ՏԵՂԵԿԱՏՎԱԿԱՆ ՏԵԽՆՈԼՈԳԻԱՆԵՐ

HUMAN-CENTERED AI AND EXPLAINABLE DECISION SUPPORT

https://doi.org/10.59982/18294359-25.2-hc-18

Arman Grigoryan
Phd in Technical Sciences, Associate Professor
EUA, Chair of Information Technologies and Applied Mathematics
a.grigoryan@eua.am

Abstract

The deployment of artificial intelligence (AI) systems in consequential domains, such as healthcare, criminal justice, finance, and others, has raised urgent questions about transparency, fairness, and human agency. While Explainable AI (XAI) techniques offer methods to interpret model predictions, they have often been designed primarily for model developers rather than for end users making critical decisions. Human-centered AI (HCAI) advocates a fundamentally different approach: systems should be designed around human needs, values, and autonomy, with explanation as a means to support human reasoning rather than merely to justify algorithmic outputs.

This paper synthesizes current research at the intersection of HCAI and explainable decision support. It examines what explanations mean in socio-technical contexts, reviews empirical evidence on how different explanation types and interfaces affect user understanding and decision quality, and proposes design principles for systems that preserve meaningful human control while leveraging AI's analytical strengths. The paper argues that genuinely human-centered explainable decision support requires moving beyond technical interpretation of models toward collaborative dialogue that respects human expertise, acknowledges uncertainty, and enables people to maintain their judgment capacities over time.

Keywords: Explainable AI (XAI), Human-Centered AI (HCAI), Explainable Decision Support, Human-AI Collaboration, Domain-Specific Explainability, Adaptive Explanation.

Introduction

Nowadays, AI systems influence high-stakes decisions affecting millions of people. Machine learning models assess creditworthiness, predict criminal recidivism, diagnose diseases, and determine hiring and promotion outcomes [Selbst and Barocas]. While these systems can achieve impressive predictive accuracy, they often operate as “black boxes” whose internal logic is difficult to understand, even for technical experts. When such systems make errors or perpetuate biases, those affected by the decisions have little recourse. They cannot easily understand why they were denied a loan, flagged as high-risk, or rejected for employment. On the other end, regulators and organizations have responded by demanding explainability. The European Union’s General Data Protection Regulation grants individuals a “right to explanation” for automated decision-making [Selbst and Barocas]. The U.S. Equal Employment Opportunity Commission has issued guidance on algorithmic discrimination. Industry leaders and researchers have called for interpretable and explainable AI as a matter of both ethics and practical necessity [Gunning and Aha].

However, much work in XAI has focused on making models intelligible to developers, data scientists, and regulators (those with technical backgrounds and responsibility for building systems). Less attention has been paid to whether and how explanations actually help people who receive decisions from these systems, or who use them in professional contexts such as clinical care or loan underwriting. A radiologist does not want to understand the mathematical operations of a deep convolutional neural network but wants to know whether the system’s suggestion to further investigate a lesion is trustworthy for the specific patient, and whether the judgment based on the model is reliable or seeks additional evidence [Caruana et al.]. Similarly, a loan officer needs to understand not just why an applicant was flagged as high-risk, but whether that assessment aligns with his institution's lending principles and whether there are extenuating circumstances [Selbst and Barocas].

This paper is motivated by the persistent gap between technical explainability and effective human-centered decision support. In this paper, we adopt the perspective of human-centered AI (HCAI), which argues that AI systems should be designed around human abilities, values, and autonomy rather than around what is technically elegant or computationally efficient [Shneiderman]. In this regard, explanations are not merely artifacts that expose model mechanics, they are interactions within a socio-technical system that include humans, institutions, data, algorithms, and governance structures. Thus, an explanation's value depends on whether it actually helps people understand relevant uncertainties, make better decisions, maintain their own expertise, and hold systems accountable [Miller].

From Interpretability to Human-Centered Design

XAI and its Limitations. XAI refers to techniques and practices aimed at making AI system behavior understandable to humans [Doshi-Velez and Kim]. These techniques span a spectrum from intrinsically interpretable models to post-hoc explanation methods. On the one hand, intrinsically interpretable approaches favor simpler models (such as linear regressions, decision trees, or sparse rule lists) that humans can directly understand [Caruana et al.]. On the other hand, post-hoc methods attempt to explain complex "black-box" models by approximating their behavior with simpler surrogates (such as LIME), computing feature-importance scores (such as SHAP), or generating counterfactual scenarios [Lundberg and Lee], [Ribeiro et al.].

