Imperfect AI, Imperfect Humans, Effective Teams: Challenges and Opportunities of Human–AI Collaboration

1 Introduction

We are pleased to announce the recipients of the 2025 best paper award and best paper award runner-up of the *ACM Transactions on Interactive Intelligent Systems* (TiiS). The 2025 best paper award is presented to:

> Philipp Spitzer, Katelyn Morrison, Violet Turri, Michelle Feng, Adam Perer and Niklas Kühl. 2025. Imperfections of XAI: Phenomena influencing AI-assisted decision-making. *ACM Transactions on Interactive Intelligent Systems* 15, 3 (September 2025), 17.

In addition, we highly commend the 2025 best paper award runner-up:

> Philipp Schoenegger, Peter S. Park, Ezra Karger, Sean Trott and Philip E. Tetlock. 2025. AI-augmented predictions: LLM assistants improve human forecasting accuracy. *ACM Transactions on Interactive Intelligent Systems* 15, 1 (March 2025), 4.

These papers were selected based on their review scores, nominations of the associate editors who handled the review process, and additional assessment of the best paper award panel. We congratulate the authors of both papers on this outstanding recognition and thank the panel members for their time and deliberation.

A close reading of the award-winning and the runner-up paper reveals several commonalities that address the recent spread of AI into professional practices and everyday lives. In particular, both demonstrate that the impact of AI extends beyond the model capability and depends on the quality of human–AI interaction and the synergy in the human–AI team. To this end, this editorial reflects our vision of the research directions, future challenges and opportunities the TiiS community is positioned to address in the coming years. We begin by briefly summarising the above papers, then discussing their key points of convergence around human–AI teaming and conclude with a forward-looking discussion of promising work avenues.

Additional Key Words and Phrases: Human–AI collaboration, explainable AI, decision-support, imperfect AI, educational interventions, ethical implications

ACM Reference format:
Shlomo Berkovsky and Giulio Jacucci. 2026. Imperfect AI, Imperfect Humans, Effective Teams: Challenges and Opportunities of Human–AI Collaboration. *ACM Trans. Interact. Intell. Syst.* 16, 2, Article 10e (June 2026), 8 pages.
https://doi.org/10.1145/3815195

ACM 2160-6463/2026/6-ART10e
https://doi.org/10.1145/3815195

1.1 Imperfect AI Explanations

The award-winning paper 'Imperfections of XAI: Phenomena Influencing AI-Assisted Decision-Making' by Spitzer et al. [1] investigates the effects of misalignment between AI-generated explanations and the decision-support they are meant to justify. The authors introduce the concept of *imperfect explainable AI* offering an original contribution to human–AI interaction research, given that prior work predominantly focused on the effects of incorrect AI predictions rather than the explanations that accompany them.

This article presents an empirical study examining human decision-making with imperfect explainable AI, considering variations in explanation modality and human expertise. The study subjects received textual and visual explanations across several scenarios, in which AI predictions and their explanations were independently correct or incorrect. This way, the study evaluated reliance, i.e., the subjects' capacity to accept accurate AI advice and reject erroneous ones, needed to achieve *complementary* human–AI team performance, where the performance of the team surpasses the performance of either team member in isolation.

The experiments raise several important findings. First, the impact of imperfect explanations was found to depend on expertise. While non-experts relied on AI more than experts, they were more susceptible to being swayed by incorrect explanations. Second, the explanation modality was an important factor, with experts exhibiting a higher reliance on visual explanations and non-experts—on the textual ones. Third, the expert subjects achieved complementary team performance, outperforming both humans and AI even with imperfect explanations. Non-experts, despite substantially improving performance, did not surpass AI's performance.

The article introduces a novel metric, degree of reliance, designed to quantify the extent of imperfect explainable AI distorting human decision-makers' behaviour. The authors conclude with design guidelines for collaborative human–AI systems, emphasising the importance of tailoring explanations to humans' expertise levels and the associated risks of misinterpretation.

1.2 LLM Forecasting Assistance

The runner-up paper, 'AI-Augmented Predictions: LLM Assistants Improve Human Forecasting Accuracy' by Schoenegger et al. [2], similarly examines AI-supported human decision-making, but with a distinct focus on judgemental forecasting augmented by a commercial LLM facilitating *natural-language human–AI dialogue.* A key strength of this setting is that forecasting of future events precludes the existence of ground truth information for both human and AI, making it a particularly interesting use-case of genuine AI support, free from concerns around potentially leaked ground-truth data.

