believe the system behaves in a given context [70]. Indeed, letting users play around and try out the system is not always enough to build a sufficient understanding. Yet, there is an absence of studies investigating how users' understanding of AI systems changes over time [71,72]. Whether users' understanding plateaus after a while, potentially quite rapidly, or not needs more investigation.

As user understanding changes, so does what the best explanation is. One temporal aspect to consider is how the same kind of explanation facilitates understanding over time. Kulesza and colleagues [73] is one of few examples that investigates how explanations influence users' mental models over five days. Overall, more research is needed regarding how users' understanding of an AI system evolves with time using longitudinal studies.

4. Discussion

In this paper, we argue that explanations of AI go beyond the methods typically considered by the XAI community. Moreover, attempts to explain AI systems can benefit from insights from other fields that study human-machine interaction (HMI). However, we acknowledge that we have not specified what an AI system is, nor have we discussed related concepts such as interpretability, comprehensibility, intelligibility, and transparency. Nevertheless, our perspective aligns with Gunning's original XAI vision [11], which outlines two (connected) key objectives: *(i)* to "create a suite of ... machine learning techniques that produce explainable models that..." *(ii)* "enable end users to understand, appropriately trust, and effectively manage the emerging generation of AI systems." (p.1). This paper illustrates that achieving *(ii)* can be done not only with algorithmic-centric methods but also with other methods.

If the aim of XAI is to make systems more understandable to people [1,2,3,4,5,6, 7], HRI and HCI researchers have already developed methods that make this possible. These fields emphasize user understanding through approaches like onboarding, cues, and communicating system states or beliefs. Some of these, align closely with principles of increasing SA in the context of HMI [42,43,44,74,75], where increased automation often lead to loss in SA. Many of the issues that are observed in automation research apply to working with AI systems [74]. By improving users' SA, they can more readily make sense of the system's behavior.

Most methods mentioned in this paper are not applicable all the time or to all systems. For instance, ST is only applicable when someone has used the system in a similar situation before. Similarly, it only makes sense to communicate future actions or goals when they are not always the same. Moreover, for some systems, like in the autonomous car example, it is easier to communicate the system's state or beliefs. Meanwhile, traditional XAI methods may be needed to elucidate similar information for other systems.

However, the two aspects directly relevant to tech-centric explanations and which are rarely investigated in XAI user studies are user onboarding and longitudinal effects. Disregarding user onboarding is often a methodological flaw, as the initial presentation of a system influences how users interact and interpret its behavior. A more transparent description of how the system, model, or XAI method is initially presented to partici-

pants would strengthen the comparability between studies. For instance, if LIME is used as an explainer in a user study with laypeople, LIME must be introduced to the participants. How it is done will have an impact on participants' understanding and interpretation. Similarly, longitudinal effects, which explore how user understanding evolves over time [71,72], remain largely underexplored. Such research could reveal how explanations need to adapt to users' evolving mental models.

While we have focused on understanding as a primary motivation for XAI, another common motivation is trust (and trust calibration) [11,18,26,30,76,77,78]. We agree that transparency and interpretability are crucial for assessing whether a system is trustworthy. However, we do not believe that end-users' trust in systems primarily comes from system interpretability. To exemplify, none of the authors knows exactly how an airplane works, yet we trust the system around the airplane enough to fly. The captain or cabin crew does not explain to the passengers how the airplane works to get them to trust the system, nor is the airplane, to us, interpretable. Instead, we trust the regulations around this system. Similarly, we believe that transparency and interpretability are not prerequisites for end-user trust. However, we believe algorithm-centric XAI methods that increase interpretability are essential for developers and regulators to ensure a similar safety net to users as standards and flight regulations do for flight passengers. Moreover, this example illustrates that it is about more than just high/low-stake scenarios. When we are passengers in an airplane, the stakes are quite high. Yet we trust the system enough to put ourselves in that situation.

Future studies should address these gaps. For instance, longitudinal research could explore how user understanding and trust develop over repeated interactions with AI systems. Additionally, testing HCI, HRI, and UX-inspired methods, such as user onboarding or system state cues, in diverse real-world applications would validate their effectiveness and broaden the scope of XAI.