XAI has made genuine progress in developing technical methods for interpretation. However, research in human-computer interaction and social science reveals important limitations [Miller]. First, many XAI techniques were developed primarily for model developers, regulators, and auditors, rather than for end users receiving automated decisions. A feature-importance score that makes sense to a machine-learning engineer may be opaque or even misleading to a clinician or loan officer [Caruana et al]. Second, explanations are not self-interpreting. Even when a system outputs an explanation, whether a human understands it, trusts it, and acts upon it depends on countless factors: the user's background knowledge, the clarity of language, the visual design, the context, and the user's expectations about what the system can do [Liao et al.]. A technically sound explanation can fail to communicate if it does not align with how users think about the problem. Third, focusing narrowly on model interpretability can obscure other critical questions about fairness, accountability, and power. An explanation that helps a bank understand why its lending algorithm rejected an applicant does not necessarily help the applicant contest that decision or remedy a biased system [Selbst and Barocas].

Human-Centered AI as a Different Paradigm. Human-centered AI starts from a different premise, i. e. AI systems should be designed around human needs, values, and capabilities rather than around what is technically possible [Shneiderman]. This perspective emphasizes several key commitments. First, humans should remain meaningfully in control. Rather than seeking to maximize automation, HCAI advocates for systems where automation and human control coexist at high levels [Shneiderman]. A clinician should not simply defer to an AI system's diagnosis but should be able to understand the reasoning, compare it to their own clinical judgment, and make an informed treatment decision. Second, HCAI emphasizes reliability and trustworthiness. This goes beyond technical accuracy to encompass robustness across different contexts, alignment with stated purposes, and transparent communication of limitations and uncertainties [Shneiderman]. Third, systems should support human agency and responsibility. Rather than shifting blame to "the algorithm," human-centered design preserves space for human judgment, professional discretion, and moral deliberation. Finally, HCAI considers governance, regulation, and institutional embedding. No algorithm exists in isolation; it is embedded in organizations with norms, power structures, and accountability mechanisms. Thus, human-centered design must address these broader contexts [Selbst and Barocas].

In the context of decision support, HCAI implies that explanations should do more than expose model mechanics. They should support human reasoning about uncertainty, trade-offs, and alternatives. They should help users understand when the system is likely to be reliable and when they should be skeptical. They should enable users to learn from interactions with the system, improving their own domain knowledge over time. And they should support accountability by making decisions contestable and facilitating oversight.

Human-Centered Explainable Decision Support

Explainable decision support can serve multiple, sometimes competing goals depending on context. Distinguishing these goals clarifies what explanations should accomplish and how they should be evaluated.

Understanding and mental models. A primary goal is to help users develop accurate mental models of what the system does and does not do [Miller]. This includes understanding what features the system considers, what patterns it recognizes, and what kinds of cases it handles well versus poorly. Users with more accurate mental models make better decisions about when to rely on the system and when to seek additional information or override its recommendations [Liao et al].

Trust calibration and appropriate reliance. Research on automation in aviation and other safety-critical domains shows that both over-reliance and under-reliance on automated systems create risks [Parasuraman and Riley]. Ideally, users should rely on AI when it is likely correct and ignore it when it is likely wrong. However, people often either blindly defer to systems (automation bias) or reflexively distrust them (algorithm aversion) [Dietvorst et al.]. Explanations can support calibrated trust by revealing uncertainty, explaining limitations, and demonstrating that the system has made errors in the past [Bansal et al].

A simple formalization of calibrated trust can be expressed as:

$$T_{\text{cal}} = \frac{1}{1 + e^{-\alpha(p-\theta)}}, \tag{1}$$

where p is the AI's predicted probability of a correct outcome, θ - a decision-maker-specific trust threshold, and α- controls the steepness of the trust response. When p exceeds θ, T_{cal} approaches 1 (high trust); when p falls below θ, T_{cal} approaches 0 (low trust). This equation captures the idea that trust should be a smooth, probabilistic function of the system's confidence rather than a binary "trust/don't trust" judgment.

Usability and cognitive efficiency. Users often make decisions under time pressure and cognitive constraints. Explanations that are too dense or complex can overwhelm rather than inform, leading users to ignore them or rely on simple heuristics [Buçinca et al.]. Human-centered explanations must balance completeness with conciseness, offering key information without imposing excessive cognitive load. A linear model of cognitive load C might be:

$$C = \beta_0 + \beta_1 N_{\text{feat}} + \beta_2 D, \tag{2}$$

where N_{feat} is the number of features presented, D- the depth (or number of layers) of the explanation, and β_0, β_1, β_2 are empirically estimated coefficients. Designers can use this relationship to limit N_{feat} and D so that C stays within acceptable bounds for the target user group.

Fairness, accountability, and contestation. In contexts where AI-supported decisions have significant consequences for individuals (such as criminal justice, hiring, or credit decisions), transparency serves a moral and legal purpose. Explanations enable affected individuals to understand decisions, identify potential discrimination, and exercise recourse [Selbst and Barocas]. They also support institutional auditing and regulatory compliance.