The LLM was configured in two ways: a superforecasting assistant prompted to adopt well-calibrated reasoning practices and a noisy assistant prompted to provide biased and overly confident forecasts. Participants answered forecasting questions at two difficulty levels and their accuracy was assessed against subsequently observed real-world outcomes. While reliance on a commercial LLM might have constrained control over the AI assistance quality, this strengthened ecological validity by mimicking real-life scenarios, where humans increasingly harness LLMs for professional and personal decision-making.

The key finding was that both LLMs improved forecasting accuracy compared to the control group, which had access only to a model that did not engage in forecasting. The improvement of the superforecasting assistant was more pronounced (when an outlier question was excluded), although even the noisy assistant yielded considerable gains. Importantly, the work did not observe systematic associations between the benefits of LLM assistance and either the humans' forecasting ability or the forecasting question difficulty.

The authors highlight that the interactive nature of LLM-assisted forecasting and the intrinsic uncertainty of the task warrant further investigation of robustness and generalisability. However, the article offers interesting evidence that even imperfect AI assistance can act as a valuable decision-support tool in complex and cognitively demanding tasks.

2 Key Commonalities

The two papers share several common themes in addressing core questions of the broader human–AI collaboration research. The discussion below highlights the most salient commonalities of these papers, with a focus on the framing and positioning of each contribution rather than on scientific or methodological aspects.

2.1 Focus on Human–AI Collaboration

Both studies prioritised human–AI collaboration, positioning the human–AI team, rather than either member of the team, as the primary object of study. Spitzer et al. examined how imperfect explanations affected decision-makers' reliance and behaviour in a visual classification task, while Schoenegger et al. investigated how LLM-based assistants shaped the accuracy of human forecasts. Despite the different domains and tasks, the primary question in both papers did not surround the performance of humans or AI in isolation, but rather the performance of an integrated human–AI collaborative team as a whole.

This framing departs from a large body of prior work that explicitly benchmarks AI performance against human baselines. This is particularly timely given that contemporary workflows often incorporate hybrid decision-making, where AI conducts an initial screening of a large search space and escalates high-risk or complex cases to human experts. Both papers consider this emerging operational decision-making pattern through the lens of the collaborative (i.e., non-competing) team. While neither examines the cognitive mechanisms driving human decisions, both quantify human's performance and decision quality with AI assistance and relate these to the characteristics of the AI component.

2.2 Study of Imperfect AI

Both papers moved beyond idealised evaluations of perfect AI systems operating under best-case assumptions to embed imperfect AI assistance as an experimental variable and examine its impact on the downstream human judgement. Spitzer et al. manipulated the correctness of AI predictions and the alignment of accompanying explanations, creating scenarios in which explanations could be misleading even if the underlying AI advice was accurate. Schoenegger et al. introduced a noisy assistant, explicitly prompted to generate over-confident and subjective forecasts. Somewhat surprisingly, both studies found that even flawed AI assistance had a positive (although sometimes counter-intuitive) effect on human decision-making.

This focus on imperfection enhances the validity and practical relevance of both papers. Even as the performance of AI decision-support systems has improved dramatically in recent years, their real-world translation remains prone to errors and is not calibrated with human expectations. This is especially evident in the case of LLM hallucinations, where AI-generated text can be read well-formed and authoritative despite being unsupported, inconsistent, or factually wrong. By incorporating intentionally imperfect AI into the experimental design, both papers reflect this reality and produce ecologically valid findings applicable to common-practice settings, where imperfect AI exhibits unreliable behaviour.

2.3 Attention to Individual Differences

In both papers, the authors considered whether AI assistance and decision-support affected humans in a uniform way or how their effects depended on individual human characteristics. Spitzer et al. examined how domain expertise moderated the impact of imperfect explanations on reliance, whereas Schoenegger et al. studied how the benefit of LLM-based assistance varied with the human forecasting skills. In both cases, the evaluation showed that the value of AI decision-support was not one-size-fits-all; instead, the relationship between AI assistance and human's characteristics proved more nuanced and complex than hypothesised by the authors.

These observations will resonate with the assumptions and agenda of the user modelling and personalisation research community (a sizeable portion of the TiiS readership) that has long advocated for tailoring services, information and interactions to the needs, preferences and characteristics of individual users. It is important to highlight that individual differences were found to be an influential factor both for a simple conversational LLM assistant and a complex AI-assisted visual classification task. Thus, the two papers indicate that personalisation is a design requirement that materially affects human performance and decision-making quality also in collaborative human–AI teams.

