Received: 23 June 2025 | Revised: 4 November 2025 | Accepted: 17 November 2025

DOI: 10.1002/aaai.70043

ARTICLE

Human at the Center: A Framework for Human-Driven AI Development

Danniell Hu[1] | **Diana Acosta Navas**[2] | **Susanne Gaube**[3] | **Hussein Mozannar**[4] | **Matthew E. Taylor**[5,6] | **Krishnamurthy Dvijotham**[7] | **Elizabeth Bondi-Kelly**[1]

[1]Computer Science and Engineering, University of Michigan, Ann Arbor, Michigan, USA

[2]Quinlan School of Business, Loyola University Chicago, Chicago, Illinois, USA

[3]Global Business School for Health, University College London, London, UK

[4]AI Frontiers, Microsoft Research, Redmond, Washington, USA

[5]Computing Science, University of Alberta, Edmonton, AB, Canada

[6]Alberta Machine Intelligence Institute (Amii), Edmonton, AB, Canada

[7]Science and Strategic Initiatives, Google DeepMind, Mountain View, California, USA

Correspondence
Danniell Hu, Computer Science and Engineering, University of Michigan, Ann Arbor, MI, USA.
Email: dannihu@umich.edu

Funding information
Google; Alberta Innovates; Alberta Machine Intelligence Institute; CIFAR AI Chair, Amii; Digital Research Alliance of Canada; Mitacs; National Science and Engineering Research Council; University of Michigan Computer Science and Engineering Fellowship

Abstract

Artificial Intelligence (AI) systems increasingly shape many aspects of daily life, influencing our jobs, finances, healthcare, and online content. This expansion has led to the rise of human–AI systems, where humans communicate, collaborate, or otherwise interact with AI, such as using AI outputs to make decisions. While these systems have shown potential to enhance human capabilities and improve performance on benchmarks, evidence suggests that they often underperform compared to AI-only or human-only approaches in experiments and real-world applications. Here, we argue that human–AI systems should be developed with a greater emphasis on human-centered factors—such as usability, fairness, trust, and user autonomy—within the algorithmic design and evaluation process. We advocate for integrating human-centered principles into AI development through human-centered algorithmic design and contextual evaluation with real users. Drawing on interdisciplinary research and our tutorial at two major AI conferences, we highlight examples and strategies for AI researchers and practitioners to embed these principles effectively. This work offers a systematic synthesis that integrates technical, practical, and ethical insights into a unified framework. Additionally, we highlight critical ethical considerations, including fairness, labor, privacy, and human agency to ensure that systems meet performance goals while serving broader societal interests. Through this work, we aim to inspire the field to embrace a truly human-centered approach to algorithmic design and deployment.

INTRODUCTION

From traditional predictive models to modern generative models, AI systems now shape decisions in critical domains such as healthcare, employment, and civic life, raising urgent questions about how to align these systems with human values. Human-AI systems—which we define as any AI system involving human interaction, whether as decision support or in collaborative contexts—have emerged as a promising approach to ensure human guidance, involvement, oversight, and correction. They also help align decisions with societal values, promote fairness through additional context, and respect individual and community autonomy in decision making (Floridi et al. 2018; Selbst et al. 2019; Shneiderman 2022; Wu et al. 2022). For these reasons, experts have signaled the potential of these systems to improve outcomes, over the expected performance of pure AI-driven or pure human-driven decisions (Shneiderman 2022).

This potential has been realized in several applications. For instance, in medical diagnostics, physicians using AI assistance provided better diagnoses than either humans or AI alone (Hu et al. 2024; McDuff et al. 2025; Reverberi et al. 2022; Yun et al. 2023). In peer-to-peer mental health support, AI-enhanced collaborations increased conversational empathy, greatly empowering participants in therapeutic settings (Sharma et al. 2023). Similarly, in online political discussions, human–AI systems improved conversation quality, democratic reciprocity, and tone (Argyle et al. 2023).

However, success in human–AI systems is not universal. In some cases, the combination of AI and human oversight has led to worse outcomes than relying on either alone (Buçinca et al. 2021; McDuff et al. 2025). Users may over-trust AI recommendations even when they are clearly incorrect (Buçinca et al. 2021), or abandon helpful input entirely after a single error due to “algorithm aversion” (Klingbeil et al. 2024). Others may under-rely on accurate AI guidance, reducing overall effectiveness by disregarding correct recommendations (Gaube et al. 2024).

In addition, these interaction failures can combine with AI systems’ preexisting biases in ways that generate misalignment between the system’s outcomes and the goals or values of its human users. For instance, Obermeyer et al. (2019) found that a widely used healthcare algorithm underestimated the health risks of Black patients, exacerbating disparities in access to care. This bias originated from the algorithm’s use of healthcare cost as a proxy for health needs, demonstrating the breakdown between developing the algorithm and ultimately deploying it in a sociotechnical system.

These challenges raise a fundamental question for researchers and developers: **While human-AI systems hold great promise to enhance human capabilities, why do they still struggle with alignment and/or performance across different applications?** We argue that building effective and ethically aligned human–AI systems requires moving beyond traditional performance and speed benchmarks by incorporating human-centered approaches to algorithmic development and testing. These approaches should take into consideration what it means for humans to be involved in the development and deployment of AI systems while also being their overarching aim, instead of being merely a means to their improvement (Floridi 2022). This requires addressing questions related to ethical labor conditions for crowd-workers, data privacy, fairness, and how interactions with AI over time can impact human autonomy and agency. We refer to *human agency* as a person’s capacity to act intentionally, make meaningful choices, and exert control over actions and outcomes in their environment (Bandura 2006).

By emphasizing these concerns, human-centered methods can advance well-being, rights, and values throughout the entire AI lifecycle—from problem identification and data preparation to design, testing, deployment, monitoring, scaling, and optimization.

Specifically, we propose:

1. **Human-Centered Algorithmic Design** to better align systems with how humans actually interact with AI, and
2. **Testing with Real Users** to empirically define what “success” means in specific domains and real-world settings.

In this article, we synthesize existing literature to integrate concepts that have been independently validated and offer an evidence-based, structured framework of human-centered algorithmic design and user evaluation strategies, weaving together technical, practical and ethical considerations. We advocate for human-centered development and deployment, and particularly emphasize three interconnected elements: **algorithmic design**, **evaluation with real users**, and **ethical reflection**. Design shapes how systems interact with and influence people; evaluation grounds this in real-world contexts; and ethical reflection contributes to make both design and evaluation more responsive to broader questions of fairness, labor ethics, privacy, and human agency. Together, these elements provide a foundation for building AI systems that advance human interests and values. By adopting these human-centered approaches, we aim to advocate for human–AI systems that are both effective and deeply aligned with human values and societal needs.

RELATED WORK

Human-centered design (HCD) has a rich history rooted in participatory design and Human–computer interaction (HCI), where the end users’ needs, values, and contexts are prioritized in system development. Research on user-centered design, with its emphasis on usability and user experience (Mao et al. 2005; Norman 2013), laid the foundation for building systems that align with human capabilities and expectations. Participatory design movements have also emerged, which advocate for the inclusion of stakeholders—particularly marginalized groups—in the co-creation of technologies to ensure empowerment and equity (Asaro 2000; Bjögvinsson et al. 2012; Costanza-Chock 2020; Spinuzzi 2005). Over time, these principles have been formalized in frameworks such as ISO 9241-210 (Mirnig et al. 2015), which defines HCD as an iterative process involving understanding user needs and requirements, followed by prototyping and testing. This approach has been successfully applied across diverse domains, from consumer software to healthcare systems, demonstrating its ability to improve adoption, satisfaction, and equity (Steen 2012).

The field of AI ethics has made substantial advances by furthering conceptual specifications for human-centered algorithmic design. These advances address multiple dimensions, including fairness in algorithmic decision-making, respect for data subject privacy, and the legitimacy of automated decisions (Barocas et al. 2023; Nissenbaum 2009). Developments in AI ethics have been adopted in regulations and recommendations as varied as the European AI Act and the UNESCO standards for AI (both of which center on the notion that AI would prioritize human rights) to the Vatican's AI Ethics guidelines, which prioritize human dignity, common good, and responsibility in the development and use of AI. Other fields, such as science and technology studies, have significantly advanced our understanding of human-centered AI by highlighting the societal contexts shaping technology. Research in value-sensitive design has likewise advanced our understanding of how human values may be embedded and encoded into technological design.

As AI systems have proliferated, adapting HCD principles to address unique algorithmic challenges has become increasingly critical. Historically, while HCI and social sciences have embraced participatory and value-sensitive design, the AI community has prioritized technical performance metrics, such as accuracy and generalization (Birhane et al. 2022; Bommasani et al. 2021; Chang et al. 2024; Ethayarajh and Jurafsky 2020; Liang et al. 2022).

Furthermore, many AI researchers and developers rely on simulations, benchmarks, or historical data for testing, typically due to the difficulties in testing with real-world systems, humans, or physical environments (Afzal et al. 2020; Birhane et al. 2022; Eriksson et al. 2025; Singh et al. 2025).

However, recent literature underscores ethical and practical motivations for integrating human-centered principles into algorithmic design, emphasizing fairness, transparency, and accountability to mitigate harms and align with societal goals. Researchers have advocated participatory approaches to identify and mitigate biases in AI systems (Bondi et al. 2021; Chen et al. 2023). Others emphasize the importance of transparency and trust in enabling effective human–AI collaboration (Endsley 2023; Vössing et al. 2022). Additionally, domain-specific evaluations have been highlighted as essential for addressing the contextual nature of user interactions (Bondi et al. 2022; Haque et al. 2023). Frameworks for human–AI collaboration (Fragiadakis et al. 2024; Gomez et al. 2025) and human–AI interaction have also emerged (Amershi et al. 2019).

Despite growing recognition of the need for human-centered approaches, adoption across the AI community remains uneven. This article synthesizes practices and strategies to provide researchers and practitioners with structured, actionable guidance to embed human-centered algorithmic design and evaluation methods across a broad spectrum of human–AI contexts. We emphasize effectiveness, ethical alignment, and responsiveness to real-world human needs.

IMPLEMENTING HUMAN-CENTERED ALGORITHMIC DESIGN

Building on the diverse strategies explored in prior work, we introduce a prescriptive framework to guide researchers in implementing human-centered algorithmic design. This framework categorizes human–AI collaboration based on (1) the level of interaction between users and AI systems and (2) how decision (or final outcome, e.g., prediction, recommendation, action) authority is distributed between them. The framework also considers the *agenticness* of a system, defined as the degree to which a system can adaptably achieve complex goals in complex environments with limited direct supervision (Shavit et al. 2023).

By structuring design decisions around these questions, developers can more deliberately determine when and how to involve users, domain experts, or stakeholders, depending on the demands of the task or application. While these may blur in practice, they offer a useful conceptual foundation for analyzing and designing human-centered systems. We identify four example modes that illustrate this space, shown in Figure 1.

1. **AI Supports Human Decisions**: In this mode, AI generates predictions, recommendations, or insights to aid human decision-making, but the human is accountable for the final decision.
2. **AI Decides Unless Uncertain**: In this mode, AI primarily functions independently but defers to a human or abstains from making a recommendation altogether when uncertain.
3. **AI Decides with Human Input**: In this mode, AI actively seeks additional input from humans or other sources to refine its decision-making when confidence is low.
4. **AI and Humans Decide Jointly**: In this mode, humans and AI engage in continuous, bidirectional

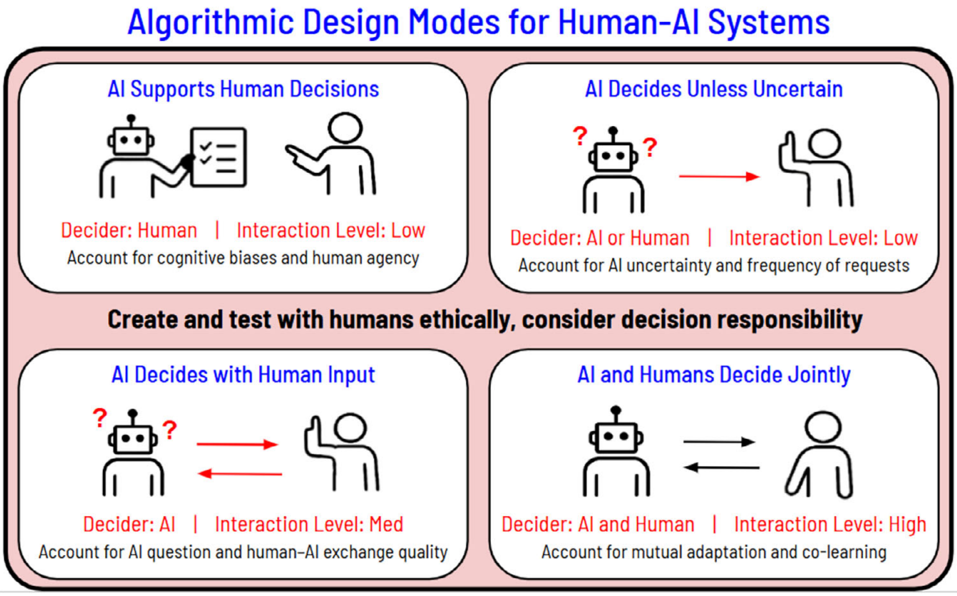

FIGURE 1 Algorithmic design modes for human–AI systems.

interaction with each other and the environment in real-time without requiring explicit prompts.