5. Conclusion

This paper calls for the XAI community to broaden its approach to facilitating user understanding, particularly for end-users. While algorithm-centric methods have dominated the field, insights from HCI and HRI demonstrate that non-algorithmic strategies—such as social transparency, onboarding, and communicating system states—can be effective. Indeed, not all methods mentioned are applicable or useful to all AI systems. Depending on the system, context, and stakeholders, we argue that facilitating understanding requires a diverse toolkit, whereas conventional algorithm-centric XAI methods represent just one approach among many. Traditional XAI methods are necessary for some stakeholders, such as developers and regulatory bodies. But most users just want the system to work as expected and know how to get it to do so if an error occurs.

Acknowledgements

We gratefully acknowledge the grant from the Swedish Research Council project XPECT, How to tailor explanations from AI systems to user's expectations (VR 2022-03180). We also thank the reviewers for their insightful comments.

References

[1] Liao QV, Varshney KR. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. arXiv. 2021 10. Available from: https://arxiv.org/abs/2110.10790v5.

[2] Adadi A, Berrada M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access. 2018 9;6:52138-60.

[3] Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang GZ. XAI-Explainable artificial intelligence. Science Robotics. 2019 12;4(37).

[4] Ehsan U, Passi S, Liao QV, Chan L, Lee IH, Muller M, et al. The Who in XAI: How AI Background Shapes Perceptions of AI Explanations. Conference on Human Factors in Computing Systems - Proceedings. 2024 5. Available from: https://dl.acm.org/doi/10.1145/3613904.3642474.

[5] Ehsan U, Riedl MO. Human-Centered Explainable AI: Towards a Reflective Sociotechnical Approach. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2020;12424 LNCS:449-66. Available from: https://link.springer.com/chapter/10.1007/978-3-030-60117-1_33.

[6] Ehsan U, Tambwekar P, Chan L, Harrison B, Riedl MO. Automated rationale generation: A technique for explainable AI and its effects on human perceptions. International Conference on Intelligent User Interfaces, Proceedings IUI. 2019;Part F147615:263-74. Available from: https://dl.acm.org/doi/10.1145/3301275.3302316.

[7] Longo L, Brcic M, Cabitza F, Choi J, Confalonieri R, Ser JD, et al. Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. Information Fusion. 2024 6;106:102301. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1566253524000794.

[8] Miller T. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence. 2019 2;267:1-38.

[9] Miller T, Howe P, Sonenberg L. Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences. arXiv. 2017 12. Available from: https://arxiv.org/abs/1712.00547v2.

[10] Ehsan U, Vera Liao Q, Muller M, Riedl MO, Weisz JD. Expanding Explainability: Towards Social Transparency in AI systems. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. vol. 19. Association for Computing Machinery; 2021. Available from: https://doi.org/10.1145/3411764.3445188.

[11] Gunning D, Vorm E, Wang JY, Turek M. DARPA's explainable AI (XAI) program: A retrospective. Applied AI Letters. 2021 12;2(4):e61. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/ail2.61https://onlinelibrary.wiley.com/doi/abs/10.1002/ail2.61https://onlinelibrary.wiley.com/doi/10.1002/ail2.61.

[12] Larsson S, Heintz F. Transparency in artificial intelligence. Internet Policy Review. 2020 5;9(2):1-16.

[13] Larsson S, Haresamudram K, Högberg C, Lao Y, Nyström A, Söderlund K, et al. Four facets of AI transparency. In: Handbook of Critical Studies of Artificial Intelligence. Edward Elgar Publishing Ltd.; 2023. p. 445-55.

[14] Saranya A, Subhashini R. A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends. Decision Analytics Journal. 2023 6;7:100230.

[15] Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A Survey of Methods for Explaining Black Box Models. ACM Computing Surveys. 2019 9;51(5):1-42.

[16] Burkart N, Huber MF. A Survey on the Explainability of Supervised Machine Learning. Journal of Artificial Intelligence Research. 2021 1;70:245-317. Available from: https://www.jair.org/index.php/jair/article/view/12228.

[17] Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion. 2020 6;58:82-115.

[18] Ribeiro MT, Singh S, Guestrin C. "Why should i trust you?" Explaining the predictions of any classifier. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016 8;13-17-August-2016:1135-44. Available from: https://dl.acm.org/doi/10.1145/2939672.2939778.

[19] Lundberg SM, Allen PG, Lee SI. A Unified Approach to Interpreting Model Predictions. In: Conference on Neural Information Processing Systems; 2017. Available from: https://github.com/slundberg/shap.

[20] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. International Journal of Computer Vision. 2020 2;128(2):336-59. Available from: https://link.springer.com/article/10.1007/s11263-019-01228-7.