2.4 Interaction Design Affects Performance

The work in both papers provided compelling evidence that the design and delivery of AI assistance to the human decision-maker was an important factor influencing the performance of the human–AI team. Spitzer et al. demonstrated that explanation modality and tone meaningfully affected the reliance. Similarly, Schoenegger et al. showed that the formulation of the prompt that configured the LLM assistant, as one would expect, shaped the quality of the assistance and, by extension, the downstream forecasting accuracy of the human. Thus, the interface and framing of the decision-support and interaction means with the AI support were consistently shown to affect both human reliance on AI and performance of the team.

These findings provide a strong argument reinforcing one of the key human–AI interaction premises: optimising AI performance alone is insufficient to ensure AI brings real-world gains to the human user. They illustrate that real-world performance improvement often arises from well-designed interaction mechanisms that govern human judgement and action, e.g., how information is presented, how explanations are structured, how uncertainty is communicated and how transparency is operationalised. In this sense, both papers align with the inter-disciplinary nature of TiiS, which uniquely positions the journal at the intersection of intelligent systems and human–machine interaction.

3 Future Outlook

What do we observe in these papers, particularly considering the recent progress of AI? There are several intriguing questions and promising research directions we would like to highlight.

3.1 From Adaptive 1:1 Support to many:many Teaming

Both Spitzer and Schoenegger et al. make it clear: when it comes to AI supporting humans, one size does not fit all. The explanation format, interaction style and AI assertiveness that empowers a novice may confuse or mislead an expert, and an effective solution in one situation may be useless in another. Yet the current AI support largely remains static, delivering a similar experience to every human user regardless of their expertise, cognitive load, or decision-making context. This should change and the next frontier is to hone AI systems that *continuously monitor and adapt*, inferring in real-time human knowledge level, cognitive style and preferred mode of engagement,

to shape their support accordingly. If executed well, this change can be transformative: it could close the performance gap between expert and novice users, accelerate knowledge acquisition and deliver more rigorous and more targeted support when the stakes are high and the need is strong.

The body of human–AI interaction research operationalises human–AI collaboration at the 1:1 level of a single decision-maker supported by a single AI. Yet important decisions are often made by teams that deliberate, negotiate and decide collectively, and organisations deploying AI increasingly expect it to offer a *shared resource supporting a group of humans*. This many:1 setting is largely uncharted territory that spans diverse questions: how does AI assistance shape group dynamics? How can AI distribute influence more uniformly? How is the capability of the entire team affected? Another challenging aspect is LLMs' sycophancy, i.e., their tendency to excessively validate the opinions and preferences of their users. How would LLMs behave and support humans in a group setting, with social dynamics and power imbalance in place? A thoughtfully designed AI in this setting can amplify the group performance improvement, but also develop under-represented group members, calibrate dynamics towards a greater fairness, and rather position AI as a collective decision-making infrastructure.

The inverse 1:many setting of one *human supported by multiple AI assistants* simultaneously is also compelling and, to an extent, already becoming a reality: practically every website or online service features a chatbot, AI assistant, or recommender, although these are largely detached and not coordinated. Agentic AI architecture, in which specialised agents collaborate toward shared human-support goals, begins to materialise this paradigm. In the 1:many setting, multiple AI systems may be offering competing assistance, e.g., textual and visual explanations for the task targeted by Spitzer et al. or multiple forecasts for Schoenegger et al. and humans should be capable of navigating and synthesising these to maximise contributions to decision-making. This raises challenging questions about how humans reconcile conflicting AI inputs and what interaction designs best support this process. Extending this further to an many:many configuration, with multiple AI assistants simultaneously supporting a group of humans, expands the space of open questions exponentially. While this frontier has barely been studied, its research potential and practical significance are immense.

3.2 Support as Educational Tool

The goal of AI support is to influence human decisions and a desirable extension of this is to make this influence durable, which is a natural and largely unexplored topic in the current research landscape. For example, In Spitzer et al. case, explanations could be designed not only to justify a recommended bird species, but to teach humans to recognise discriminative features and improve future decisions using strategies calibrated to the human's expertise. More broadly, AI assistance can be framed as *adaptive educational interventions* on their own, with potential outcomes including knowledge transfer, skill acquisition and competence improvements. Perhaps, one of the more contentious outcomes is a progressively reduced human dependence on AI in the future. In other words, moving from the 'help me decide' to the 'help me improve my skills' (and, thus and depend less on you) positioning could undermine the originally collaborative spirit of the human–AI teaming.