These modes provide a high-level conceptual framing for system design, but their application is not one-size-fits-all. Selecting and using the appropriate mode requires thoughtful attention to the specific context, available resources, and goals of the system in question. For example, multi-agent settings may make use of multiple modes and raise additional considerations.

The following subsections present a non-exhaustive set of systems within each mode, along with human-centered considerations that are particularly relevant for each.

AI supports human decisions

We define "AI Supports Human Decisions" as a mode in which AI aids human decision making by generating predictions, recommendations, or insights, while humans maintain final decision authority. AI behaves as a static tool providing recommendations, but always defers final judgment to humans.

Systems in the "AI Supports Human Decisions" mode aim to reduce human workload, enhance decision accuracy, mitigate cognitive biases, and standardize decision processes (Howard 2019; Küper et al. 2024; Li et al. 2025).

To illustrate design components of decision support systems, we break this mode down into the four dimensions of decision scope, decision mechanisms, decision outputs, and decision explanations (Gomez et al. 2025; Lee et al. 2020). These components are relevant across many modes of AI–human interaction, but are foundational to decision support systems.

Decision Scope

Decision scope refers to the types and complexity of decisions that an AI system supports. For example, decision scope may range from narrow, well-defined tasks such as classifying images or recommending products, to more complex and open-ended decisions such as selecting medical treatments or assessing financial risk.

Decision Mechanisms

Decision mechanisms encompass the technical methodologies used to generate AI outputs. For example, rule-based systems, with explicit if-then logic defined by experts, offer high transparency and auditability, but have limited flexibility (Masri et al. 2019). Shallow machine learning models, such as logistic regression (Shipe et al. 2019), decision trees (De Ville 2013), and support vector machines (Suthaharan 2016) balance interpretability and accuracy, making them suitable for structured data with regulatory transparency requirements. Models may also be more complex, in which case decision sets can provide interpretable versions that are better suited to complex domains (Lakkaraju et al. 2016).

Decision Outputs and Explanations

Decision outputs are actionable results, such as classifications, predictions, or recommendations, that AI systems provide to users. Decision outputs range from direct recommendations (e.g., recommended medical treatments) to more abstract insights (e.g., predicted financial risk).

Communication of these outputs impacts user trust, understanding, and the eventual decision quality (Hassija et al. 2024; Miller 2019). For example, communicating uncertainty can significantly influence human decision-making (Kim et al. 2024).

Explainable AI (XAI) techniques promote transparent and understandable outputs. Transparency ranges from inherently interpretable models such as linear regression, to opaque models like neural networks, which require post-hoc explanation methods (e.g., SHAP (Salih et al. 2025; Sundararajan and Najmi 2020)).

Explanations play a critical role in shaping user trust and reliance, but their impact depends heavily on timing and context. Properly timed explanations, delivered simultaneously with or immediately following AI outputs, can significantly influence user trust and system effectiveness (Hemmer et al. 2021; Rong et al. 2023). If not delivered carefully, they can harm overall performance (Jabbour et al. 2023; Jacobs et al. 2021).

Ali et al. (2023); Dwivedi et al. (2023) offer a detailed taxonomy to guide the selection of explanation strategies based on user needs and system goals. Doshi-Velez and Kim (2018) provide a set of principles for the evaluation of interpretability and helps researchers understand factors that may make tasks similar in their explanation needs. Green and Chen (2019) also outline critical considerations for decision-support design: accuracy enhancement, reliability in calibrating user trust, and fairness to prevent reinforcing societal biases or inequities. Other algorithms show promise in supporting AI-assisted human decision-making, such as by strategically tailoring the display of AI recommendations based on the case and/or individual (Buçinca et al. 2024; Swaroop et al. 2025).

Considerations for AI Supports Human Decisions

Challenges

- Preservation of human agency: Risks of over-reliance on AI outputs (Passi and Vorvoreanu 2022) or under-reliance (Gaube et al. 2024).
- Cognitive biases: AI recommendations can introduce or amplify cognitive biases such as automation bias (Alon-Barkat and Busuioc 2023), algorithm aversion (Dietvorst et al. 2015), selective adherence (Alon-Barkat and Busuioc 2023), etc.

Design Considerations

- Prioritize transparency and explainability by communicating explanations with appropriate timing and method (Hassija et al. 2024; Hemmer et al. 2021; Kim et al. 2024; Rong et al. 2023; Miller 2019).
- Use cognitive forcing functions and adaptive algorithms to encourage critical engagement with AI outputs (Vasconcelos et al. 2023; Buçinca et al. 2021, 2024; Schemmer et al. 2023; de Jong et al. 2025).
- Avoid cognitive overload (Schemmer et al. 2022) and be sensitive to potential amplification of societal biases (Green and Chen 2019).

AI decides unless uncertain

The "AI Decides Unless Uncertain" mode differs from "AI Supports Human Decisions" systems by placing AI as the primary (but not only) decision maker. These systems independently make final decisions within well-defined confidence thresholds, but abstain from acting or defer to humans when encountering uncertainty or complexity beyond their designed capabilities.

Systems in this mode implicitly acknowledge that even highly capable AI models have limitations, such as knowledge gaps, biases, and difficulty handling edge cases. As such, humans remain essential in scenarios that require moral reasoning, regulatory compliance, domain-specific expertise, and more. Rather than constantly seeking validation, these systems allocate tasks dynamically: they act independently when confident, but either abstain or defer to a human when uncertainty is high.

Two prominent approaches under this mode are: **selective prediction (abstention)** and **deferral**. These methods allow AI systems to manage uncertainty by calibrating their decision boundaries and determining when human input is required.

Selective Prediction

Selective prediction, also known as *abstention*, allows a model to opt out of making predictions on uncertain inputs, prioritizing reliability over forced decision making (Geifman and El-Yaniv 2017). The foundational work of Chow (1970) framed abstention as a trade-off between error rate and rejection frequency, formalized through the Error–Reject Tradeoff curve. Wiener and El-Yaniv (2013) later introduced the risk-coverage function to quantify this trade-off in probabilistic models, showing that abstaining on low-confidence inputs can significantly improve overall accuracy.

Selective prediction depends on effective uncertainty estimation. For instance, recent work has proposed methods for Large Language Models (LLMs) to detect knowledge gaps and abstain from answering, reducing hallucinations and improving reliability (Feng et al. 2024). In addition, the way uncertainty and decision status are communicated to users can influence system effectiveness, as absent or unclear signals may undermine trust and lead to misuse. (Bhatt et al. 2021; Bondi et al. 2022; Prabhudesai et al. 2023).

Deferral

Deferral extends the concept of selective prediction by assuming that if the AI abstains, a human expert will be available to resolve the case (Madras et al. 2018). Rather than "doing nothing," the system actively transfers uncertain or high-risk instances to a human collaborator. Mozannar and Sontag (2020) characterized deferral as a cost-sensitive learning framework, balancing the trade-off between automation errors and the burden on human experts. The AI system must decide when to act independently and when to defer, considering both the potential impact of a mistake and the availability of human expertise. A key insight from this work is that deferral can be trained as a standard machine learning problem using consistent surrogate losses (Mozannar and Sontag 2020; Zhou 2011). In contrast to such learning-based approaches, other deferral mechanisms use simple rule-based models that defer when the AI model's confidence falls within an optimized score interval. These thresholds can be chosen via brute force search to maximize overall system accuracy while respecting a predefined constraint on the allowable deferral rate (Bondi et al. 2022). These approaches enable the AI to learn when deferral is most beneficial, minimizing both prediction errors and human burden.

Deferral is particularly relevant in domains where human expertise can mitigate the AI's limitations. Additional work has looked at deferring to (multiple) experts strategically based on their skills (Wilder et al. 2021; Verma et al. 2023; Mao et al. 2023). Another prominent implementation of this strategy is the complementarity-driven deferral to clinical workflow (CoDoC) system (Dvijotham et al. 2023). CoDoC combines AI predictions with clinician expertise to improve diagnostic accuracy. For instance, in breast cancer and tuberculosis screening, CoDoC defers uncertain cases to clinicians, seamlessly integrating into clinical workflows to optimize decision-making and improve patient outcomes. In the case where the AI is confident and does not defer, the final prediction is made by the model and a human is not in the loop—ultimately reducing clinician workload.

Considerations for AI Decides Unless Uncertain

Challenges

- Communicating uncertainty and deferral status: Missing or unclear AI confidence or deferral cues can lead to inappropriate reliance and undermine user trust (Bondi et al. 2022; Bhatt et al. 2021; Prabhudesai et al. 2023).
- Confidence calibration: How well an AI's expressed confidence in its outputs aligns with its true correctness can affect collaboration (Li et al. 2024).
- Human readiness and skill erosion: Users must be ready to take over when the system defers, and they need to maintain the skills required to do so. Skill erosion can increase the risk of errors (Rinta-Kahila et al. 2023).

Design Considerations

- Consider communication of deferral cases in the system design, e.g., show deferral status alone without the model's predictions. Communicating deferral status alone led to significantly higher human accuracy compared to providing no information, while showing the model's prediction reduced human accuracy (Bondi et al. 2022), though this should be tested in each context.
- Present model confidence using a method that aligns with context and user needs. Examples include frequency-based messages, probability scores and confidence intervals, or verbal expressions of uncertainty (Prabhudesai et al. 2023; Xu et al. 2025; Zhang et al. 2020).

AI decides with human input

The "AI Decides with Human Input" mode is characterized by a higher level of agenticness compared to "AI Decides Unless Uncertain" systems. In this setting, the AI holds primary decision authority. These systems proactively detect uncertainty, strategically engage humans through targeted interactions, incorporate responses, and then resume decision-making. We highlight two strategies for these systems: **Selective Clarification** and **Incorporating Humans in Training**.

Selective Clarification

Selective clarification enables systems to proactively identify ambiguity or uncertainty, prompt users with targeted clarifying questions, and seamlessly integrate responses into the decision process. This strategy prioritizes interactions with high informational value, ensuring that human input occurs only when it is significantly beneficial (Kuhn et al. 2023; Zhang & Choi 2023).

For instance, CLAM (Kuhn et al. 2023) applies selective clarification in natural language question-answering tasks by prompting language models to detect ambiguity, formulate specific clarifying questions, and refine answers based on user responses. Similarly, INTENT-SIM (Zhang & Choi 2023) leverages simulated interactions to estimate whether a clarification will significantly improve model performance, explicitly distinguishing between epistemic (knowledge-based) and aleatoric (inherent) uncertainty. Both approaches exemplify cooperative principles: efficiently improving performance without burdening users with unnecessary interactions.

The evaluation of selective clarification focuses on downstream performance improvements, frequency and costs of interactions, and generalizability across contexts. Designing clear, concise, and impactful questions is essential, as poorly structured queries can lead to user confusion or ineffective responses (Rahmani et al. 2024).

Incorporating Humans in Training: Concept Bottleneck Models and Interactive Machine Learning

Another potential way to implement a system in this mode is via concept bottleneck models (CBMs). CBMs predict high-level intermediate concepts, which are then used to predict the final class label. The goal is to improve interpretability and enable humans to correct intermediate concepts to improve classification performance. For example, in a bird species classifier, a model might first predict interpretable concepts such as "has red head," or "has long beak," which are then combined to predict the species label; a human can correct an intermediate concept if it was mispredicted. This is shown to increase the overall classification performance on various tasks (Koh et al. 2020). Recent work improves CBMs' accuracy, handling uncertainty, intervenability, concept discovery, and application to retrieval tasks (Balloli et al. 2024; Chauhan et al. 2023; Espinosa Zarlenga et al. 2023; Kim et al. 2023; Shang et al. 2024; Sheth and Ebrahimi Kahou 2023; Yuksekgonul et al. 2022).

Similarly, interactive machine learning is another paradigm in which users may provide feedback to AI systems during training, for example, by providing explanations of queries and decision making for potential correction (Teso and Kersting 2019). Further work on testing methods with humans could benefit such frameworks. For example, CBMs and adjacent models could improve complementarity by accounting for human strengths and weaknesses. Further correction opportunities, for both humans and AI systems to review and receive feedback, are another potential direction discussed in Section 3.4.

Incorporating Humans in Training: Reinforcement Learning with Human Feedback

Reinforcement learning with human feedback (RLHF) operationalizes adaptation by embedding human guidance directly into the learning process. Building on foundational reinforcement learning frameworks (Szepesvári 2022), RLHF introduces three principal feedback mechanisms: reward shaping through scalar evaluations (as pioneered in the TAMER framework (Knox and Stone 2009)), preference ranking via pairwise comparisons (Christiano et al. 2017a), and real-time corrective feedback during task execution (MacGlashan et al. 2017; Ouyang et al. 2022). Critical to RLHF's success is feedback efficiency, or the ratio of human judgments required per unit of behavioral improvement, which Christiano et al. (2017b) show can be optimized through recursive reward modeling. Efficiency can be improved via active preference sampling (Das et al. 2024) and reward model reuse (Ziegler et al. 2020). However, RLHF has limitations, such as the potential for misalignment and misgeneralization (Casper et al. 2023), and it can be extractive—relying heavily on human labor for feedback without fair compensation (Gonzalez-Cabello et al. 2024).

