Human-AI Trust Calibration Should Be Contextual and Continuous

STEVEN R. GOMEZ, KEVIN K. NAM, KIMBERLEE CHESTNUT CHANG, and GREGG MARCUS, MIT Lincoln Laboratory, USA

ERIN K. CHIOU, Arizona State University, USA

This paper argues that trust calibration is an important step for real-world Human-AI teaming and requires continuous calibration throughout the lifecycle of the AI. This perspective is informed by the authors' experiences working to close the gap between engineering trustworthy AI during early-stage research and design, and understanding user trust in operational technologies once fielded. We frame three kinds of *trust sources* of data for Human-AI systems and argue that these should be more deliberately utilized for trust calibration across multiple stages in the technology lifecycle. We propose a concept artifact aimed at bringing together these data in the form of a context-aware, adaptive scorecard for reasoning about AI trustworthiness and context of use.

CCS Concepts: • **Human-centered computing → HCI theory, concepts and models**; **Human computer interaction (HCI)**; • **Computing methodologies → Artificial intelligence**.

Additional Key Words and Phrases: artificial intelligence, machine learning, trust, transparency, explainability, human-computer interaction

1 INTRODUCTION

In recent years, approaches to developing trustworthy AI have focused both on enhancing model fitness through new architectures and training (e.g., adversarial training), as well as quantifying and communicating useful information about the model and its behaviors to improve transparency. There will always be a need for operators to interpret the outputs of an AI decision-support system, regardless of future improvements to model accuracy. In the authors' experience, this is especially true in government mission domains, either due to the criticality of the decisions being made or due to decision-making authority structures in place (e.g., chain of command). In the former case, there is a need to anticipate or analyze consequential AI failures, however rare. For the latter, there is a need to track elements of decision-making provenance or rationale for stakeholders at different levels of operational involvement. For AI decision-support tools to be effective at improving operational outcomes, stakeholders at multiple levels must have appropriate levels of trust in these systems.

Interpreting the AI's performance, either with tools or simply observing the outputs and mission outcomes, is central to how trust is formed and updated by users of the system. People are susceptible both to over-trusting and under-trusting AI tools, leading to different kinds of adverse outcomes, like incorrect decision-making or unjustified tool abandonment. Our position is that new techniques and evaluation processes are needed to help decision makers and other stakeholders better align their expectations of the AI's reliability or trustworthiness with its actual behaviors—to better *calibrate trust*—to avoid these outcomes. For the remainder of the paper, we are primarily considering recommender systems that help human decision-makers to accept, reject, or modify the output of the AI support tool, though our position may apply to other models of Human-AI interaction as well. We aim to shine a light on two aspects of Human-AI trust calibration:

(1) The sources of data that impact human trust are highly contextual and have not been operationalized fully up to this point in AI interpretability research.
(2) The lifecycle through which AI-based decision-support tools are developed, evaluated, and deployed may have multiple or continuous times at which trust calibration is critical.

Table 1. Human-AI Trust Sources

Source Type	Examples	Variability Between Missions	Potential Sources that Cause Trust Variability	Interpretation Complexity	When Trust is Formed or Updated
Static	▪ AI model performance measures ▪ Documented bias, errors and bounds ▪ Reputation of developer ▪ Guiding laws and policies	Low	▪ Newly discovered developer mishap or recall ▪ Updated laws and policies ▪ Any updates to the model since initial validation	Low	▪ Document review ▪ Education and training ▪ Best practice development
Dynamic	▪ Past mission outcome and results of similar decision making tasks ▪ Degree and subject of accountability ▪ Mission risks and uncertainty ▪ Costs of failure	Mid to High	▪ Differences from known mission parameters ▪ Human teammate's cognitive and workload	Mid to High	▪ Comparative analysis against past missions ▪ Task exercises modeled on or relevant to the intended mission
Interactive	▪ Experience through trial and error with AI in specific decision-making tasks ▪ AI explanations	Mid to High	▪ Task variability ▪ AI behavior changes due to new learning and adaptation ▪ Previously unseen AI quirks ▪ Human emotions ▪ Degree of anthropomorphism	High	▪ Repeated interaction with the AI ▪ Interaction during illustrative exemplars for the mission

Furthermore, we discuss opportunities for artifacts like AI "scorecards" that incorporate varied trust sources and can serve different stakeholders for calibrating AI trust during the lifecycle stages of a decision-support tool.