Human learning and expertise development. Over time, repeated interactions with an AI system can either enhance or degrade human expertise. If explanations are designed to help users understand underlying domain structure, recognize patterns they might miss, and refine their own heuristics, the system can support human learning [Buçinca et al.]. Conversely, if users simply defer to the system without understanding it, they may become passive and dependent.

Organizational integration and legitimacy. AI systems do not operate in isolation but within organizational workflows, professional norms, and institutional cultures. Explanations that align with how domain professionals (such as clinicians, engineers, managers) naturally think about problems are more likely to be trusted and integrated into practice [Caruana et al.]. Explanations that feel alien or imposed can generate resistance regardless of their technical quality.

Human-centered explainable decision support must attend to all these goals, recognizing that they may sometimes be in tension. An explanation that is highly detailed and trustworthy for a technical auditor may overwhelm a frontline worker, making time-pressured decisions. An explanation that preserves human learning by encouraging reflection may reduce short-term decision speed. Designing effective explanation systems requires understanding the specific context, stakeholders, and trade-offs at hand.

Different types of explanations have been studied in decision-support research. Feature-based explanations indicate which input variables most influenced a prediction, based on methods such as LIME or SHAP [Lundberg and Lee], [Ribeiro et al.]. Uncertainty-focused explanations explicitly communicate what the system is unsure about, rather than presenting predictions as deterministic [Liao et al.]. This is especially important in domains such as medicine, where uncertainty is inevitable, and users need to understand the limits of algorithmic confidence.

Empirical studies comparing explanation types show mixed results, with effectiveness depending on task, user, and context [Bansal et al.], [Buçinca et al.]. No single explanation type works best across all situations. This suggests that human-centered systems should offer multiple explanation modalities and allow users to request the type that best suits their needs and reasoning style.

Emerging Frameworks for Human-AI Collaboration

Recent work proposes richer conceptualizations of how humans and AI can collaborate in decision-making, moving beyond the simple "system recommends, human decides" dynamic [Ma et al.].

Deliberative frameworks. Ma et al. propose that AI systems should support structured deliberation, where humans and AI discuss different dimensions of a decision, explore disagreements, and iteratively refine their judgments. Rather than AI producing a final recommendation and humans accepting or rejecting it, the system engages in dialogue, asking clarifying questions, and revising its position based on human input. Such deliberative approaches have shown promise in supporting both better decisions and deeper human learning [Ma et al.].

Information complementarity. Guo focuses on explanations that highlight information the AI system possesses that humans tend to miss or underweight. Rather than explaining how the model works in general, these explanations emphasize what the model has discovered about this particular case that should inform the human's reasoning. This shifts focus from model transparency to decision relevance [Guo].

Conversational and question-driven explanation. Molaei et al. emphasize that users have diverse information needs. Rather than presenting predetermined explanations, systems should support user-driven questioning, allowing people to ask what they want to know about decisions. This requires more sophisticated natural-language understanding and generation, but can provide more tailored, relevant explanations [Molaei et al.].

Human-in-the-loop learning. Rather than treating the trained AI model as fixed, human-centered systems can update and adapt based on human feedback. When a human overrides or questions a recommendation, the system can learn from that, improving its future behavior. This creates a feedback loop where human expertise and AI learning reinforce one another [Liao et al.].

These frameworks suggest that genuinely human-centered decision support involves more active, ongoing collaboration rather than one-off interactions where AI produces output and humans consume it.

Design Principles for Human-Centered Explainable Decision Support

Based on the above-mentioned review, several design principles can be proposed.

1. ***Start from user needs, not algorithmic capabilities.*** Grounded explanation design in a careful study of what users actually need to know to make good decisions in their specific context. Involve target users as collaborators, shaping what gets explained and how.
2. ***Design for calibrated trust, not maximum trust.*** Aim for a reliability that appropriately reflects actual system reliability. Be transparent about uncertainty, acknowledge past errors, and help users understand the boundaries of the system's competence.
3. ***Manage cognitive load through layering and adaptation.*** Provide concise, high-level explanations by default, with optional drill-down into greater detail for users who want or need it. Adapt depth and style to user expertise, time constraints, and past behavior. Use interactive rather than purely static explanations when users have time to engage.
4. ***Align explanations with domain reasoning and professional practice.*** Co-design with domain professionals so that explanations reflect how they naturally think about problems (e.g., patient trajectories in medicine, precedent in law).
5. ***Preserve space for human judgment and learning.*** Design explanations that highlight unexpected patterns, prompt reflection, and encourage users to refine their own heuristics over time.
6. ***Build in contestation and feedback mechanisms.*** Enable two-way dialogue so that affected individuals can question decisions, provide corrective information, and improve the system.
7. ***Integrate explanations into broader governance and accountability structures.*** Support not only individual decision-makers but also managers, auditors, and regulators, recognizing that different stakeholders require different explanatory content.