Considerations for AI Decides with Human Input

Challenges

- Formulating easily understandable questions: Vague or overly complex questions may reduce the likelihood of accurate human input (Rahmani et al. 2024).
- Minimizing unnecessary interruptions: Avoiding over-questioning while reserving the benefits of human input. Interaction mechanisms must be optimized to prevent excessive interruptions and response fatigue (Zou et al. 2020, 2023).
- Respecting human contributors: Ensure contributors are respected, compensated, and representative of end-user populations (Bondi et al. 2021; Hawkins and Mittelstadt 2023).

Design Considerations

- Use well-calibrated uncertainty estimators or interaction triggers to minimize unnecessary interruptions (Kim et al. 2021; Kuhn et al. 2023; Mu et al. 2024; Testoni and Fernández 2024).
- Design concise, high-impact exchanges (e.g., clarification questions) that are easy for users to interpret and act upon (Rahmani et al. 2024).

AI and humans decide jointly

The "AI and Humans Decide Jointly" mode combines high AI agenticness with sustained human engagement. Rather than alternating control at discrete checkpoints, humans and AI agents adapt to each other in real time, continuously updating strategies and mental models. We highlight two prominent design strategies in this mode: **AI- and**

Human-in-the-Loop and **Mixed-Initiative Collaboration**.

AI- and Human-in-the-Loop

As we have seen, humans play a central role in many human–AI systems and can interact with AI in varied ways, including interactive learning, preference elicitation, and probabilistic model learning (Natarajan et al. 2025). These interactive forms of learning enable bidirectional collaboration, where both the AI and human adapt over time. For example, Amershi et al. (Amershi et al. 2014) describe a system that translates video of arm movements into sound. Through interaction, the system is tuned to match an expert user's preferences for how different movement parameters (such as position, speed, or rotation) map to sound, while simultaneously providing implicit feedback that helps the user refine their movement precision.

Other promising examples include iterative preference elicitation for personalized decision support (De Toni et al. 2022), and dynamic fine-tuning of multi-armed bandit resource allocation policies by public health decision-makers (Behari et al. 2024).

Mixed-Initiative Collaboration

Mixed-initiative (MI) collaboration is an interaction style in which humans and other agents can both take the initiative to start, steer, interrupt, or relinquish control, allowing each partner to contribute what it does best at any given moment (Hearst et al. 1999; Horvitz 1999). Four design hallmarks characterize effective MI systems:

- **Dynamic initiative switching** enables either party to propose actions, ask clarifying questions, or defer.
- **Negotiation of roles** means that responsibilities are continually re-allocated rather than fixed in advance.
- **Complementary strengths** generally allows humans to provide context, judgment, creativity, and so forth, while agents provide speed, memory, computation, and so forth.
- **Transparency & intention recognition** encourages each side to model the other's goals well enough to decide when to act and when to stay silent.

Originally explored in dialogue and planning research in the late 1990s, MI collaboration can now be found in many human–AI systems. Code assistants such as GitHub Copilot [1], productivity suites like Notion AI [2], and proactive LLM chat assistants (Chen et al. 2025) surface context-aware suggestions that users can accept, edit, or ignore. Other MI AI tools such as *Coalesce* help civic leaders craft locally relevant survey and interview questions (Overney et al. 2025). In human–robot teaming, MI systems dynamically adjust control between manual, shared, and autonomous modes to suit real-time task demands and human preferences (Chanel et al. 2020; Jiang and Arkin 2015).

Considerations for AI and Humans Decide Jointly

Challenges

- Misaligned mental models: Humans and AI may develop inaccurate understandings of each other's goals or capabilities over time, which can reduce coordination and system effectiveness (Bansal et al. 2019).
- Risks of over-personalization: Excessive adaptation to user behavior can encourage existing biases and reduce exposure to diverse options. (Kirk et al. 2024).

Design Considerations

- Promote mutual adaptation and co-learning through bidirectional feedback loops and sustained transparency (Huang et al. 2019; Kumar et al. 2024; Lu et al. 2025). Design systems to support continual learning on both sides: AI models should adapt to user behavior while helping users build accurate mental models of the system (Andrews et al. 2023).
- Carefully calibrate personalization to avoid reinforcing user biases or creating filter bubbles (Areeb et al. 2023; Stray 2023). Personalization can improve user experience, but should be implemented with safeguards to prevent narrowing perspectives or amplifying existing biases.

EVALUATION WITH REAL USERS

While direct evaluation with users provides invaluable insights, it can be challenging due to costs and iteration timelines (Natarajan et al. 2025). As a result, many researchers and practitioners rely on proxies such as simulations and benchmarks for evaluation (Birhane et al. 2022).

Unfortunately, these methods have clear limitations. Simulations often fail to capture the complex emotional, social, and strategic responses of real users (Bondi et al. 2022; Gaube et al. 2021; Jacobs et al. 2021; Montemayor et al. 2022; Schröder et al. 2025). This disconnect can result in unexpected outcomes after deployment, as human behavior diverges from the model's assumptions (Beede et al. 2020). Historical datasets may reflect outdated norms and systemic biases. For example, Amazon's automated

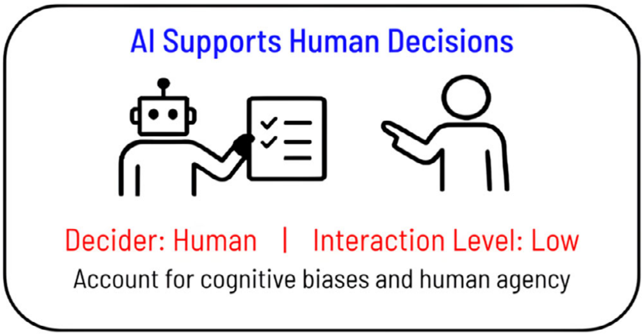

FIGURE 2 AI supports human decisions.

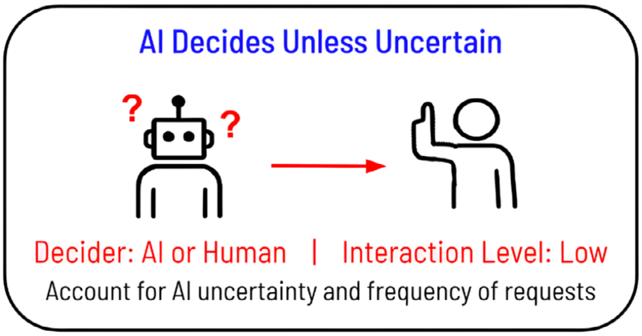

FIGURE 3 AI decides unless uncertain.

resume screener penalized female applicants (Dastin 2022), highlighting how historical data can perpetuate discrimination when applied to new contexts.

Moreover, important constructs such as trust, perceived fairness, and cognitive burden cannot be directly inferred from system logs. These aspects must be discovered through qualitative methods such as think-aloud protocols, structured interviews, or others (Adams 2015; Adeoye-Olatunde and Olenik 2021; Jaspers et al. 2004; Jääskeläinen 2012).

To underscore the importance of human evaluation, consider parallels from other established domains. In medicine, regulation mandates test in animal models and then rigorous human trials of drugs to demonstrate safety and efficacy FDA (c, b). In the cosmetics industry, the seemingly innocuous, low-risk products can cause allergic reactions, disrupt hormonal balance, or produce long-term skin damage (Bilal and Iqbal 2019; Khan and Alam 2019), necessitating testing (Barthe et al. 2021; FDA a). In short, we do not accept simulations of human biology as a substitute for actual evidence of human impact. These examples illustrate a critical point: even when a system or product shows promising performance in a simulation or benchmark, its effects on real humans can be significant and possibly harmful (Figures 2–5).

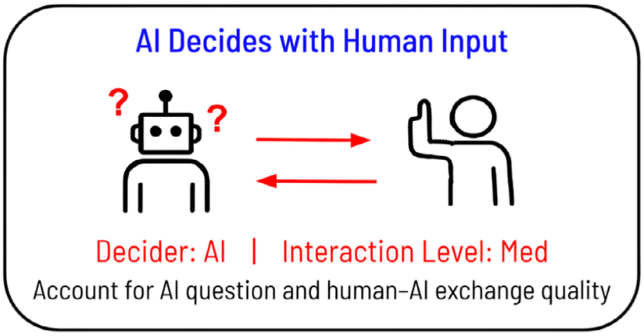

FIGURE 4 AI decides with human input.

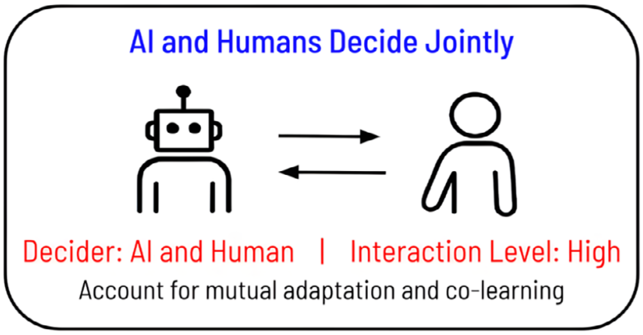

FIGURE 5 AI and humans decide jointly.

We argue that critical human–AI systems must also be evaluated with real users before large-scale deployment.

While approaches may vary depending on context, certain methodological considerations consistently shape the quality and reliability of human–AI evaluations. In Figure 6, we provide a non-exhaustive guide to designing user studies. These guidelines draw from best practices and tutorials in human–computer interaction (HCI), AI, and social science research (Amershi et al. 2019; Gray et al. 2023; Lazar et al. 2017; Pargent et al. 2024; Walliman 2021). We also analyze a human study by Gaube et al. (2024) as a case study to illustrate the guidelines, using it simply as one example among many possible ways to conduct human–AI studies. In sharing this overview, we aim to support AI researchers in understanding common steps toward developing robust, reproducible, and responsible human–AI evaluations, while acknowledging that the context of individual systems can significantly vary, and direct collaboration of AI researchers with HCI and social science researchers is highly valuable.

A case study: Gaube et al

To illustrate the guidelines in Figure 6, we turn to a recent example in the healthcare domain. Gaube et al. (2024) investigated how medical experts and novices interact

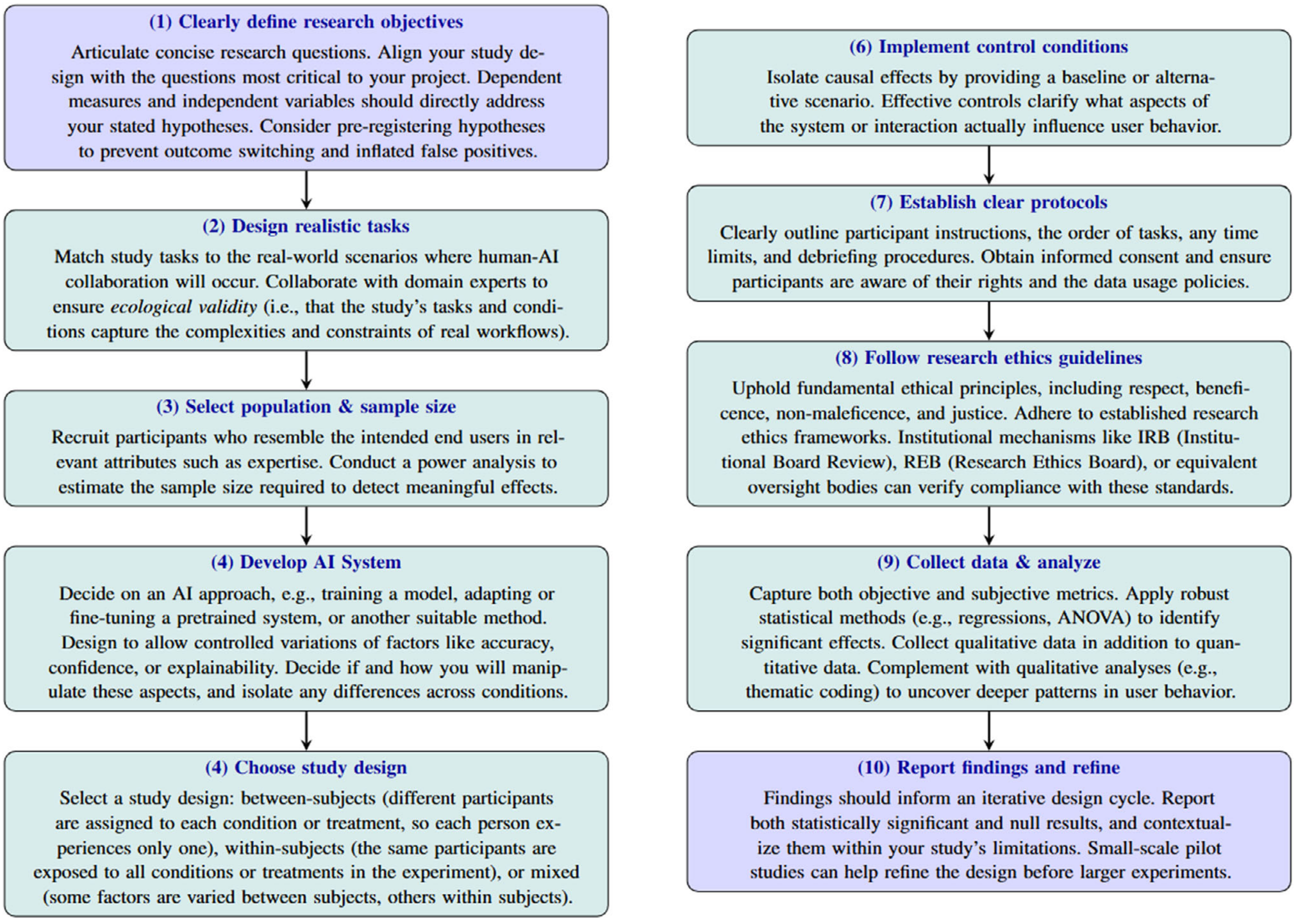

FIGURE 6 Flowchart of study design process.

with AI-generated advice in the context of diagnosing intracranial hemorrhage (ICH) from head CT scans.