The remainder of this paper is organized as follows: We first outline sources of data that inform stakeholders about the trustworthiness of the AI components in operational Human-AI workflows; these data are produced and interpreted across different stages of the technology lifecycle. Next we review findings about trust formation between people and AI support tools, then discuss why current explainable AI (XAI) techniques may be insufficient for helping stakeholders calibrate trust. Finally, we propose a "living scorecard" concept to support trust calibration and discuss opportunities and limitations for its design.

2 CALIBRATING TRUST IN HUMAN-AI OPERATIONS

A key aspect of trust calibration in real-world scenarios is that trust in the AI by its hands-on operators and other stakeholders is influenced by sources of data that originate across the system lifecycle. We refer to these data sources as *trust sources* in the remainder of this paper. For example, a meteorologist using an AI weather tool may consider the fact that the tool has been validated by an organization like the US National Weather Service prior to deployment, along with their own experiences training to use the tool. A posteriori retrospectives on how the tool performed during operations ("26 days in the last month had weather as predicted, and 4 days were predicted incorrectly") could support or be in conflict with the user's formative estimates of the AI's trustworthiness, leading to recalibration. In fact, the user's own knowledge in the decision-making domain, or of AI technologies, is likely to help them interpret and scrutinize outputs from the system situationally [11, 24]. Understanding the sources of what influences trust estimates by AI users is an important step in helping to develop processes and tools to support trust calibration.

2.1 Varied Data Informs AI Trust

In Table 1, we synthesize different types of data that might impact human-AI trust calibration—informed by recent work in trustworthy AI, XAI, and teaming—and argue that supporting effective trust calibration between people and AI

decision-support tools must account for these sources. The table shows example sources, variability, and how and when trust sources are interpreted by people. The table is not exhaustive but an initial attempt at cataloging the data that inform AI trust judgments during real operations, which is an important step for designing better support tools for trust calibration ("What inputs are available for improved reasoning about model behaviors with my tasks?"). Here we identify three types of trust sources:

- **Static sources** help provide single, fixed estimates of the trustworthiness of the AI system. This includes data gathered from one-off evaluations during testing, reputational information about the developers, etc. Formal verification with well-defined task boundaries and known datasets might be considered a static source.
- **Dynamic sources** may be invalidated or updated over time, and include observing mission outcomes that may have been impacted by correct or incorrect AI guidance, changing mission needs and data sources that can impact the fitness of the particular AI model or its task, etc.
- **Interactive sources** of trust data are the means through which operators can probe, influence the model, or interpret any explanations it provides, especially with their use cases in mind. We distinguish this from dynamic sources in that they are task-specific and provide new data in the moment of decision making or shortly thereafter. Interaction here could also be bilateral, including nudges from the AI system to operator, or other back-and-forth sensemaking with the model.

The temporal lens of analyzing these sources is complementary to other taxonomies, like the sources of variability Hoff and Bashir identified in their survey of empirical trust research—human operator, environment, and automated system [11]—and useful for understanding how reliability information gets produced and carried forward in real-world operations. We might expect that in live operational environments, operators' trust is likely to be most impacted by interactive, dynamic, and static sources––in that order—-because the relative timeliness of interactive sources compared to static ones may reflect more task-specific data and also trigger recency bias in the operator. Of course, not all scenarios will follow that pattern, nor is there a single "correct" way we can prescribe for any stakeholder to weigh all these data.

Furthermore, a single trust source (like AI performance results from a field deployment and post-mortem analysis) may serve multiple stakeholder roles. For example, a human factors researcher could modify the experimental design of a controlled experiment based on observations about how operators interacted with their AI tools in the field. Another decision maker could use the same performance information as a summary evaluation to define new tool requirements (creating a new static source for AI developers).