[21] Chattopadhay A, Sarkar A, Howlader P, Balasubramanian VN. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. Proceedings - 2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018. 2018 5;2018-January:839-47.

[22] Arya V, Bellamy RKE, Chen PY, Dhurandhar A, Hind M, Hoffman SC, et al. One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques. arXiv. 2019 9. Available from: https://arxiv.org/abs/1909.03012v2.

[23] Hind M. Explaining explainable AI. XRDS: Crossroads, The ACM Magazine for Students. 2019 4;25(3):16-9. Available from: https://dl.acm.org/doi/10.1145/3313096.

[24] Laato S, Tiainen M, Najmul Islam AKM, Mäntymäki M. How to explain AI systems to end users: a systematic literature review and research agenda. Internet Research. 2021;32(7):1-31.

[25] Cambria E, Malandri L, Mercorio F, Mezzanzanica M, Nobani N. A survey on XAI and natural language explanations. Information Processing & Management. 2023 1;60(1):103111.

[26] Abdul A, Vermeulen J, Wang D, Lim BY, Kankanhalli M. Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. Conference on Human Factors in Computing Systems - Proceedings. 2018 4;2018-April. Available from: https://dl.acm.org/doi/10.1145/3173574.3174156.

[27] Wang Z, Huang C, Yao X. A Roadmap of Explainable Artificial Intelligence: Explain to Whom, When, What and How? ACM Transactions on Autonomous and Adaptive Systems. 2024 11;19(4). Available from: https://dl.acm.org/doi/10.1145/3702004.

[28] Mohseni S, Zarei N, Ragan ED. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. ACM Transactions on Interactive Intelligent Systems. 2021 12;11(3-4):1-45.

[29] Mueller ST, Veinott ES, Hoffman RR, Klein G, Alam L, Mamun T, et al. Principles of Explanation in Human-AI Systems. arXiv. 2021 2. Available from: https://arxiv.org/abs/2102.04972v1.

[30] Ali S, Abuhmed T, El-Sappagh S, Muhammad K, Alonso-Moral JM, Confalonieri R, et al. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. Information Fusion. 2023 11;99:101805.

[31] Wang D, Yang Q, Abdul A, Lim BY. Designing theory-driven user-centric explainable AI. Conference on Human Factors in Computing Systems - Proceedings. 2019 5. Available from: https://dl.acm.org/doi/10.1145/3290605.3300831.

[32] Buçinca Z, Malaya MB, Gajos KZ. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. Proc ACM Hum-Comput Interact. 2021;5:21. Available from: https://doi.org/10.1145/3449287.

[33] Bansal G, Nushi B, Kamar E, Lasecki WS, Weld DS, Horvitz E. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. In: The Seventh AAAI Conference on Human Computation and Crowdsourcing; 2019. Available from: www.aaai.org.

[34] Bansal G, Fok R, Ribeiro MT, Wu T, Zhou J, Kamar E, et al. Does the whole exceed its parts? The effect of ai explanations on complementary team performance. Conference on Human Factors in Computing Systems - Proceedings. 2021 5:16. Available from: https://dl.acm.org/doi/10.1145/3411764.3445717.

[35] Bertrand A, Belloum R, Eagan JR, Maxwell W. How cognitive biases affect XAI-Assisted decision-making: A systematic review. AIES 2022 - Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society. 2022 7:78-91. Available from: https://dl.acm.org/doi/10.1145/3514094.3534164.

[36] Miller T. Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven Decision Support using Evaluative AI. ACM International Conference Proceeding Series. 2023 6;(23):333-42. Available from: https://dl.acm.org/doi/10.1145/3593013.3594001.

[37] Le T, Miller T, Sonenberg L, Singh R. Towards the New XAI: A Hypothesis-Driven Approach to Decision Support Using Evidence. arXiv. 2024 2. Available from: https://arxiv.org/abs/2402.01292v3.

[38] Riveiro M, Thill S. "That's (not) the output I expected!" On the role of end user expectations in creating explanations of AI systems. Artificial Intelligence. 2021 9;298:103507.

[39] Riveiro M, Thill S. The challenges of providing explanations of AI systems when they do not behave

like users expect ACM Reference Format. In: Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization. ACM; 2022. Available from: https://doi.org/10.1145/3503252.3531306.