1. Define research Objectives and Hypotheses

The authors formulated a set of hypotheses and research questions to guide their investigation across four outcome domains: diagnostic performance, perceived advice quality, diagnostic confidence, and review time. They hypothesized that correct AI advice would enhance performance and confidence, and be perceived as higher quality than incorrect advice. Task experts were expected to be more confident and faster than novices. Central to the study was an examination of how explainability (XAI vs. basic advice), advice accuracy (correct vs. incorrect), and user expertise influenced these outcomes. Importantly, the study also aimed to understand how these effects mapped onto overreliance (following incorrect advice) and underreliance (rejecting correct advice), as two distinct but critical failure modes in human–AI collaboration.

2. Design realistic tasks

Participants reviewed real-world patient cases containing head CT scans of varying complexity (easy, medium, and difficult) and determined the presence or absence of ICH. The interactive user interface closely resembled clinical workflows, and the advice was framed as generated by an AI-based decision support system. This setup ensured an ecologically valid scenario comparable to real-world decision-making in healthcare.

3. Select population

125 participants from 10 countries were recruited and divided into two main groups:

- Experts: Radiologists and radiology residents (year two and above).
- Novices: Non-radiologist physicians, interns, and medical students.

This differentiation enabled the researchers to investigate how domain expertise influences reliance (including overreliance and underreliance) on AI advice and its impact on XAI for both groups. Sample size was determined using a tailored simulation-based approach for generalized linear mixed models (GLMMs), accounting for the study's hierarchical structure and various outcome metrics (Pargent et al. 2024). Adequate sample planning is important for ensuring robust and informative results.

4. Develop AI System
The study did not employ a live AI model in order to systematically manipulate the advice to be either correct or incorrect (in an 80:20 ratio), and delivered either as basic predictions or as XAI advice, including localization annotations. This allowed the study to isolate the effects of advice accuracy and explainability on user decisions.

5. Choose study design
A multi-session crossover design was implemented. Each participant reviewed the same 50 CT cases in three sessions (with a minimum 14-day washout period, or time in between sessions), under three experimental conditions (within-subject design):

- Control: No AI advice.
- Basic Advice: Prediction only.
- XAI Advice: Prediction with bounding-box annotation.

Randomization ensured balance in the order of exposure across participants. To complement the quantitative data, a subset of participants took part in a think-aloud (novices) or an eye-tracking (experts) study arm. These additions provided valuable insight into the cognitive and perceptual mechanisms underpinning overreliance and underreliance, enhancing the interpretability of behavioral outcomes.

6. Implement control conditions
The study included baseline (no advice) sessions and controlled the accuracy and presentation of advice across all experimental conditions.

7. Establish clear protocols
Participants received standardised instructions. In each session, they completed all 50 cases, recorded a binary diagnosis (ICH: yes/no), rated their confidence, and, in advice conditions, evaluated the usefulness of the advice. A subset completed think-aloud or eye-tracking protocols for deeper process insight. The full study protocol, including hypotheses, design, and analysis plan, was preregistered to ensure transparency and reduce analytical flexibility.

8. Follow research ethics guidelines
All participants provided informed consent and were informed about the study's objectives, including its focus on human–AI interaction in diagnostic decision-making. The study received institutional ethics approval. AI advice was simulated for the purpose of the experiment and did not affect clinical care. Participants who expressed interest received individual feedback on their performance after completing the study.

9. Collect data and analyze
The study collected:

- Objective measures: Diagnostic accuracy, reading time.
- Subjective measures: Confidence ratings, perceived usefulness.
- Process measures: Eye-tracking fixations (experts) and verbal reasoning (novices).

Mixed-effects models were used to assess how advice accuracy, format, and participant expertise affected decision-making outcomes (see Gaube et al. 2024 for more details).

10. Report findings and refine
The study revealed that underreliance occurred more frequently than overreliance due to the higher base rate of correct advice, and it had a greater negative impact on diagnostic performance. Explainable AI (XAI) was effective in reducing underreliance when it stemmed from uncertainty, particularly in difficult, true positive cases. However, it had limited impact when underreliance was driven by distrust or disengagement. Notably, novices benefited the most from XAI, and in some difficult cases, outperformed experts when supported by correct XAI advice. These findings suggest that XAI can enhance human–AI collaboration but must be tailored to specific user groups and contexts. The authors recommend designing human–AI systems with mechanisms to mitigate both overreliance and underreliance, and to move beyond binary agree/disagree interfaces that oversimplify clinical decision-making.

This case study illustrates how user studies that adhere to structured methodological guidelines can generate meaningful insights into human-AI collaboration (in this case, in a clinical setting).

ETHICAL CONSIDERATIONS

Although integrating humans into AI systems can improve performance and provide deeper insights into human–AI interaction, it should be sensitive to a broad range of ethical considerations. For example, labor and privacy issues become more pronounced when humans participate in tasks such as data annotation, domain-specific evaluations, or continuous interactive engagement with AI systems. Recognizing and addressing these concerns helps align AI development with societal values and underscores our broader call to build more consistent, trustworthy human–AI collaborations. Although this section cannot address these issues in full depth and breadth, it emphasizes important contributions in the literature and adds context to some of the crucial considerations discussed throughout this article.

Participation

Due to the large-scale and high-stakes impact of AI systems in society, it is critically important to consider the aggregate effects of AI systems and how their impact is distributed among different social groups, particularly in light of existing social inequalities and structural oppression. This section lays out some of the considerations that are most central for human–AI collaboration.

Fairness in Participation, Process, and Outcomes

Fairness is a multidimensional concept central to the ethical integration of AI systems, encompassing participation, process, and outcomes:

- **Fairness in participation** refers to the opportunity given to all stakeholders to contribute to decision-making processes (see: Costanza-Chock (2020)).
- **Procedural Fairness** emphasizes the procedures and methods employed in making consequential decisions (Morse et al. 2022; Wang et al. 2024).
- **Fairness in outcomes** refers to scenarios in which results are impartial and avoid disproportionately benefiting or burdening specific groups (see Barocas et al. (2023)).

While these categories are analytically independent, there are important connections between them. Including diverse human voices in human–AI system development mitigates the risks of bias introduced by narrow or homogeneous perspectives (for a discussion of risks, see the works of Bender et al. 2021 and Weidinger et al. 2021). Providing an opportunity for all relevant stakeholders to have a voice can help mitigate both procedural and outcome unfairness. For example, involving diverse annotators in data labeling tasks can uncover implicit biases in training data that might otherwise perpetuate inequities in AI models. To illustrate this, the WinoQueer benchmark (Felkner et al. 2023) recruited survey respondents from the LGBTQ+ community—the very group affected by the biases being measured—to develop a benchmark for identifying anti-queer bias in LLMs. By incorporating real-world concerns and lived experiences from this community, the benchmark revealed harmful patterns in LLMs, such as misrepresentation, stereotyping, and exclusion. Including humans with domain expertise and diverse lived experiences during the evaluation phase can also help identify and address the inequities resulting from AI systems.

However, relying on human judgment introduces its own challenges. Human evaluations are inherently noisy and heterogeneous: even domain experts often disagree, and aggregate measures such as simple averages can conceal systematic polarization or bias in opinion (Kahneman 2011). Rather than treating "the human perspective" as a single source of ground truth, evaluations should characterize the full distribution of human responses. Reporting dispersion statistics (e.g., standard deviations, confidence intervals) and inter-rater agreement metrics (McDonald et al. 2019) makes this variability visible and allows researchers to assess the consistency and diversity of human perspectives. Such methodological practices are well established in HCI research (Lazar et al. 2017). Building on this recognition, recent work explores not only how to measure disagreement but also how to mediate and synthesize it. Emerging approaches to structured preference aggregation, such as AI-assisted deliberation systems like the Habermas Machine (Tessler et al. 2024) demonstrate that disagreement among humans has the potential to be transformed into a more representative collective judgment.

Participation, by itself, is insufficient

While human involvement is an important step in creating better human–AI systems, it is important to highlight that it is not, by itself, sufficient to guarantee that the systems will be fair or legitimate In fact, systems that employ human feedback can encode biases depending on how they are built and what kind of feedback they seek. A prominent case is that of MIT's Moral Machine Experiment. The Nature article (Awad et al. 2018) presents findings from a large-scale online survey that collected over 40 million moral decisions from people in 233 countries and territories about autonomous vehicle crash scenarios. The study reveals significant cultural variations in ethical preferences, for example, some countries prioritize young people over the elderly or pedestrians over passengers, highlighting the complexity of creating universally accepted moral guidelines for AI systems. Although this study was conducted to show how ethical judgment varies in such scenarios, it was strongly criticized (Jaques 2019) for its methodological focus on individual decision-making, overlooking societal biases and the potential structural effects of aggregating such individual choices.

More broadly, participatory approaches have been criticized for being a form of "ethics washing" or "participatory washing" (Bietti 2020; Birhane et al. 2022). Traditional participatory design is based on extensive qualitative stakeholder consultations, which are localized and sensitive to specific communities and their context. This context-sensitivity is what provides developers with valuable information and ensures the systems deployed have sufficient legitimacy in the communities where they are deployed. However, due to the rapid scalability of AI, the impact of the systems is likely to reach far beyond the contexts in which stakeholders were consulted. Hence, both the information and legitimacy gains obtained through the consultation are quickly diluted (Sloane et al. 2022).

Enabling Democratic Governance

The ideal of democratic governance of AI is to shape the development, deployment, and regulation of AI systems to reflect the values and priorities of a diverse range of stakeholders. This approach promotes deliberation among parties with a broad range of values and interests, acknowledges value pluralism, and leverages collective intelligence (The Collective Intelligence Project 2024) for a better alignment of AI systems. In this way, it upholds democratic principles (Ovadya 2023), and creates necessary conditions for a better distribution of agenda-setting and decision-making power (Acosta-Navas 2025).

To the extent that AI systems are likely to have large-scale impact on matters of public interest, it is of fundamental importance to establish procedures that shift decision power toward the public. Incorporating large-scale stakeholder consultations can lead to AI systems that better serve societal interests, insofar as it provides developers with a more fine-grained, nuanced understanding of the opinion landscape among parties who may be impacted by these systems. For instance, OpenAI launched an initiative that sought to emphasize the importance of public input in shaping the goals and constraints of its AI systems (OpenAI 2023). Similarly, Anthropic's "Collective Constitutional AI" initiative involves aligning language models with broad public values (Huang 2024). These examples illustrate how more democratic forms of governance, facilitated by AI systems, can enhance the alignment of AI systems with the values of a broader and more diverse set of stakeholders.

Labor and crowdsourcing

Human labor remains indispensable in many "last mile" tasks that support AI development, including data labeling, content moderation, dataset curation, and much of the human evaluation we discuss in this article. Researchers frequently enlist large-scale annotation services or rely on crowdsourcing platforms such as Amazon Mechanical Turk (MTurk) Turk (Turk) and Prolific Prolific (2024) to recruit workers for these tasks. Despite their central role, these workers often remain invisible within organizational structures, with little public understanding of the labor that underpins AI systems. Studies show that workers on MTurk, sometimes called "Turkers," can receive wages well below minimum wage thresholds, frequently hovering around $2 per hour (Hara et al. 2018; Newman 2019; Toxtli et al. 2021). The term "ghost work" has been coined to describe this hidden pool of workers, whose contributions are fundamental to the performance and credibility of AI (Casilli 2025; Gray and Suri 2019; Roberts 2019; Williams et al. 2022). Beyond low wages, wage theft and insufficient recourse options for tasks that are rejected without explanation further strain the well-being of workers (Alkhatib and Bernstein 2019).