2.2 Trust Building with AI

Many factors that affect trust building in AI have been identified such as mission risk, uncertainty, individual differences, reputation [22], self confidence in the task [12], and interaction structure [10, 21]. Interactions between these factors remain hazy. For example, how do we reconcile automation bias with the benefits of transparency in interactive machine learning [10]? The goal of this work is not to enumerate all factors of trust, but to draw attention to different trust sources that may be more or less influential in different parts of the technology lifecycle. For example, controlled user studies or benchmarking during the research and development of an AI system may be the best way to gauge its reliability prior to deploying it. But those results might have limited influence on stakeholders' trust later on when interactive or more longitudinal dynamic sources of trust (e.g., field-performance lessons learned) are available.

Many user studies evaluating Human-AI teamwork are conducted with simplified use cases or in gaming environments, so it may be difficult to extrapolate results to operational scenarios in which the mission environment or system

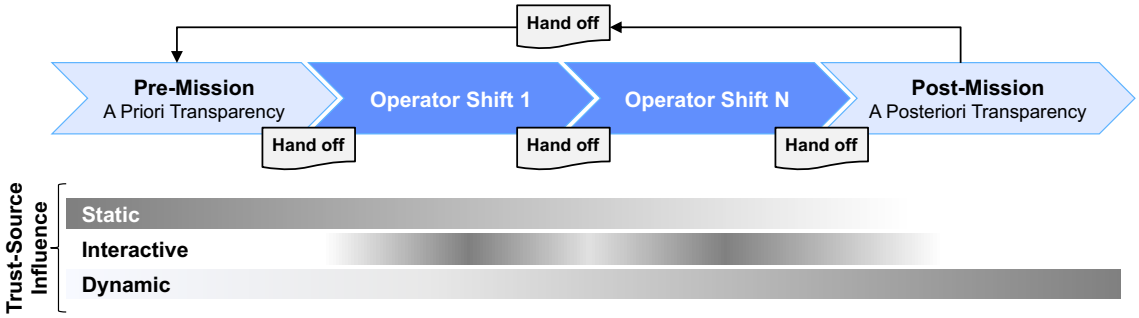

Fig. 1. Notional Mission Cycle Showing Pre- and Post-mission Stages and Potential Influences of Trust Sources by Type

boundaries are dynamic or unclear. Measuring the right thing can also be challenging for evaluating a Human-AI system. For example, using proxy tasks with people to evaluate AI models or explanations can lead to inappropriate take-aways about the tool's performance in its intended task [3]. Under-trusting should be distinguished from having appropriately-low trust in the technology, which might be justified when the technology is not reliable in an existing workflow. Furthermore, skilled operators who are able to complete a task well on their own without the supporting technology may choose not to use it, even in cases where they have high trust in it.

Figure 1 shows a portion of the technology lifecycle centered around mission planning and operations, along with notional relative influences of different trust-source types during different stages. We consider a mission loop that involves pre-mission planning, then operators performing the mission with the AI system with possible shift hand-offs between individuals, and finally post-mission assessment. In the first iteration of this loop, static sources may be most influential for mission planners who have limited field data. Interactive sources like AI explanations or reflecting on real-time errors may be the most visible to operators, though some information from these sources may be lost when operators change shifts unless there are effective tools to externalize and share learned operational knowledge. Other dynamic trust sources collected during the mission may be aggregated for post-mission analysis.

We argue that re-evaluating trust in the AI system should happen at each stage more deliberately by utilizing all available trust sources, and that these hand-off points represent boundaries at which contexts change—and where new tools or processes that support trust calibration in the AI could be developed.

This framing complements earlier work by Miller that argues different levels of AI decision-making transparency (*a priori transparency* before observing the system in operations, and *a posteriori transparency* afterward) at these stages play a key role in establishing human trust [16]. He notes challenges for people reasoning about transparency in the "moment of execution" where operators may have other timing or cognitive demands that interfere with interpreting even a perfectly transparent AI teammate. Entin and Serfaty found that high-performing human teams could make due with terse or otherwise limited communication (and therefore, interactive or in-the-moment transparency) in the moment of execution by leveraging shared mental models built up outside of operations [8]. Transparency in different lifecycle phases facilitates trust calibration in various trust dimensions, including affective, analogic, and analytic-based trust [13]. Interactive Team Cognition (ITC) further argues that teaming should be understood as an interactive activity that is highly context-dependent [5], and this focus on interactivity including concepts like coordination and cooperation, rather than independent trust factors, has been further specified for trust in automation more recently [4].