[40] Dodge J, Vera Liao Q, Zhang Y, Bellamy RKE, Dugan C. Explaining models: An empirical study of how explanations impact fairness judgment. International Conference on Intelligent User Interfaces, Proceedings IUI. 2019;Part F147615:275-85. Available from: https://doi.org/10.1145/3301275.3302310.

[41] Holmberg L, Riveiro M, Thill S. Leveraging large language models for tailored and interactive explanations in AI systems. In: Olofsson J, Jernsäther-Ohlsson T, Thunberg S, Holm L, Billing E, editors. Proceedings of the 19th Swecog Conference. Stockholm; 2024. p. 27-31. Available from: https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1908743&dswid=-1211.

[42] Endsley MR. Situation Awareness in Aircraft Systems: Symposium Abstract. Proceedings of the Human Factors Society Annual Meeting. 1988 10;32(2):97-101.

[43] Endsley MR, Connors ES. Situation Awareness: State of the Art. In: 2008 IEEE Power and Energy Society General Meeting-Conversion and Delivery of Electrical Energy in the 21st Century; 2008. .

[44] Endsley MR. Supporting Human-AI Teams:Transparency, explainability, and situation awareness. Computers in Human Behavior. 2023 3;140:107574.

[45] Endsley MR, Kaber DB. Level of automation effects on performance, situation awareness and workload in a dynamic control task. Ergonomics. 1999 3;42(3):462-92.

[46] Nielsen J. Enhancing the explanatory power of usability heuristics. Conference on Human Factors in Computing Systems - Proceedings. 1994:152-8. Available from: https://dl.acm.org/doi/10.1145/191666.191729.

[47] Nielsen J. Ten Usability Heuristics; 2005.

[48] Shneiderman B, Plaisant C, Cohen M, Jacobs S, Elmqvist N, Diakopoulos N. Designing the User Interface: Strategies for Effective Human-Computer Interaction. 6th ed. Pearson; 2016.

[49] Roy L, Croft EA, Kulic D. Learning to Communicate Functional States With Nonverbal Expressions for Improved Human-Robot Collaboration. IEEE Robotics and Automation Letters. 2024 6;9(6):5393-400.

[50] Baraka K, Rosenthal S, Veloso M. Enhancing human understanding of a mobile robot's state and actions using expressive lights. 25th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2016. 2016 11:652-7.

[51] Knight H, Simmons R. Laban head-motions convey robot state: A call for robot body language. Proceedings - IEEE International Conference on Robotics and Automation. 2016 6;2016-June:2881-8.

[52] Skantze G. Turn-taking in Conversational Systems and Human-Robot Interaction: A Review. Computer Speech & Language. 2021 5;67:101178.

[53] Lala D, Inoue K, Kawahara T. Smooth turn-taking by a robot using an online continuous model to generate turn-taking cues. ICMI 2019 - Proceedings of the 2019 International Conference on Multimodal Interaction. 2019 10;19:226-34. Available from: https://dl.acm.org/doi/10.1145/3340555.3353727.

[54] Tesla. Full Self-Driving (Supervised) | Tesla; 2024. Available from: https://www.youtube.com/watch?v=TUDiG7PcLBs.

[55] Watanabe A, Ikeda T, Morales Y, Shinozawa K, Miyashita T, Hagita N. Communicating robotic navigational intentions. IEEE International Conference on Intelligent Robots and Systems. 2015 12;2015-December:5763-9.

[56] Searle JR. Minds, brains, and programs. Behavioral and Brain Sciences. 1980;3(3):417-24. Available from: https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/abs/minds-brains-and-programs/DC644B47A4299C637C89772FACC2706A.

[57] Ziemke T. Understanding Social Robots: Attribution of Intentional Agency to Artificial and Biological Bodies. Artificial Life. 2023 8;29(3):351-66. Available from: https://dx.doi.org/10.1162/artl_a_00404.

[58] Thellman S, De Graaf M, Ziemke T. Mental State Attribution to Robots: A Systematic Review of Conceptions, Methods, and Findings. ACM Transactions on Human-Robot Interaction. 2022 9;11(4). Available from: https://dl.acm.org/doi/10.1145/3526112.

[59] Vernon D, Thill S, Ziemke T. The Role of Intention in Cognitive Robotics. Intelligent Systems Reference Library. 2016 3;105:15-27. Available from: https://link.springer.com/chapter/10.1007/978-3-319-31056-5_3.