A striking example of the difficult conditions faced by these workforces was exposed when OpenAI outsourced sensitive content annotation to the Kenyan firm Sama, where employees were paid less than 2 USD per hour to label explicit or disturbing content, risking psychological harm for employees (Perrigo 2023). In a similar vein, Roblox, a massively popular gaming platform, profits from the unpaid or underpaid labor of children who create games on the platform. Although marketed as a tool for creativity and entrepreneurship, the system heavily favors Roblox financially, offering young developers minimal compensation despite their significant contributions (Parkin 2022). More broadly, the use of human labor creates the risk of treating human crowd workers as mere means of improving AI systems (Altenried 2020).

Some platforms have taken steps to address these issues and serve as examples of more equitable practices. Prolific, for example, sets minimum pay thresholds aligned with local minimum wages, offers transparency in task requirements, and limits arbitrary rejections of completed work (Prolific 2024). Providing fair compensation, outlining explicit guidelines, and instituting mechanisms for dispute resolution is critical for sustaining a committed and high-quality workforce. Offering mental health resources for workers exposed to graphic or disturbing content can mitigate long-term psychological harm. Acknowledging the contributions of annotators, moderators, and participants, rather than hiding them behind opaque organizational structures, not only respects the individuals performing these tasks but also ensures the integrity and trustworthiness of AI systems in the eyes of users and the broader public.

Automation and the Future of Alignment Labor

The growing demand for human effort across labeling, evaluation, and oversight has made the sustainability of alignment labor a pressing concern. As AI systems expand in scope and complexity, so too does the volume of human input required. In response, researchers and organizations increasingly turn to automation to reduce repetitive workload and accelerate iteration in the alignment process.

Recent advances in LLM-based classifiers, policy-aware prompts, and model-assisted review have improved the speed, scalability, and consistency of moderation and evaluation (Chen et al. 2025; Huang 2025; Palla et al. 2025). Automated systems can pre-filter content, prioritize cases for human review, and generate structured feedback that accelerates retraining. (Lykouris and Weng 2024)

However, automation introduces its own limitations. Current systems still struggle with nuances such as author intent, shifting social norms, and contextual interpretation beyond short text. Intent, in particular, remains central to most platform policies yet is poorly captured in datasets and models, leading to false positives and negatives in the absence of richer conversational, user, and policy context (Wang et al. 2025).

To address these limitations, hybrid approaches that combine algorithmic triage with human-in-the-loop review (Li et al. 2025) have been created in the hopes of combining the best of both worlds. However, such hybrid systems remain, at their core, human–AI systems. As such, they are subject to the same risks outlined in this paper (Section 1), and therefore require careful design and evaluation through human-centered methods.

Privacy

Privacy concerns intensify when humans become more directly involved in the AI lifecycle. Every interaction with an AI system—be it by annotators, domain experts, evaluators, or end-users—can result in data collection that may reveal sensitive personal or contextual information. Researchers and practitioners often rely on techniques such as anonymization and encryption to safeguard these datasets. However, the effectiveness of these measures can vary widely. Even when data is partially anonymized, there is a risk that it could be combined with publicly or privately available datasets to reveal information that contributors did not intend to disclose (Crawford and Schultz 2014a; Véliz 2020).

Ordinarily, informed consent is used as a way of preserving autonomy and control over personal data. However, Barocas and Nissenbaum (2014) and Crawford and Schultz (2014b) identify two key limitations of informed consent in big data systems. First, the fidelity–simplicity trade-off, where accurate representations of data use are too complex for users to understand, and simplified notices fail to convey meaningful information. Second, the tyranny of the minority, where AI systems can infer information about individuals who did not consent to data sharing by drawing on patterns learned from those who did, as well as from data the individual consented to share for unrelated purposes. These limitations reveal that informed consent, as traditionally conceived, cannot meaningfully protect privacy in the context of large-scale data analytics.

The iterative and adaptive nature of many AI deployments compounds these risks. Context-based interactions in real-world scenarios often involve clarifying questions or follow-up prompts that elicit additional data, which may be sensitive. Traditional consent processes can become less effective in these settings, particularly if users are not fully informed about how their data can be used, aggregated, or inferred. Ensuring privacy requires consistent, robust technical safeguards, including encryption at rest and in transit, restricted access to sensitive datasets, and rigorous privacy audits to identify and remediate vulnerabilities over time. Data minimization, illustrated by projects such as CooP (Dvijotham et al. 2023), offers an important framework in which AI systems only request the information needed to improve specific decisions or predictions by asking clarifying questions or additional information only when needed. By narrowing the scope of the data that are collected, these systems can build trust without sacrificing performance. Truly informed consent also benefits from clear explanations of how data might be used or shared, enabling contributors to weigh the potential benefits against the risks and to decline participation if they deem those risks unacceptable.

Other responses to these risks include (1) federated learning, in which central models are trained on local user devices and only model updates are shared centrally for aggregation and central model refinement, and (2) differential privacy, a mathematical framework that limits the risk of revealing information about any individual by introducing carefully calibrated noise into aggregate outputs. Both approaches are discussed in (Yan et al. 2024).

Labor and privacy practices directly shape the reliability, trustworthiness, and overall acceptance of AI systems. Exploitative labor arrangements fail to respect the dignity of workers, while weak privacy safeguards deter people from engaging meaningfully with AI, limiting both the diversity of data and the scope of possible improvements. Conversely, strategies such as paying fair wages, providing psychological support for workers, implementing robust encryption and anonymization protocols, and minimizing data collection can sustain healthy interactions between humans and AI.

Human agency

Another benefit of involving humans in the development and evaluation of AI systems is the preservation of space for human choice (Costanza-Chock 2020). For example, protecting spaces where human judgment is necessary from being invaded by automated decisions and allowing humans to determine where exercising discretion is preferable to delegating decisions to automated systems or being passive recipients of AI outcomes.

This is particularly important given the potential ways in which sustained interaction with AI systems can reshape human agency. On the one hand, excessively outsourcing judgment to AI systems may have a detrimental impact on human agency, both in the short term and through the cumulative impact of outsourcing decisions to AI systems. For instance, as AI is increasingly used for content moderation and fact-checking in social media platforms, users may increasingly distrust their own judgment on the contents they are exposed to (Coeckelbergh 2023; Navas 2024).

On the other hand, AI systems have the potential to support human agency by allowing users to be more efficient and perform tasks in a way that further advances their goals and ends, potentially improving human decision-making, much like assisted driving systems. More agentic systems may further enhance individuals' ability to promote their own goals as well as those of broader society. However, ethical challenges may arise if these systems behave in a paternalistic manner—supplementing, guiding, or overriding user choices. As we have argued throughout this article, this tension raises critical questions about who should decide which values and goals AI systems amplify, especially since individual users may lack complete information about how their choices aggregate at scale.

CONCLUSION

Achieving trustworthy, effective, and ethically grounded human–AI systems requires more than improving model performance or technical sophistication. It demands a fundamental reorientation of the AI development process around human-centered values, goals, and contexts. In this article, we have argued for two core strategies to support this shift: **human-centered algorithmic design** and **evaluation with real users**. We have also highlighted the ethics of involving humans in AI systems, emphasizing the need to design these interactions responsibly. Together, these approaches offer a structured foundation for building systems that not only function well but also align with the needs, capabilities, and rights of the people they are meant to serve.

Through our framework of interaction modes: AI Supports Human Decisions, AI Decides Unless Uncertain, AI Decides with Human Input, and AI and Humans Decide Jointly, we outlined how different modes of human-AI collaboration present distinct design challenges and opportunities. By explicitly considering factors such as the degree of interaction between human and AI, where decision authority falls, and the agenticness of the AI system, developers can more deliberately match system behavior to human expectations and domain demands. We also offered practical guidance for evaluating these systems through real-world user studies, emphasizing the importance of ecological validity, iterative refinement, and empirical definitions of success grounded in human outcomes.

Importantly, we argue that human involvement is not only a methodological necessity but also an ethical imperative. Decisions about system behavior, data collection, labor, fairness, and accountability reflect value judgments that must be made transparently and inclusively. Ethical participation and democratic governance are not optional add-ons but core requirements for legitimate AI deployment in society.

Looking forward, we hope this work contributes to a broader movement toward integrative, interdisciplinary approaches to human–AI systems, bridging AI, HCI, ethics, social sciences, and domain expertise. Only by centering human values and real-world impact throughout the design and evaluation lifecycle can we ensure that AI augments, rather than undermines, human judgment, autonomy, and dignity.

ACKNOWLEDGMENTS

We thank all those who collaborated with us throughout the tutorials at AAMAS and NeurIPS. We greatly appreciate your support, helpful feedback, and ongoing dialogue, which continue to inform and inspire this work. We also thank Google for supporting tutorial panelist travel to NeurIPS. Part of this work has taken place in the Intelligent Robot Learning (IRL) Lab at the University of Alberta, which is supported in part by research grants from Alberta Innovates; Alberta Machine Intelligence Institute (Amii); a Canada CIFAR AI Chair, Amii; the Digital Research Alliance of Canada; Mitacs; and the National Science and Engineering Research Council (NSERC).

CONFLICT OF INTEREST STATEMENT

None of the authors have a conflict of interest to disclose.

ORCID

Danniell Hu https://orcid.org/0009-0000-1868-3004
Susanne Gaube https://orcid.org/0000-0002-1633-4772
Elizabeth Bondi-Kelly https://orcid.org/0000-0002-8459-8403

ENDNOTES

[1] Github Copilot: https://github.com/features/copilot
[2] Notion AI: https://www.notion.com/help/guides/category/ai

REFERENCES

Acosta-Navas, Diana. 2025. "On Foundations and Foundation Models: What Lessons Can AI and Philanthropy Learn from One Another?." In *The Routledge Handbook of Artificial Intelligence and Philanthropy*, edited by Giuseppe Ugazio and Milos Maricic, London: Routledge.

Adams, William C. 2015. "Conducting semi-structured interviews." *Handbook of practical program evaluation* 492–505.

A Adeoye-Olatunde, Omolola, and Nicole L Olenik. 2021. "Research and scholarly methods: Semi-structured interviews." *Journal of the American College of Clinical Pharmacy* 4(10): 1358–67.

Afzal, Afsoon, Deborah S. Katz, Claire Le Goues, and Christopher S. Timperley. 2020. "A study on the challenges of using robotics simulators for testing." *arXiv preprint arXiv:2004.07368*.

Ali, Sajid, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. 2023. "Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence." *Information fusion* 99: 101805.

Alkhatib, Ali, and Michael Bernstein. 2019. "Street–Level Algorithms: A Theory at the Gaps Between Policy and Decisions." In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, 1–13.

Alon-Barkat, Saar, and Madalina Busuioc. 2023. "Human–AI interactions in public sector decision making:"automation bias" and "selective adherence" to algorithmic advice." *Journal of Public Administration Research and Theory* 33(1): 153–69.

Altenried, Moritz. 2020. "The platform as factory: Crowdwork and the hidden labour behind artificial intelligence." *Capital & Class* 44(2): 145–58.

Amershi, Saleema, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. "Power to the people: The role of humans in interactive machine learning." *AI magazine* 35(4): 105–20.

Amershi, Saleema, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. "Guidelines for human-AI interaction." In *Proceedings of the 2019 Chi Conference on Human Factors in Computing Systems*, 1–13.

W Andrews, Robert, J Mason Lilly, Divya Srivastava, and Karen M Feigh. 2023. "The role of shared mental models in human-AI teams: A theoretical review." *Theoretical Issues in Ergonomics Science* 24(2): 129–75.

Mohammad Areeb, Qazi, Mohammad Nadeem, Shahab Saquib Sohail, Raza Imam, Faiyaz Doctor, Yassine Himeur, Amir Hussain, and Abbes Amira. 2023. "Filter bubbles in recommender systems: Fact or fallacy—A systematic review." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 13(6): e1512.

P Argyle, Lisa, Christopher A Bail, Ethan C Busby, Joshua R Gubler, Thomas Howe, Christopher Rytting, Taylor Sorensen, and David Wingate. 2023. "Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale." *Proceedings of the National Academy of Sciences* 120(41): e2311627120.

Asaro, Peter M. 2000. "Transforming society by transforming technology: the science and politics of participatory design." *Accounting, Management and Information Technologies* 10(4): 257–90.

Awad, Edmond, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. "The Moral Machine Experiment." *Nature* 563(7729): 59–64.

Balloli, Vaibhav, Sara Beery, and Elizabeth Bondi-Kelly. 2024. "Are they the same picture? adapting concept bottleneck models for human-AI collaboration in image retrieval." In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 7824–32.

Bandura, Albert. 2006. "Toward a psychology of human agency." *Perspectives on Psychological Science* 1(2): 164–80.

Bansal, Gagan, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. "Beyond accuracy: The role of mental models in human-AI team performance." In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol 7, 2–11.

Barocas, Solon, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and Machine Learning: Limitations and Opportunities*, Cambridge, MA: MIT Press.

Barocas, Solon, and Helen Nissenbaum. 2014. "Big Data's End Run Around Procedural Privacy Protections." *Communications of the ACM* 57(11): 31–33.