The extent to which transparency modifies the operators' prior beliefs about the AI's reliability is difficult to measure. If those modifications are valid, how can they be captured to inform tool design and decision-making stakeholders outside of hands-on operations? We discuss this further in Section 3.

2.3 XAI is Not a Magic Bullet for Trust Calibration

Explainable AI is an area of research and techniques generally aimed at discovering and communicating to people how an AI model (often an opaque Deep Neural Network, or DNN) produces its outputs. While the field is advancing rapidly, there remain open challenges, like:

- aligning representations of model behavior to a specific user's task and knowledge (e.g., is feature attribution useful to a person if the features are not individually human-interpretable?);
- integrating useful global information about the model (e.g., performance on benchmark datasets) alongside local explanations;
- preventing cognitive biases while interpreting XAI outputs;
- making explanations that are both truthful and digestible during operations (e.g., not consuming the time and cognitive resources needed for the operator's own decision making); and others.

The rest of this section discusses several of these issues in more depth, and argues broadly that other trust sources are needed to contextualize XAI outputs for accurate trust calibration.

Much of the focus in the XAI research community thus far has centered on supporting ML engineers in debugging models (e.g., [25])—earlier in the lifecycle than mission planning—but some work is aimed at helping operators interpret live AI outputs and potentially catch erroneous recommendations. Model-agnostic XAI frameworks like LIME [19] and SHAP [15] may have dual purposes in 1) spot-checking model behaviors by evaluating individual task instances or producing batches of saliency maps, and 2) helping an operator produce an explanation, in the moment of execution. Other forms of explanations (e.g., user-readable decision trees [18, 20] have aimed to support reasoning about projected situations for more operationally-focused stakeholders. But explanations at interactive moments may be more or less effective at correctly helping a person calibrate trust depending on additional context during operations.

The insight that highly-effective teams can streamline in-the-moment communication as shared mental models are developed [8] suggests that more efficient representations of explanations could evolve over time, as the AI's behaviors are better understood and conditioned on who the operator is (e.g., their goals and knowledge) and relational factors in the mission environment (response tempo, shared attention, etc.). Tim Miller describes the pitfalls of operationalizing XAI without incorporating relational information about the operator, task, and task environment [17]. De Bruijn et al. summarize ongoing challenges that XAI faces in achieving trust and transparency among stakeholders, including situational needs for explanations such as dynamic data and decisions and context dependent explanations [6].

A broader challenge for the research community is to encourage case studies on effective explanation methods for specialized systems, at the risk of producing less generalizable tools or academic findings, because demonstrating ways to elicit and use rich contextual information for trust calibration would be helpful for applied researchers and practitioners.

Even with XAI techniques that incorporate context, their role in helping stakeholders calibrate trust in decision-support tools should be cautiously evaluated. Explanation techniques may vary in their robustness. For example, perturbation-based saliency techniques may be more unstable than gradient-based methods [2]. Gilpin et al. describe the need for multiple levels of explainability tied to the functional role of the stakeholder [9]. They also discuss the