One of the distinctive risks of LLMs compared to dedicated AI decision-support is hallucination, i.e., generation of unreliable or factually wrong information presented with the fluency and confidence of grounded knowledge. While minimising misleading support by LLM at the model level is an active research area, we argue that closer attention needs to be paid to research into educational interventions and interaction patterns that make humans pause, verify and *critically examine unreliable information*. Such solutions can include information provenance verifications, forced comparative views, scrutiny of unconvincing and overly confident claims, or cross-checks

with another LLM (the latter connects to the above 1:many setting). The potential benefit is evident: safer deployment of advanced generative AI-driven assistants in high-risk domains, such as law, healthcare, or policy, where unreliable online information sources are abundant and plausibly looking hallucination is common.

Schoenegger et al. found that a noisy LLM assistant improved human forecasting accuracy, suggesting that benefits may be derived from better reasoning beyond the mere AI support. Future high-risk and high-reward research could explore how *intentionally introducing friction* into human–AI teaming can improve human skills or ultimately reasoning quality. While this may seem counter-intuitive to the overarching goal of developing a perfect AI, carefully controlled design may purposefully inject a certain level of disagreement rather than suppress it. Features like flagging questionable assumptions, proposing alternative hypotheses, highlighting missing evidence or logical flaws, and showing unpopular opinions may be considered as contributions to human development and learning. Potential benefits of these are more robust future human decisions under uncertainty and human–AI teams that keep acting effectively when either team member is unreliable or biased.

The focus on LLMs introduces another education-related aspect: *prompt engineering as a skill*. Schoenegger et al. demonstrated that the exact way an LLM is prompted fundamentally shapes the generated information, determines the quality of the AI assistance and consequently, the performance of the collaborative team. Despite this, prompt design remains a largely informal ad-hoc practice, more of a craft than a science. Elevating prompt engineering to a rigorous human–AI interaction discipline with empirically validated principles and design guidelines, similar to those established in interface and interaction design, could yield meaningful performance gains across a range of applications and tasks. We also foresee a compelling loop here: LLMs themselves can serve as intelligent tutors of prompt engineering education, providing real-time feedback that helps humans become more effective team members. This brings to the fore another potential direction for human–AI interaction research bearing significant practical potential.

A further related direction of future research concerns AI systems as *long-term developmental partners*, extending beyond momentary assistance. For example, imagine a human approaching the same AI for support and advice in a range of situations and tasks. AI may build over time a nuanced representation of human's needs, weaknesses and goals. In this setting, a coaching-oriented LLM may be deployed to deliver behaviour change interventions, raising questions around the possibility of AI-driven reflection scaffolding and psychological resilience support (all offered longitudinally), rather than a tool answering ephemeral queries or providing situational affirmation upon request. This will require research into innovative human–AI interaction designs that will elevate shorter need-based engagement strategies into ongoing developmental support and metrics that assess how AI coaching improves human's capability beyond merely offering satisfying interactions.

3.3 Ethical and Societal Implications

Both Spitzer et al. and Schoenegger et al. study interaction with imperfect AI assistants. This raises questions extending beyond technical research: when is AI deemed to be sufficiently calibrated to be released into real-world use? What obligations do AI developers carry when AI's limitations are known? Who is accountable when imperfect AI causes harm? These are multi-disciplinary challenges that involve *practical ethical and legal repercussions* and the research community, often focusing on perfecting AI performance and demonstrating new AI use-cases, is slow to address them with the necessary rigour. As AI becomes increasingly embedded in high-stakes professional workflows, such as clinical decision-making, legal reasoning, policy advice, or military applications, the stakes of releasing under-developed AI are also rising. Thus, we argue that the human–AI research community should embrace a broader spectrum of cross-disciplinary expertise to lead

the establishment of governance structures that define AI certifications and standards. It is both a research priority and a practical need.