Barthe, Manon, Clarisse Bavoux, Francis Finot, Isabelle Mouche, Corina Cuceu-Petrenci, Andy Forreryd, Anna Chérouvrier Hansson, Henrik Johansson, Gregory F Lemkine, Jean-Paul Thénot, et al. 2021. "Safety testing of cosmetic products: overview of established methods and new approach methodologies (NAMs)." *Cosmetics* 8(2): 50.

Beede, Emma, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M Vardoulakis. 2020. "A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy." In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–12.

Behari, Nikhil, Edwin Zhang, Yunfan Zhao, Aparna Taneja, Dheeraj Nagaraj, and Milind Tambe. 2024. "A decision-language model (DLM) for dynamic restless multi-armed bandit tasks in public health." In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, 3964–4002.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?." In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, 610–23.

Bhatt, Umang, Javier Antorán, Yunfeng Zhang, Q Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, et al. 2021. "Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty." In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 401–13.

Bietti, Elettra 2020. "From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy." In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 210–19.

Bilal, Muhammad, and Hafiz MN Iqbal. 2019. "An insight into toxicity and human-health-related adverse consequences of cosmeceuticals—A review." *Science of the Total Environment* 670: 555–68.

Birhane, Abeba, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. "Power to the people? Opportunities and challenges for participatory AI." In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–8.

Birhane, Abeba, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. "The values encoded in machine learning research." In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, 173–84.

Bjögvinsson, Erling, Pelle Ehn, and Per-Anders Hillgren. 2012. "Design things and design thinking: Contemporary participatory design challenges." *Design Issues* 28(3): 101–16.

Bommasani, Rishi, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. "On the opportunities and risks of foundation models." *CoRR*, abs/2108.07258.

Bondi, Elizabeth, Raphael Koster, Hannah Sheahan, Martin Chadwick, Yoram Bachrach, Taylan Cemgil, Ulrich Paquet, and Krishnamurthy Dvijotham. 2022. "Role of human-AI interaction in selective prediction." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 5286–94.

Bondi, Elizabeth, Lily Xu, Diana Acosta-Navas, and Jackson A Killian. 2021. "Envisioning communities: a participatory approach towards AI for social good." In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 425–36.

Buçinca, Zana, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. "To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making." *Proceedings of the ACM on Human–Computer Interaction* 5(CSCW1): 1–21.

Buçinca, Zana, Siddharth Swaroop, Amanda E Paluch, Susan A Murphy, and Krzysztof Z Gajos. 2024. "Towards optimizing human-centric objectives in AI-assisted decision-making with offline reinforcement learning." *arXiv preprint arXiv:2403.05911*.

Casilli, Antonio A. 2025. *Waiting for Robots: The Hired Hands of Automation*, Chicago, IL: University of Chicago Press.

Casper, Stephen, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. 2023. "Open problems and fundamental limitations of reinforcement learning from human feedback." *Transactions on Machine Learning Research*.

PC Chanel, Caroline, Raphaëlle N Roy, Frédéric Dehais, and Nicolas Drougard. 2020. "Towards mixed-initiative human–robot interaction: Assessment of discriminative physiological and behavioral features for performance prediction." *Sensors* 20(1): 296.

Chang, Yupeng, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. "A survey on evaluation of large language models." *ACM Transactions on Intelligent Systems and Technology* 15(3): 1–45.

Chauhan, Kushal, Rishabh Tiwari, Jan Freyberg, Pradeep Shenoy, and Krishnamurthy Dvijotham. 2023. "Interactive concept bottleneck models." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, 5948–55.

Chen, Cong, Wei Qu, Si Su, Yukun Feng, and Tao Li. 2025. "A Comprehensive Review of LLM-based Content Moderation: Advancements, Challenges, and Future Directions." *Knowledge-Based Systems* 114689.

Chen, Valerie, Alan Zhu, Sebastian Zhao, Hussein Mozannar, David Sontag, and Ameet Talwalkar. 2025. "Need Help? Designing Proactive AI Assistants for Programming." In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–18.

Chen, You, Ellen Wright Clayton, Laurie Lovett Novak, Shilo Anders, and Bradley Malin. 2023. "Human-centered design to address biases in artificial intelligence." *Journal of Medical Internet Research* 25: e43251.

Chow, C. 1970. "On optimum recognition error and reject tradeoff." *IEEE Transactions on Information Theory* 16(1): 41–46.

Christiano, Paul F, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017a. "Deep Reinforcement Learning from Human Preferences." In *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, vol. 30, Curran Associates, Inc.

Christiano, Paul F, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017b. "Deep reinforcement learning from human preferences." *Advances in Neural Information Processing Systems* 31: 4302–10.

Coeckelbergh, Mark. 2023. "Democracy, Epistemic Agency, and AI: Political Epistemology in Times of Artificial Intelligence." *AI and Ethics* 3(4): 1341–50.

Costanza-Chock, Sasha. 2020. *Design Justice: Community-Led Practices to Build the Worlds We Need*, Cambridge, MA: MIT Press.

Crawford, Kate, and Jason Schultz. 2014a. "Big data and due process: Toward a framework to redress predictive privacy harms." *Boston College Law Review* 55: 93.

Crawford, Kate, and Jason Schultz. 2014b. "Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms." *Boston College Law Review* 55(1): 93–128.

Das, Nirjhar, Souradip Chakraborty, Aldo Pacchiano, and Sayak Ray Chowdhury. 2024. "Active preference optimization for sample efficient RLHF." In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* 96–112.

Dastin, Jeffrey. 2022. "Amazon scraps secret AI recruiting tool that showed bias against women." In *Ethics of Data and Analytics*, 296–99. Auerbach Publications.

de Jong, Sander, Ville Paananen, Benjamin Tag, and Niels van Berkel. 2025. "Cognitive forcing for better decision-making: reducing overreliance on AI systems through partial explanations." *Proceedings of the ACM on Human-Computer Interaction* 9(2): 1–30.

De Toni, Giovanni, Paolo Viappiani, Stefano Teso, Bruno Lepri, and Andrea Passerini. 2022. "Personalized algorithmic recourse with preference elicitation." *Transactions on Machine Learning Research*.

De Ville, Barry. 2013. "Decision trees." *Wiley Interdisciplinary Reviews: Computational Statistics* 5(6): 448–55.

J Dietvorst, Berkeley, Joseph P Simmons, and Cade Massey. 2015. "Algorithm aversion: people erroneously avoid algorithms after seeing them err." *Journal of Experimental Psychology: General* 144(1): 114.

Doshi-Velez, Finale, and Been Kim. 2018. "Considerations for evaluation and generalization in interpretable machine learning." *Explainable and Interpretable Models in Computer Vision and Machine Learning* 3–17.

Dvijotham, Krishnamurthy, Jim Winkens, Melih Barsbey, Sumedh Ghaisas, Robert Stanforth, Nick Pawlowski, Patricia Strachan, Zahra Ahmed, Shekoofeh Azizi, Yoram Bachrach, et al. 2023. "Enhancing the reliability and accuracy of AI-enabled diagnosis via complementarity-driven deferral to clinicians." *Nature Medicine* 29(7): 1814–20.

Dwivedi, Rudresh, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, et al. 2023. "Explainable AI (XAI): Core ideas, techniques, and solutions." *ACM Computing Surveys* 55(9): 1–33.

Endsley, Mica R. 2023. "Supporting Human-AI Teams: Transparency, explainability, and situation awareness." *Computers in Human Behavior* 140: 107574.

Eriksson, Maria, Erasmo Purificato, Arman Noroozian, Joao Vinagre, Guillaume Chaslot, Emilia Gomez, and David Fernandez-Llorca. 2025. "Can We Trust AI Benchmarks? An Interdisciplinary Review of Current Issues in AI Evaluation." In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 8(1): 850–64.

Mateo Espinosa, Zarlenga,, Katie Collins, Krishnamurthy Dvijotham, Adrian Weller, Zohreh Shams, and Mateja Jamnik. 2023. "Learning to receive help: Intervention-aware concept embedding models." *Advances in Neural Information Processing Systems* 36: 37849–75.

Ethayarajh, Kawin, and Dan Jurafsky. 2020. "Utility is in the eye of the user: A critique of NLP leaderboards." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* 4846–53.

U.S Food and Drug Administration (FDA). 2025. "Cosmetics." https://www.fda.gov/cosmetics.

U.S Food and Drug Administration (FDA). "Drug Quality Sampling and Testing Programs." https://www.fda.gov/drugs/science-and-research-drugs/drug-quality-sampling-and-testing-programs.

U.S Food and Drug Administration (FDA). 2025. "The FDA's Drug Review Process: Ensuring Drugs Are Safe and Effective." https://www.fda.gov/drugs/information-consumers-and-patients-drugs/fdas-drug-review-process-ensuring-drugs-are-safe-and-effective.

Felkner, Virginia K, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. 2023. "Winoqueer: A community-in-the-loop benchmark for anti-lgbtq+ bias in large language models." In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* 1: 9126–40.

Feng, Shangbin, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. "Don't Hallucinate, Abstain: Identifying LLM Knowledge Gaps via Multi-LLM Collaboration." In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* 1: 14664–90.

Floridi, Luciano. 2022. *The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities*, Oxford, UK: Oxford University Press.

Floridi, Luciano, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, et al. 2018. "AI4People—An ethical framework for a good AI society: opportunities, risks, principles, and recommendations." *Minds and Machines* 28: 689–707.

Fragiadakis, George, Christos Diou, George Kousiouris, and Mara Nikolaidou. 2024. "Evaluating human-ai collaboration: A review and methodological framework." *CoRR* abs/2407.19098.

Gaube, Susanne, Ekaterina Jussupow, Eesha Kokje, Jowaria Khan, Elizabeth Bondi-Kelly, Andreas Schicho, Felipe C Kitamura, Timo K. Koch, Timur Ezer, Jürgen Mottok, et al. 2024. "Underreliance Harms Human-AI Collaboration More Than Overreliance in Medical Imaging." OSF preprint.

Gaube, Susanne, Harini Suresh, Martina Raue, Alexander Merritt, Seth J Berkowitz, Eva Lermer, Joseph F Coughlin, John V Guttag, Errol Colak, and Marzyeh Ghassemi. 2021. "Do as AI say: susceptibility in deployment of clinical decision-aids." *NPJ digital Medicine* 4(1): 31.

Geifman, Yonatan, and Ran El-Yaniv. 2017. "Selective classification for deep neural networks." *Advances in Neural Information Processing Systems* 30: 4878–87.

Gomez, Catalina, Sue Min Cho, Shichang Ke, Chien-Ming Huang, and Mathias Unberath. 2025. "Human-AI collaboration is not very collaborative yet: A taxonomy of interaction patterns in AI-assisted decision making from a systematic review." *Frontiers in Computer Science* 6: 1521066.

Gonzalez-Cabello, Martin, Auyon Siddiq, Charles J. Corbett, and Catherine Hu. 2024. "Fairness in crowdwork: Making the human AI supply chain more humane." *Business Horizons* 68(5): 645–57.

Gray, Mary L., and Siddharth Suri. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*, New York: Houghton Mifflin Harcourt.

Gray, Richard, Daniel Bressington, David R Thompson, and Martin Jones. 2023. "Why pre-registration of research must be taken more seriously."

Green, Ben, and Yiling Chen. 2019. "The principles and limits of algorithm-in-the-loop decision making." *Proceedings of the ACM on Human-Computer Interaction* 3(CSCW): 1–24.

Bahalul Haque, AKM, AKM Najmul Islam, and Patrick Mikalef. 2023. "Explainable Artificial Intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research." *Technological Forecasting and Social Change* 186: 122120.

Hara, Kotaro, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P Bigham. 2018. "A data-driven analysis of workers' earnings on Amazon Mechanical Turk." In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14.

Hassija, Vikas, Vinay Chamola, Atmesh Mahapatra, Abhinandan Singal, Divyansh Goel, Kaizhu Huang, Simone Scardapane, Indro Spinelli, Mufti Mahmud, and Amir Hussain. 2024. "Interpreting black-box models: a review on explainable artificial intelligence." *Cognitive Computation* 16(1): 45–74.

Hawkins, Will, and Brent Mittelstadt. 2023. "The ethical ambiguity of AI data enrichment: Measuring gaps in research ethics norms and practices." In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 261–70.

A Hearst, Marti, J Allen, C Guinn, and Eric Horvitz. 1999. "Mixed-initiative interaction: Trends and controversies." *IEEE Intelligent Systems* 14(5): 14–23.

Hemmer, Patrick, Max Schemmer, Michael Vössing, and Niklas Kühl. 2021. "Human-AI complementarity in hybrid intelligence systems: A structured literature review." *PACIS* 78: 118.

Horvitz, Eric. 1999. "Principles of mixed-initiative user interfaces." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 159–66.

Howard, John. 2019. "Artificial intelligence: Implications for the future of work." *American Journal of Industrial Medicine* 62(11): 917–26.

Hu, Bin, Zhao Shi, Li Lu, Zhongchang Miao, Hao Wang, Zhen Zhou, Fandong Zhang, Rongpin Wang, Xiao Luo, Feng Xu, et al. 2024. "A deep-learning model for intracranial aneurysm detection on CT angiography images in China: a stepwise, multicentre, early-stage clinical validation study." *Lancet Digital Health* 6(4): e261–71.