[60] Thill S, Ziemke T. The role of intentions in human-robot interaction. ACM/IEEE International Con-

ference on Human-Robot Interaction. 2017 3:427-8. Available from: https://dl.acm.org/doi/10.1145/3029798.3029802.

[61] Pascher M, Gruenefeld U, Schneegass S, Gerken J. How to Communicate Robot Motion Intent: A Scoping Review. Conference on Human Factors in Computing Systems - Proceedings. 2023 4. Available from: https://dl.acm.org/doi/10.1145/3544548.3580857.

[62] Chauvergne E, Hachet M, Prouzeau A. User Onboarding in Virtual Reality: An Investigation of Current Practices. Conference on Human Factors in Computing Systems - Proceedings. 2023 4. Available from: https://dl.acm.org/doi/10.1145/3544548.3581211.

[63] Grossman T, Fitzmaurice G, Attar R. A survey of software learnability: Metrics, methodologies and guidelines. Conference on Human Factors in Computing Systems - Proceedings. 2009:649-58. Available from: https://dl.acm.org/doi/10.1145/1518701.1518803.

[64] Renz J, Staubitz T, Pollak J, Meinel C. Improving The Onboarding User Experience in MOOCS. EDULEARN14 Proceedings. 2014 7:3931-41. Available from: http://library.iated.org/view/FELIP2014MON.

[65] Linja-aho M. Creating a framework for improving the learnability of a complex system. Human Technology. 2006 10;2(2):202-24. Available from: https://ht.csr-pub.eu/index.php/ht/article/view/38.

[66] Furhat. The Furhat Robot | Furhat Robotics; 2024. Available from: https://www.furhatrobotics.com/furhat-robot.

[67] Abubshait A, Wiese E. You look human, but act like a machine: Agent appearance and behavior modulate different aspects of human-robot interaction. Frontiers in Psychology. 2017 8;8(AUG):277299. Available from: www.frontiersin.org.

[68] Kwon M, Jung MF, Knepper RA. Human expectations of social robots. ACM/IEEE International Conference on Human-Robot Interaction. 2016 4;2016-April:463-4.

[69] Furhat. Users and Attention - Furhat Developer Docs; 2025. Available from: https://docs.furhat.io/users/.

[70] Honig S, Oron-Gilad T. Understanding and resolving failures in human-robot interaction: Literature review and model development. Frontiers in Psychology. 2018 6;9(JUN).

[71] Bigras E, Jutras MA, Sénécal S, Léger PM, Fredette M, Black C, et al. Working with a recommendation agent: How recommendation presentation influences users' perceptions and behaviors. Conference on Human Factors in Computing Systems - Proceedings. 2018 4;2018-April. Available from: https://dl.acm.org/doi/10.1145/3170427.3188639.

[72] Tullio J, Dey AK, Chalecki J, Fogarty J. How it works: A field study of non-technical users interacting with an intelligent system. Conference on Human Factors in Computing Systems - Proceedings. 2007:31-40. Available from: https://dl.acm.org/doi/10.1145/1240624.1240630.

[73] Kulesza T, Stumpf S, Burnett M, Kwan I. Tell me more? the effects of mental model soundness on personalizing an intelligent agent. Conference on Human Factors in Computing Systems - Proceedings. 2012:1-10. Available from: https://dl.acm.org/doi/10.1145/2207676.2207678.

[74] Endsley MR. Ironies of artificial intelligence. Ergonomics. 2023;66(11):1656-68.

[75] Chiou EK, Demir M, Buchanan V, Corral CC, Endsley MR, Lematta GJ, et al. Towards Human–Robot Teaming: Tradeoffs of Explanation-Based Communication Strategies in a Virtual Search and Rescue Task. International Journal of Social Robotics. 2022 7;14(5):1117-36. Available from: https://link.springer.com/article/10.1007/s12369-021-00834-1.

[76] Kaur D, Uslu S, Rittichier KJ, Durresi A. Trustworthy Artificial Intelligence: A Review. ACM Computing Surveys. 2023 2;55(2):1-38.

[77] Zhang Y, Vera Liao Q, Bellamy RKE. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 2020 1:295-305. Available from: https://dl.acm.org/doi/10.1145/3351095.3372852.

[78] Papenmeier A, Englebienne G, Seifert C. How model accuracy and explanation fidelity influence user trust. arXiv. 2019 7. Available from: https://arxiv.org/abs/1907.12652v1.