Conclusion

Despite progress, substantial gaps remain in our understanding of how to design explainable decision support that genuinely serves human needs. Key research directions demonstrate that XAI and human-centered system design are not opposed but complementary. Technical methods for interpreting models are valuable, but insufficient. What ultimately matters is whether explanations help humans understand decisions, maintain appropriate reliance, learn over time, and exercise meaningful control.

This requires shifting focus from explaining algorithms to supporting human reasoning within organizational and societal contexts.

The research and practice reviewed in this paper point toward a conception of explainable decision support as collaborative dialogue. Rather than AI systems producing final answers that humans accept or reject, humans and AI engage in ongoing interaction where each contributes complementary strengths. AI brings pattern recognition across vast datasets, consistency, and freedom from some human biases. Humans bring contextual understanding, moral judgment, accountability, and the ability to recognize when situations fall outside the system's competence. Explanations mediate this collaboration, enabling each partner to understand and learn from the other.

Building such systems demands sustained interdisciplinary effort. Computer scientists must develop new explanation methods and interfaces that serve human reasoning, not just model interpretation. Human-computer interaction researchers must study how explanations actually influence decision-making in real-world contexts. Cognitive scientists must investigate human reasoning processes and how to support them. Domain professionals must participate in design, bringing their expertise and tacit knowledge. Ethicists and policy scholars must ensure that systems align with the values of fairness, autonomy, and accountability.

As AI systems influence ever more important decisions (e.g., who receives medical treatment, credit, employment, and freedom), getting explainable decision support right becomes increasingly urgent. The alternative (opaque systems that people either blindly trust or reject) serves neither human interests nor organizational integrity. Human-centered explainable decision support offers a path forward, one that harnesses AI's strengths while preserving human judgment, learning, and accountability as central to consequential decision-making.

References

1. Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., & Horvitz, E., "Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance", *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2021, 1-16.
2. Buçinca, Z., Malaya, M. B., & Gajos, K. Z., "Too Much Assistance? A Framework for Understanding the Impact of AI Assistance on Human Decision-Making", *Proceedings of the International Conference on Intelligent User Interfaces*, 2021, 309-319.
3. Caruana, R., Lou, Y., Johansson, F., Snelson, E., & Saria, S., "Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-Day Readmission", *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, 1721-1730.
4. Dietvorst, B. J., Simmons, J. P., & Massey, C., "Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err", *Journal of Experimental Psychology: General*, 2015, 144(1), 114-126.
5. Doshi-Velez, F., & Kim, B., "Towards a Rigorous Science of Interpretable Machine Learning", *arXiv preprint*, 2017, arXiv:1702.08608.
6. Guo, Z., "The Value of Information in Human-AI Decision-Making", *arXiv preprint*, 2025, arXiv:2502.06152.
7. Gunning, D., & Aha, D., "DARPA'S Explainable Artificial Intelligence (XAI) Program", *AI Magazine*, 2019, 40(2), 44-58.
8. Liao, Q. V., Gruen, D., & Miller, S., "Questioning the AI: Informing Design Practices for Explainable AI User Experiences", *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2020, 1-15.
9. Lundberg, S. M., & Lee, S.-I., "A Unified Approach to Interpreting Model Predictions", *Advances in Neural Information Processing Systems*, 2017, 30, 4765-4774.

10. Ma, S., Chen, Q., Wang, X., Zheng, C., Peng, Z., Yin, M., & Ma, X., “Towards Human-AI Deliberation: Design and Evaluation of LLM-Empowered Deliberative AI for AI-Assisted Decision-Making”, *Proceedings of CHI*, 2025.
11. Miller, T., “Explanation in Artificial Intelligence: Insights from the Social Sciences”, *Artificial Intelligence*, 2019, 267, 1-38.
12. Molaei, S., Robert, L. P., & Banovic, N., “What Do People Want to Know about Artificial Intelligence (AI)? The Importance of Answering End-User Questions to Explain Autonomous Vehicle (AV) Decisions”, *Proceedings of the ACM on Human-Computer Interaction*, 2025, 9(7), CSCW256.
13. Parasuraman, R., & Riley, V., “Humans and Automation: Use, Misuse, Disuse, Abuse”, *Human Factors*, 1997, 39(2), 230-253.
14. Ribeiro, M. T., 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, 1135-1144.
15. Selbst, A. D., & Barocas, S., “The Watered Down Lemonade Stand: Algorithmic Fairness Soup”, *Data & Society Research Institute*, 2019.
16. Shneiderman, B., “Human-Centered Artificial Intelligence: Reliable, Safe and Trustworthy”, *International Journal of Human-Computer Interaction*, 2020, 36(6), 495-504.

Submitted on 15.10.2025.
Sent for review on 09.12.2025.