Huang, Saffron, Divya, Siddarth, Liane, Lovitt, Thomas I., Liao, Esin, Durmus, Alex, Tamkin, and Deep, Ganguli. 2024. "Collective constitutional AI: Aligning a language model with public input." In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1395–417.

Huang, Tao. 2025. "Content moderation by llm: From accuracy to legitimacy." *Artificial Intelligence Review* 58(10): 1–32.

Huang, Yi-Ching, Yu-Ting Cheng, Lin-Lin Chen, and Jane Yung-jen Hsu. 2019. "Human-AI Co-learning for data-driven AI." *arXiv preprint arXiv:1910.12544*.

Jääskeläinen, Riitta. 2012. "Think-aloud protocol." In *Handbook of Translation Studies: Volume 1*, 371–73. John Benjamins Publishing Company.

Jabbour, Sarah, David Fouhey, Stephanie Shepard, Thomas S Valley, Ella A Kazerooni, Nikola Banovic, Jenna Wiens, and Michael W Sjoding. 2023. "Measuring the impact of AI in the diagnosis of hospitalized patients: a randomized clinical vignette survey study." *JAMA* 330(23): 2275–84.

Jacobs, Maia, Melanie F Pradier, Thomas H McCoy Jr, Roy H Perlis, Finale Doshi-Velez, and Krzysztof Z Gajos. 2021. "How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection." *Translational Psychiatry* 11(1): 108.

Jaques, Abby Everett. 2019. "Why the moral machine is a monster." *Robot Law* 2: 17–34.

WM Jaspers, Monique, Thiemo Steen, Cor Van Den Bos, and Maud Geenen. 2004. "The think aloud method: a guide to user interface design." *International Journal of Medical Informatics* 73(11-12): 781–95.

Jiang, Shu, and Ronald C Arkin. 2015. "Mixed-initiative human-robot interaction: definition, taxonomy, and survey." In *2015 IEEE International Conference on Systems, Man, and Cybernetics*, 954–61. IEEE.

Kahneman, Daniel. 2011. *Thinking, fast and slow*, macmillan.

Khan, Azhar Danish, and Mohammad Niyaz Alam. 2019. "Cosmetics and their associated adverse effects: A review." *Journal of Applied Pharmaceutical Sciences and Research* 1–6.

Kim, Eunji, Dahuin Jung, Sangha Park, Siwon Kim, and Sungroh Yoon. 2023. "Probabilistic concept bottleneck models." In *Proceedings of the 40th International Conference on Machine Learning* 16521–40.

Kim, Joo-Kyung, Guoyin Wang, Sungjin Lee, and Young-Bum Kim. 2021. "Deciding whether to ask clarifying questions in large-scale spoken language understanding." In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 869–76. IEEE.

SY Kim, Sunnie, Q Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. ""I'm Not Sure, But…": examining the impact of large language models' uncertainty expression on user reliance and trust." In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 822–35.

Rose Kirk, Hannah, Bertie Vidgen, Paul Röttger, and Scott A Hale. 2024. "The benefits, risks and bounds of personalizing the alignment of large language models to individuals." *Nature Machine Intelligence* 6(4): 383–92.

Klingbeil, Artur, Cassandra Grützner, and Philipp Schreck. 2024. "Trust and reliance on AI—An experimental study on the extent and costs of overreliance on AI." *Computers in Human Behavior* 160: 108352.

Bradley Knox, W, and Peter Stone. 2009. "Interactively shaping agents via human reinforcement: The TAMER framework." In *Proceedings of the fifth international conference on Knowledge capture*, 9–16.

Wei Koh, Pang, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. "Concept bottleneck models." In *International Conference on Machine Learning*, 5338–48. PMLR.

Kuhn, Lorenz, Yarin Gal, and Sebastian Farquhar. 2023. "Clam: Selective clarification for ambiguous questions with generative language models." *ICML Workshop on Deployment Challenges for Generative AI*.

Kumar, Shruti, Xiaoyu Chen, and Xiaomei Wang. 2024. "Mapping Human-Agent Co-Learning and Co-Adaptation: A Scoping Review." *Human Factors and Ergonomics Society*.

Küper, Alisa, Georg Lodde, Elisabeth Livingstone, Dirk Schadendorf, and Nicole Krämer. 2024. "Mitigating cognitive bias with clinical decision support systems: An experimental study." *Journal of Decision Systems* 33(3): 439–58.

Lakkaraju, Himabindu, Stephen H Bach, and Jure Leskovec. 2016. "Interpretable decision sets: A joint framework for description and prediction." In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1675–84.

Lazar, Jonathan, Jinjuan Heidi Feng, and Harry Hochheiser. 2017. *Research Methods in Human-Computer Interaction*, Morgan Kaufmann.

Hun Lee, Min, Daniel P Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 2020. "Co-design and evaluation of an intelligent decision support system for stroke rehabilitation assessment." *Proceedings of the ACM on Human-Computer Interaction* 4(CSCW2): 1–27.

Li, Haiwen, Soham De, Manon Revel, Andreas Haupt, Brad Miller, Keith Coleman, Jay Baxter, Martin Saveski, and Michiel A Bakker. 2025. "Scaling Human Judgment in Community Notes with LLMs." *arXiv preprint arXiv:2506.24118*.

Li, Jingshu, Yitian Yang, Renwen Zhang, and Yi-chieh Lee. 2024. "Overconfident and Unconfident AI Hinder Human-AI Collaboration." *arXiv preprint arXiv:2402.07632*.

Li, Zhuoyan, Hangxiao Zhu, Zhuoran Lu, Ziang Xiao, and Ming Yin. 2025. "From text to trust: empowering ai-assisted decision making with adaptive LLM-powered analysis." In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–18.

Liang, Percy, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. "Holistic evaluation of language models." *arXiv preprint arXiv:2211.09110*.

Lu, Jinwei, Yikuan Yan, Keman Huang, Ming Yin, and Fang Zhang. 2025. "Do we learn from each other: Understanding the human-ai co-learning process embedded in human-AI collaboration." *Group Decision and Negotiation* 34(2): 235–71.

Lykouris, Thodoris, and Wentao Weng. 2024. "Learning to defer in content moderation: The human-AI interplay." *arXiv preprint arXiv:2402.12237*.

MacGlashan, James, Mark K. Ho, Robert Loftin, Bei Peng, Guan Wang, David L. Roberts, Matthew E. Taylor, and Michael L. Littman. 2017, 06–11 Aug. "Interactive Learning from Policy-Dependent Human Feedback." In *Proceedings of the 34th International Conference on Machine Learning*, edited by Doina Precup and Yee Whye Teh, Vol 70 of *Proceedings of Machine Learning Research*, 2285–94. PMLR.

Madras, David, Toni Pitassi, and Richard Zemel. 2018. "Predict responsibly: improving fairness and accuracy by learning to defer." *Advances in Neural Information Processing Systems* 31: 6147–57.

Mao, Anqi, Christopher Mohri, Mehryar Mohri, and Yutao Zhong. 2023. "Two-stage learning to defer with multiple experts." *Advances in Neural Information Processing Systems* 36: 3578–606.

Mao, Ji-Ye, Karel Vredenburg, Paul W Smith, and Tom Carey. 2005. "The state of user-centered design practice." *Communications of the ACM* 48(3): 105–9.

Masri, Naser, Yousef Abu Sultan, Alaa N Akkila, Abdelbaset Almasri, Adel Ahmed, Ahmed Y Mahmoud, Ihab Zaqout, and Samy S Abu-Naser. 2019. "Survey of rule-based systems." *International Journal of Academic Information Systems Research (IJAISR)* 3(7): 1–23.

McDonald, Nora, Sarita Schoenebeck, and Andrea Forte. 2019. "Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice." *Proceedings of the ACM on Human-Computer Interaction* 3(CSCW): 1–23.

McDuff, Daniel, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, et al. 2025. "Towards accurate differential diagnosis with large language models." *Nature* 1–7.

Miller, Tim. 2019. "Explanation in artificial intelligence: Insights from the social sciences." *Artificial Intelligence* 267: 1–38.

G Mirnig, Alexander, Alexander Meschtscherjakov, Daniela Wurhofer, Thomas Meneweger, and Manfred Tscheligi. 2015. "A formal analysis of the ISO 9241-210 definition of user experience." In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, 437–50.

Montemayor, Carlos, Jodi Halpern, and Abrol Fairweather. 2022. "In principle obstacles for empathic AI: why we can—t replace human empathy in healthcare." *AI & Society* 37(4): 1353–59.

Morse, Lily, Mike Horia M Teodorescu, Yazeed Awwad, and Gerald C Kane. 2022. "Do the ends justify the means? Variation in the distributive and procedural fairness of machine learning algorithms." *Journal of Business Ethics* 181(4): 1083–95.

Mozannar, Hussein, and David Sontag. 2020. "Consistent estimators for learning to defer to an expert." In *International Conference on Machine Learning*, 7076–87. PMLR.

Mu, Fangwen, Lin Shi, Song Wang, Zhuohao Yu, Binquan Zhang, ChenXue Wang, Shichao Liu, and Qing Wang. 2024. "Clarifygpt: A framework for enhancing llm-based code generation via requirements clarification." *Proceedings of the ACM on Software Engineering* 1(FSE): 2332–54.

Mathur Natarajan, Saurabh, Sriraam Sahil Sidheekh, Wolfgang Stammer, and Kristian Kersting. 2025. "Human-in-the-loop or AI-in-the-loop? Automate or Collaborate?." In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol 39, 28594–600.

Navas, Diana Acosta. 2024. "In Moderation: Automation in the Digital Public Sphere." *Journal of Business Ethics*.

Newman, Andy. 2019, November. "I Found Work on an Amazon Website. I Made 97 Cents an Hour." *The New York Times*.

Nissenbaum, Helen. 2009. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*, Stanford, CA: Stanford University Press.

Norman, Don. 2013. *The design of everyday things: Revised and Expanded Edition*, Basic books.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. "Dissecting racial bias in an algorithm used to manage the health of populations." *Science* 366(6464): 447–53.

OpenAI. 2023, May. "Democratic inputs to AI.".

Ouyang, Long, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. "Training language models to follow instructions with human feedback." In *Advances in Neural Information Processing Systems*, edited by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Vol 35, 27730–44. Curran Associates, Inc.

Ovadya, Aviv. 2023. "Reimagining democracy for AI." *Journal of Democracy* 34(4): 162–70.

Overney, Cassandra, Daniel T Kessler, Suyash Pradeep Fulay, Mahmood Jasim, and Deb Roy. 2025. "Coalesce: An Accessible Mixed-Initiative System for Designing Community-Centric Questionnaires." In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, 366–89.

Palla, Konstantina, José Luis Redondo García, Claudia Hauff, Francesco Fabbri, Andreas Damianou, Henrik Lindström, Dan Taber, and Mounia Lalmas. 2025. "Policy-as-prompt: Rethinking content moderation in the age of large language models." In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 840–54.

Pargent, Florian, Timo K Koch, Anne-Kathrin Kleine, Eva Lermer, and Susanne Gaube. 2024. "A Tutorial on Tailored Simulation-Based Sample-Size Planning for Experimental Designs With Generalized Linear Mixed Models." *Advances in Methods and Practices in Psychological Science* 7(4): 25152459241287132.

Parkin, Simon. 2022. "The Trouble with Roblox: The Video Game Empire Built on Child Labour." *The Guardian*.

Passi, Samir, and Mihaela Vorvoreanu. 2022. "Overreliance on AI literature review." *Microsoft Research* 339: 340.

Perrigo, Billy. 2023. "OpenAI used Kenyan workers on less than $2 per hour: Exclusive."

Prabhudesai, Snehal, Leyao Yang, Sumit Asthana, Xun Huan, Q Vera Liao, and Nikola Banovic. 2023. "Understanding uncertainty: how lay decision-makers perceive and interpret uncertainty in human-AI decision making." In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 379–96.

Prolific. 2024, August. "What is Prolific and how does it work.", https://participant-help.prolific.com/en/article/dc132c.

Rahmani, Hossein A, Xi Wang, Mohammad Aliannejadi, Mohammadmehdi Naghiaei, and Emine Yilmaz. 2024. "Clarifying the path to user satisfaction: An investigation into clarification usefulness." *arXiv preprint arXiv:2402.01934*.

Reverberi, Carlo, Tommaso Rigon, Aldo Solari, Cesare Hassan, Paolo Cherubini, and Andrea Cherubini. 2022. "Experimental evidence of effective human–AI collaboration in medical decision-making." *Scientific Reports* 12(1): 14952.

Rinta-Kahila, Tapani, Esko Penttinen, Antti Salovaara, Wael Soliman, and Joona Ruissalo. 2023. "The vicious circles of skill erosion: A case study of cognitive automation." *Journal of the Association for Information Systems* 24(5): 1378–412.

Roberts, Sarah T. 2019. *Behind the Screen: Content Moderation in the Shadows of Social Media*, New Haven, CT: Yale University Press.