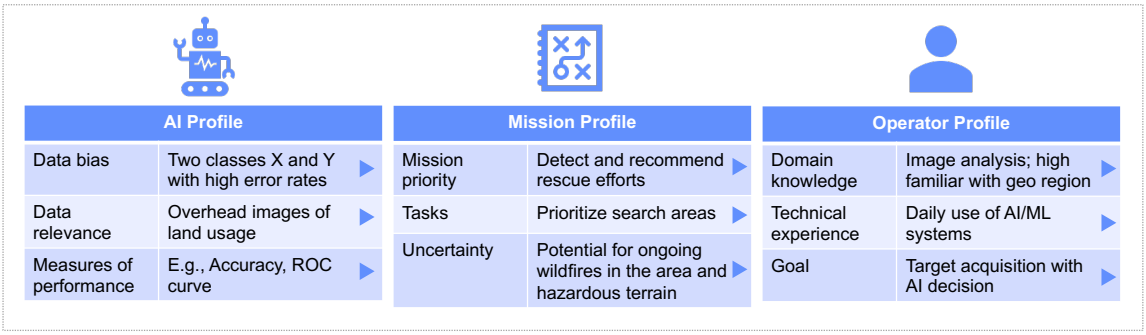

Fig. 2. Concept Sketch for AI Trust Scorecard

dangers of XAI dark patterns, like focusing on increasing perceived trust in a system rather than ensuring that the system is trustworthy. Earlier work by Ehsan and Riedl highlighted explainability pitfalls and proactive mitigations for them at different levels [7]. Creating effective context-dependent explanations is not just difficult in one specific instance, but could lead to systems that are overfitted to a specific context in environments that change rapidly. For example, consider an instance of the mission loop shown in Figure 1, where shift changes for an operator like an intelligence analyst necessitate a hand-off of responsibilities or work products. Meanwhile, as the AI system persists across the mission, ideally it is aware of the new analyst's knowledge and prior beliefs (and skepticism of the AI tool) and adapts its own representations and affordances.

3 MOTIVATING A LIVING SCORECARD FOR HUMAN-AI TRUST

We believe new information artifacts are needed for Human-AI teaming in real-world operations that go beyond localized AI explanations or model summary evaluations, and utilize other trust sources and mission factors. We propose the concept of a *trust scorecard* that helps stakeholders at different phases of the AI system's lifecycle—including but not limited to developers, evaluators, and operators—more accurately align their expectations of the AI system to its performance in the operational context. The scorecard analogy is made only in terms of its function, rather than its implementation or design, to help people interpret a system's capabilities as concisely and accurately as possible. The artifact could take the form of a single, centralized portal of information about the AI, or be composed of smaller modules.

3.1 A Boundary Object to Support Trust Calibration

Trust scorecards can be used as boundary objects [23] between the stages of the AI decision-support lifecycle and between groups of stakeholders who have specific goals and tasks. Developing, validating, operating, and maintaining AI decision-support systems is a cooperative endeavor. People supporting a common mission need to have consensus on an AI's reliability and behaviors, even if those people do not interact closely or have the same visibility into the system or mission needs. Therefore, the boundary object needs to inform different stakeholders' understanding of the AI's trustworthiness by conveying useful information gained from earlier stages that informs the current stage [14]. To achieve this, new technology functions for the scorecard object are discussed below.

Maintaining a rich data model. In our concept, scorecards are continually updated as they go through different stages of the lifecycle. Each scorecard instance will capture relevant information about a tuple of profiles: the AI system,

mission context, and human agent. Other profiles, like the task environment, may be useful to fully contextualize this snapshot of the AI system's trustworthiness. For example, relevant information could include:

- measures of AI performance and known biases,
- task priorities,
- uncertainties,
- previous outcomes, and
- domain expert knowledge and experience.

Figure 2 shows examples of these data fields in a notional scorecard concept for a disaster-response decision-support AI. The field values in each profile may be annotations or other kinds of expressive, data-driven visualizations. Aspects of the scorecard system may work automatically in some cases, like comparing and displaying benchmark performance relevant to current tasks. The data are likely to be multi-modal and multi-dimensional—for example, time-series measurements of Human-AI team performance in a baseline environment. Opportunities for research and design in this area range from identifying best practices for soliciting tacit operator knowledge through interactive user interfaces, to developing efficient, mission-specific representations of tasks, to automatically synthesizing insights from user feedback.

Enabling mission-relevant reasoning. The scorecard is meant to be extensible to support organic development of AI components to suit particular practitioners' and mission needs [14]. It should contain relevant information to be handed off across user and domain boundaries, and support interoperability between the individual organizational units that may be in charge of a particular stage of development or operations. Stakeholders of the scorecard could use its current approximation of information about the trustworthiness of the system to evaluate it and make decisions for the next iteration (e.g., "ready to deploy", "should be cautious in these cases", "needs additional training before the next mission"). We envision the scorecard is also able to automatically identify patterns in the AI behaviors that are annotated as failures between mission cycles, and can visualize these for stakeholders. For example, in a vision-based object-detection system, it could present examples of visual features that are commonly present in inputs that are misclassified. This dynamic knowledge about the types of mistakes the AI makes may be updated as the sample size of tasks it performs grows and maps out more of the AI model's task domain.