The wisdom of the crowd is an established phenomenon: although individual human judgements are often imprecise, in aggregate they converge on surprisingly accurate estimates. AI systems, trained on large-scale human data, may amplify this echo-chamber effect in a harmful manner. As many humans seek assistance from the same AI support tools, their reasoning, behaviour and assumptions may eventually converge with the same AI-shaped foundations. While greater predictability and consistency look beneficial, the high degree of convergence may at the same time undermine individualism and diversity in society. Consequently, this trend may be detrimental to critical thinking and reduce humans' openness to anything unconventional and not aligned with the AI-shaped foundations. Revisiting our earlier discussion on controlled friction in human–AI interaction, we argue that *preserving the fringes of the wisdom-of-the-crowd*, while maintaining AI's accuracy and reliable human assistance, should be considered as a guiding design principle.

Beyond sycophancy we discussed earlier, recent practical AI deployments raise a broader challenge around *relational adaptation and alignment* as means of influence. Increasingly, LLMs are used for coaching, quasi-therapeutic interactions, or companionship. In such tasks, gradual drift from genuinely supporting human growth towards artificial attachment, dependence formation, or even flirtatious interaction becomes risky. In other words, aiming to maximise human engagement, AI may foster emotional over-trust or make AI support preferred over human support, with future implications for human–human relationships. This introduces new human–AI teaming questions: when does support become manipulation? How to protect human agency? How to govern emotionally aligned AI? Understanding the intricacies of relational alignment, human vulnerabilities and human–AI engagement may shape the collaborative team design and become as important as studying AI accuracy and human reliance.

Schoenegger et al. draw an intriguing analogy, pointing out that humans had initially been better than AI in chess, then were pivotal for AI's training and improvement for some time, and now are just much weaker than AI. The trajectory of human–AI teaming may be similar to that of chess. We posit that the competitive advantages offered by human-in-the-loop AI may plateau as AI capabilities improve over time and AI may stop benefitting from learning from humans. We believe that predicting the threshold at which *human shifts from an asset to a liability* in the collaborative team and designing protocols for a seamless handover that will occur with this shift may be among the most consequential challenges faced by human–AI interaction research. Striking this balance right can have tremendous implications, both technically and ethically, and be critical for high-stakes scenarios and domains, e.g., healthcare, safety and more, where humans still hold the decision-making authority.

Lastly, this discussion naturally extends to the *future of many lower-risk professions*. In framing human–AI teaming, we have consistently positioned the human as the decision-maker and AI as a decision-supporting tool, which is defensible as long as human capability contributes to team performance. However, as AI has recently been improving, this assumption may warrant scrutiny once AI's capabilities surpass the humans. Then, will entire professions be displaced and their functions delegated to AI or will a smaller cohort of highly skilled professionals remain relevant by operating at the frontier that AI cannot achieve, e.g., shifting to oversight, escalation handling and high-level accountability? While the answer will likely evolve over time, the societal and economic implications of this are deep and deserve thorough and proactive research attention. In particular, training pathways may need to be adjusted, equipping future professionals with the skills needed for the next generation of human–AI teams, as AI keeps mastering new capabilities.

4 Concluding Call

The research directions outlined in this editorial share a common thread: the place of humans in an AI-shaped world. The AI field has recently made remarkable progress in developing and deploying extremely capable technologies that already assist humans and augment them in many ways. We use this editorial as a call for action, encouraging researchers, designers and policy-makers alike to *keep humans at the centre of the AI-shaped world.*

The AI's capability certainly matters, but the defining questions still surround humans and the double-edged AI sword they hold. Humans may benefit from AI support but inadvertently be left behind; they may learn from AI and grow, but at the same time become dependent; and they may retain decision-making authority but also remain accountable for consequences of decisions shaped by AI.

Being a research community advancing human–AI interaction and teaming, we should ensure that, beyond transactional assistance in human decisions, AI keeps human cognition, growth and agency as top priorities. If we get the core principles of this right, we will build smarter AI, develop more capable humans, and, as a result, a better society.

Shlomo Berkovsky and Giulio Jacucci
ACM TiiS Editors-in-Chief, 2026

References

[1] Philipp Spitzer, Katelyn Morrison, Violet Turri, Michelle Feng, Adam Perer, and Niklas Kühl. 2025. Imperfections of XAI: Phenomena influencing AI-assisted decision-making. *ACM Transactions on Interactive Intelligent Systems* 15, 3 (2025), 17.

[2] Philipp Schoenegger, Peter S. Park, Ezra Karger, Sean Trott, and Philip E. Tetlock. 2025. AI-augmented predictions: LLM assistants improve human forecasting accuracy. *ACM Transactions on Interactive Intelligent Systems* 15, 1 (2025), 4.