Rong, Yao, Tobias Leemann, Thai-Trang Nguyen, Lisa Fiedler, Peizhu Qian, Vaibhav Unhelkar, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci. 2023. "Towards human-centered explainable ai: A survey of user studies for model explanations." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46(4): 2104–22.

M Salih, Ahmed, Zahra Raisi-Estabragh, Ilaria Boscolo Galazzo, Petia Radeva, Steffen E Petersen, Karim Lekadir, and Gloria Menegaz. 2025. "A perspective on explainable artificial intelligence methods: SHAP and LIME." *Advanced Intelligent Systems* 7(1): 2400304.

Schemmer, Max, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. "Appropriate reliance on AI advice: Conceptualization and the effect of explanations." In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 410–22.

Schemmer, Max, Niklas Kühl, Carina Benz, and Gerhard Satzger. 2022. "On the influence of explainable AI on automation bias." *arXiv preprint arXiv:2204.08859*.

Schröder, Sarah, Thekla Morgenroth, Ulrike Kuhl, Valerie Vaquet, and Benjamin Paaßen. 2025. "Large language models do not simulate human psychology." *arXiv preprint arXiv:2508.06950*.

D Selbst, Andrew, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. "Fairness and abstraction in sociotechnical systems." In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59–68.

Shang, Chenming, Shiji Zhou, Hengyuan Zhang, Xinzhe Ni, Yujiu Yang, and Yuwang Wang. 2024. "Incremental residual concept bottleneck models." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11030–40.

Sharma, Ashish, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2023. "Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support." *Nature Machine Intelligence* 5(1): 46–57.

Shavit, Yonadav, Sandhini Agarwal, Miles Brundage, Steven Adler, Cullen O'Keefe, Rosie Campbell, Teddy Lee, Pamela Mishkin, Tyna Eloundou, Alan Hickey, et al. 2023. "Practices for governing agentic AI systems." *Research Paper, OpenAI*.

Sheth, Ivaxi, and Samira Ebrahimi Kahou. 2023. "Auxiliary losses for learning generalizable concept-based models." *Advances in Neural Information Processing Systems* 36: 26966–90.

E Shipe, Maren, Stephen A Deppen, Farhood Farjah, and Eric L Grogan. 2019. "Developing prediction models for clinical use using logistic regression: an overview." *Journal of Thoracic Disease* 11(4): S574.

Shneiderman, Ben. 2022. *Human-Centered AI*, Oxford University Press.

Singh, Shivalika, Yiyang Nan, Alex Wang, Daniel D'Souza, Sayash Kapoor, Ahmet Üstün, Sanmi Koyejo, Yuntian Deng, Shayne Longpre, Noah Smith, et al. 2025. "The leaderboard illusion." *arXiv preprint arXiv:2504.20879*.

Sloane, Mona, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. 2022. "Participation Is Not a Design Fix for Machine Learning." In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '22)*, 1–6.

Spinuzzi, Clay. 2005. "The methodology of participatory design." *Technical Communication* 52(2): 163–74.

Steen, Marc. 2012. "Human-centered design as a fragile encounter." *Design Issues* 28(1): 72–80.

Stray, Jonathan. 2023. "The AI learns to lie to please you: preventing biased feedback loops in machine-assisted intelligence analysis." *Analytics* 2(2): 350–58.

Sundararajan, Mukund, and Amir Najmi. 2020. "The many Shapley values for model explanation." In *International Conference on Machine Learning*, 9269–78. PMLR.

Suthaharan, Shan. 2016. "Support vector machine." In *Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning*, 207–35. Springer.

Swaroop, Siddharth, Zana Buçinca, Krzysztof Z Gajos, and Finale Doshi-Velez. 2025. "Personalising AI assistance based on overreliance rate in AI-assisted decision making." In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, 1107–22.

Szepesvári, Csaba. 2022. *Algorithms for Reinforcement Learning*, Springer nature.

Teso, Stefano, and Kristian Kersting. 2019. "Explanatory interactive machine learning." In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 239–45.

Henry Tessler, Michael, Michiel A Bakker, Daniel Jarrett, Hannah Sheahan, Martin J Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tantum Collins, David C Parkes, et al. 2024. "AI can help humans find common ground in democratic deliberation." *Science* 386(6719): eadq2852.

Testoni, Alberto, and Raquel Fernández. 2024. "Asking the right question at the right time: Human and model uncertainty guidance to ask clarification questions." *arXiv preprint arXiv:2402.06509*.

The Collective Intelligence Project. 2024. "Introducing the Collective Intelligence Project."

Toxtli, Carlos, Siddharth Suri, and Saiph Savage. 2021. "Quantifying the invisible labor in crowd work." *Proceedings of the ACM on Human-Computer Interaction* 5(CSCW2): 1–26.

Turk, Amazon Mechanical. "Amazon Mechanical Turk." https://www.mturk.com/.

Vasconcelos, Helena, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S Bernstein, and Ranjay Krishna. 2023. "Explanations can reduce overreliance on ai systems during decision-making." *Proceedings of the ACM on Human-Computer Interaction* 7(CSCW1): 1–38.

Verma, Rajeev, Daniel Barrejón, and Eric Nalisnick. 2023. "Learning to defer to multiple experts: Consistent surrogate losses, confidence calibration, and conformal ensembles." In *International Conference on Artificial Intelligence and Statistics*, 11415–34. PMLR.

Véliz, Carissa. 2020. *Privacy Is Power: Why and How You Should Take Back Control of Your Data*, London: Bantam Press.

Vössing, Michael, Niklas Kühl, Matteo Lind, and Gerhard Satzger. 2022. "Designing transparency for effective human-AI collaboration." *Information Systems Frontiers* 24(3): 877–95.

Walliman, Nicholas. 2021. *Research methods: The basics*, Routledge.

Wang, Xinyu, Sai Koneru, Pranav Narayanan Venkit, Brett Frischmann, and Sarah Rajtmajer. 2025. "The unappreciated role of intent in algorithmic moderation of abusive content on social media."

Wang, Ziming, Changwu Huang, and Xin Yao. 2024. "Procedural fairness in machine learning." *arXiv preprint arXiv:2404.01877*.

Weidinger, Laura, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. "Ethical and Social Risks of Harm from Language Models." *arXiv preprint arXiv:2112.04359*.

Wiener, Yair, and Ran El-Yaniv. 2013. "Theoretical foundations of selective prediction."

Wilder, Bryan, Eric Horvitz, and Ece Kamar. 2021. "Learning to complement humans." In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, 1526–33.

Williams, Adrienne, Milagros Miceli, and Timnit Gebru. 2022. "The Exploited Labor Behind Artificial Intelligence." *Noema*.

Wu, Xingjiao, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. 2022. "A survey of human-in-the-loop for machine learning." *Future Generation Computer Systems* 135: 364–81.

Xu, Zhengtao, Tianqi Song, and Yi-Chieh Lee. 2025. "Confronting verbalized uncertainty: Understanding how LLM's verbalized uncertainty influences users in AI-assisted decision-making." *International Journal of Human-Computer Studies* 197: 103455.

Yan, Biwei, Kun Li, Minghui Xu, Yueyan Dong, Yue Zhang, Zhaochun Ren, and Xiuzhen Cheng. 2024. "On protecting the data privacy of large language models (llms): A survey." *arXiv preprint arXiv:2403.05156*.

Yuksekgonul, Mert, Maggie Wang, and James Zou. 2022. "Post-hoc concept bottleneck models." *arXiv preprint arXiv:2205.15480*.

Jin Yun, Tae, Jin Wook Choi, Miran Han, Woo Sang Jung, Seung Hong Choi, Roh-Eul Yoo, and In Pyeong Hwang. 2023. "Deep learning based automatic detection algorithm for acute intracranial haemorrhage: a pivotal randomized clinical trial." *npj Digital Medicine* 6(1): 61.

Zhang, Michael JQ, and Eunsol Choi. 2023. "Clarify when necessary: Resolving ambiguity through interaction with lms." *arXiv preprint arXiv:2311.09469*.

Zhang, Yunfeng, Q Vera Liao, and Rachel KE Bellamy. 2020. "Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making." In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 295–305.

Zhou, Zhi-Hua,. 2011. "Cost-sensitive learning." In *International Conference on Modeling Decisions for Artificial Intelligence*, 17–18. Springer.

Ziegler, Daniel M., Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. "Fine-Tuning Language Models from Human Preferences."

Zou, Jie, Mohammad Aliannejadi, Evangelos Kanoulas, Maria Soledad Pera, and Yiqun Liu. 2023. "Users meet clarifying questions: Toward a better understanding of user interactions for search clarification." *ACM Transactions on Information Systems* 41(1): 1–25.

Zou, Jie, Evangelos Kanoulas, and Yiqun Liu. 2020. "An empirical study on clarifying question-based systems." In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2361–64.

How to cite this article: Hu, D., D. A. Navas, S. Gaube, H. Mozannar, M. E. Taylor, K. Dvijotham, and E. Bondi-Kelly. 2025. "Human at the Center: A Framework for Human-Driven AI Development." *AI Magazine* 46: e70043. https://doi.org/10.1002/aaai.70043

AUTHOR BIOGRAPHIES

Danniell Hu is a Ph.D. student in Computer Science at the University of Michigan, where she conducts research in the Realize Lab under the supervision of Professor Elizabeth Bondi-Kelly. She received her B.S.E. in Computer Science from the University of Michigan. Her research focuses on Artificial Intelligence for Social Impact, with particular interest in public health and women's health applications, stakeholder-aligned system design, and the intersection of AI and software engineering.

Diana Acosta-Navas is an assistant professor of management specializing in business ethics at Loyola University Chicago. She was previously an Embedded EthiCS fellow at Stanford University, based in the Center for Ethics in Society (EiS) and the Institute for Human-Centered Artificial Intelligence (HAI). Diana holds a PhD in philosophy from Harvard University. Diana studied philosophy at the Universidad de Los Andes in Bogotá, Colombia, before completing a master's program at the Universidad Nacional de Colombia. Her research examines the role of digital platforms in fostering healthy public debate and the ethical implications of emerging technologies on the digital public sphere. She also explores human rights in post-conflict scenarios, focusing on the moral responsibilities of businesses to support peace-building processes.

Dr. Susanne Gaube is a Lecturer (Assistant Professor) in Human Factors in Healthcare at UCL's Global Business School for Health (GBSH). Her research focuses on (a) identifying opportunities in healthcare that could benefit from the introduction of digital technologies, (b) understanding the support needs of healthcare providers and patients and optimal ways to implement digital health technologies, and (c) improving the interaction between users and technology with a particular focus on Radiology, Infection Prevention and Control, and Mental Health. She also investigates how AI-enabled clinical decision support systems influence decision-making and proposes strategies to enhance human-AI collaboration. Susanne holds an MSc in Industrial/Organizational and Business Psychology from UCL and a PhD in Psychology from the University of Regensburg (Germany).

Hussein Mozannar is a Senior Researcher at Microsoft Research AI Frontiers working on AI agents. He obtained his PhD from MIT in Social & Engineering Systems and Statistics in 2024 and his undergraduate degree in computer engineering from the American University of Beirut in 2019. His research focuses on augmenting humans with AI to help them complete tasks more efficiently. Specifically, he focuses on building AI models and agents that complement human expertise and designing interaction schemes to facilitate human-AI interaction. The main applications of his research have been software development, computer and browser use, and healthcare.

Matthew E. Taylor (Matt) is a professor in computer science at the University of Alberta, a Fellow-in-Residence at the Alberta Machine Intelligence Institute, and remains an adjunct professor at Washington State University. He received his doctorate from the University of Texas at Austin in the summer of 2008, supervised by Peter Stone. Previously, he completed a two-year postdoctoral research position at the University of Southern California with Milind Tambe and spent 2.5 years as an assistant professor at Lafayette College. He was also an assistant professor at Washington State University where he held the Allred Distinguished Professorship in Artificial Intelligence. In 2017, he temporarily left academia to help start an artificial intelligence lab in Edmonton, Alberta, with Borealis AI, the artificial intelligence research lab for the Royal Bank of Canada.

Krishnamurthy (Dj) Dvijotham is a senior staff research scientist at Google DeepMind where he leads efforts on the development of secure and trustworthy AI agents. He previously founded the AI security research team at ServiceNow Research and co-founded the robust and verified AI team at DeepMind. His past research has received best paper awards at many leading AI conferences, including most recently at ICML and CVPR 2024. His research led to the framework used for AI security testing at ServiceNow and has been deployed in several Google products, including the Android Play Store, YouTube and Gemini. He received his PhD in Computer Science and Engineering from the University of Washington.

Elizabeth Bondi-Kelly is an assistant professor of Electrical Engineering and Computer Science at the University of Michigan. Previously, she was a Postdoctoral fellow at MIT through the CSAIL METEOR fellowship. She earned her PhD in Computer Science from Harvard University, where she was advised by professor Milind Tambe. Her research interests lie broadly in the area of artificial intelligence (AI) for social impact, particularly spanning the fields of multi-agent systems and machine learning for conservation and public health.