Supporting model debugging and improvement. A scorecard is not only a remedy for under- and over-trust in the AI, but can also play a role in improving the system's actual fitness. For example, if poor AI decisions during operations result in operators losing trust in the technology and abandoning it, an evaluator of the system could assess the use context and make a determination about the cause of the failure. Developers may then be asked to improve specific aspects of the AI model or the application(s) using it, or apply particular design interventions that better support trust calibration. A trust scorecard will enable other useful activities related to test and evaluation or early-stage prototyping. For example, we expect that large organizations in industry or government will increasingly need to test and evaluate competing AI models for a target application or mission, and the qualitative and quantitative information captured in scorecards during A/B or canary testing of the systems can aid in the selection process.

3.2 Limitations

We have argued for the benefits of identifying static, dynamic, and interactive trust sources throughout the technology lifecycle to better support understanding and calibrating Human-AI trust. Further work is needed to capture and use the information in practice. We believe the scorecard concept is a useful target, though several challenges remain.

First, some context-rich information is inherently difficult to encode. Information and knowledge capture and encoding has been an important research problem in many domains and several techniques have been suggested such as data format standardization, ontology and taxonomy development, and use of machine learning to align and integrate different schemas. All of the approaches have considerable challenges that will require communication and collaboration among disparate communities of practice. Rather than defining the scorecard in its entirety, iterative development within a focused user group may be more realistic.

Second, as the scorecard is envisioned to capture many complex data types and points, the data may require a specialized tool for viewing, using, sharing, and interpreting. Fit-for-purpose data object formats such as NetCDF [1] have been created and used by many scientific communities to capture complex multi-dimensional data in a machine agnostic way. We believe that iterative and community-specific development may lead to data formats and visualization tools that support intuitive navigation of the scorecard data.

Finally, depending on the evaluation environment, not everyone will have the means to generate data points in a scorecard. However, the scorecard is still valuable in that it highlights what is available and missing in order to calibrate human and organizational trust in an AI at multiple points in its lifecycle.

4 CONCLUSION

We presented an operations-focused perspective on the need for accurate trust calibration by people working with AI decision-support tools. Operators and other stakeholders can benefit from having insights about the strengths and weaknesses of a tool and the analytic provenance of its recommendations. There may be subtle trust sources that impact operational decision making that are not being incorporated into existing Human-AI workflows, even ones that apply current XAI techniques. We summarized these sources into static, dynamic, and interactive categories by drawing on related work, and challenge the XAI community to consider them when designing techniques to support AI trust calibration in real-world operations. We proposed one such concept aimed at providing critical insights to AI stakeholders at different stages of a system's lifecycle—a living trust scorecard that is both context-dependent and adaptive. We highlighted limitations and opportunities for this kind of artifact.

ACKNOWLEDGMENTS

DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited. This material is based upon work supported by the Under Secretary of Defense for Research and Engineering under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Under Secretary of Defense for Research and Engineering.

REFERENCES

[1] 2023. Network Common Data Form (NetCDF). https://www.unidata.ucar.edu/software/netcdf/. Accessed online: 2023-04-10.
[2] David Alvarez-Melis and Tommi S. Jaakkola. 2018. On the Robustness of Interpretability Methods. http://arxiv.org/abs/1806.08049 arXiv:1806.08049 [cs, stat].

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

[4] Erin K. Chiou and John D. Lee. 2021. Trusting automation: Designing for responsivity and resilience. *Human Factors* (April 2021). https://doi.org/10/gjvcr2 Publisher: SAGE Publications Inc.

[5] Nancy J. Cooke, Jamie C. Gorman, Christopher W. Myers, and Jasmine L. Duran. 2013. Interactive team cognition. *Cognitive Science* 37, 2 (March 2013), 255–285. https://doi.org/10/gf69qb

[6] Hans de Bruijn, Martijn Warnier, and Marijn Janssen. 2022. The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. *Government Information Quarterly* 39, 2 (April 2022), 101666. https://doi.org/10.1016/j.giq.2021.101666

[7] Upol Ehsan and Mark O. Riedl. 2021. Explainability Pitfalls: Beyond Dark Patterns in Explainable AI. https://doi.org/10.48550/arXiv.2109.12480 arXiv:2109.12480 [cs].

[8] Elliot E. Entin and Daniel Serfaty. 1999. Adaptive Team Coordination. *Human Factors* 41, 2 (June 1999), 312–325. https://doi.org/10.1518/001872099779591196 Publisher: SAGE Publications Inc.

[9] Leilani H. Gilpin, Andrew R. Paley, Mohammed A. Alam, Sarah Spurlock, and Kristian J. Hammond. 2022. "Explanation" is Not a Technical Term: The Problem of Ambiguity in XAI. https://doi.org/10.48550/arXiv.2207.00007 arXiv:2207.00007 [cs].

[10] Robert S. Gutzwiller and John Reeder. 2017. Human interactive machine learning for trust in teams of autonomous robots. In *2017 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*. 1–3. https://doi.org/10.1109/COGSIMA.2017.7929607 ISSN: 2379-1675.

[11] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 57, 3 (May 2015), 407–434. https://doi.org/10/f68kpx

[12] John D. Lee and Neville Moray. 1994. Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies* 40 (1994), 153–184. https://doi.org/10/d95kwv

[13] John D. Lee and Katrina A. See. 2004. Trust in automation: Designing for appropriate reliance. *Human Factors* (2004), 31. https://doi.org/10.1518/hfes.46.1.50_30392

[14] Susan Leigh Star. 2010. This is Not a Boundary Object: Reflections on the Origin of a Concept. *Science, Technology, & Human Values* 35, 5 (Sept. 2010), 601–617. https://doi.org/10.1177/0162243910377624 Publisher: SAGE Publications Inc.

[15] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

[16] Christopher A. Miller. 2021. Trust, transparency, explanation, and planning: Why we need a lifecycle perspective on human-automation interaction. In *Trust in Human-Robot Interaction*, Chang S. Nam and Joseph B. Lyons (Eds.). Academic Press, 233–257. https://doi.org/10.1016/B978-0-12-819472-0.00011-3

[17] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (Feb. 2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[18] Rohan Paleja, Muyleng Ghuy, Nadun Ranawaka Arachchige, Reed Jensen, and Matthew Gombolay. 2021. The Utility of Explainable AI in Ad Hoc Human-Machine Teaming. In *Advances in Neural Information Processing Systems*, Vol. 34. Curran Associates, Inc., 610–623. https://proceedings.neurips.cc/paper/2021/hash/05d74c48b5b30514d8e9bd60320fc8f6-Abstract.html

[19] 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*. ACM, San Francisco California USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[20] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (May 2019), 206–215. https://doi.org/10.1038/s42256-019-0048-x Number: 5 Publisher: Nature Publishing Group.

[21] Pouria Salehi, Erin K. Chiou, Michelle Mancenido, Ahmadreza Mosallanezhad, Myke C. Cohen, and Aksheshkumar Shah. 2021. Decision deferral in a human-AI joint face-matching task: Effects on human performance and trust. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 65. SAGE Publications Inc, 638–642. https://doi.org/10/gpg29p

[22] Kristin E. Schaefer, Jessie Y. C. Chen, James L. Szalma, and P. A. Hancock. 2016. A Meta-Analysis of Factors Influencing the Development of Trust in Automation: Implications for Understanding Autonomy in Future Systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 58, 3 (May 2016), 377–400. https://doi.org/10.1177/0018720816634228

[23] Susan Leigh Star. 1989. The Structure of Ill-Structured Solutions: Boundary Objects and Heterogeneous Distributed Problem Solving. In *Distributed Artificial Intelligence*, Les Gasser and Michael N. Huhns (Eds.). Morgan Kaufmann, San Francisco (CA), 37–54. https://doi.org/10.1016/B978-1-55860-092-8.50006-X

[24] Harini Suresh, Steven R. Gomez, Kevin K. Nam, and Arvind Satyanarayan. 2021. Beyond Expertise and Roles: A Framework to Characterize the Stakeholders of Interpretable Machine Learning and their Needs. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3411764.3445088

[25] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. 2020. The What-If Tool: Interactive Probing of Machine Learning Models. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (Jan. 2020), 56–65. https://doi.org/10.1109/TVCG.2019.2934619 Conference Name: IEEE Transactions on Visualization and Computer Graphics.