How to Explain in XAI? - Investigating Explanation Protocols in Decision Support Systems

Caterina Fregosi[1]

[1]University of Milano-Bicocca

Abstract

The objective of this proposal is to bridge the gap between Deep Learning (DL) and System Dynamics (SD) by developing an **interpretable neural system dynamics** framework. While DL excels at learning complex models and making accurate predictions, it lacks interpretability and causal reliability. Traditional SD approaches, on the other hand, provide transparency and causal insights but are limited in scalability and require extensive domain knowledge. To overcome these limitations, this project introduces a Neural System Dynamics pipeline, integrating Concept-Based Interpretability, Mechanistic Interpretability, and Causal Machine Learning. This framework combines the predictive power of DL with the interpretability of traditional SD models, resulting in both causal reliability and scalability. The efficacy of the proposed pipeline will be validated through real-world applications of the EU-funded AutoMoTIF project, which is focused on autonomous multimodal transportation systems. The long-term goal is to collect actionable insights that support the integration of explainability and safety in autonomous systems.

1. Context and Motivation

Imagine a clinician confronted with a complex diagnostic dilemma. She consults a newly deployed AI decision support system in her department, which confidently proposes a diagnosis accompanied by a seemingly persuasive explanation. After a brief hesitation, she accepts the recommendation. The decision feels rational: the system has been trained on thousands of cases and appears confident. But was her judgment genuinely improved by the AI, or merely influenced by it? Did the explanation reinforce her reasoning—or subtly circumvent it, diminishing her sense of agency? And what happens if she uses this system daily? Will she continue to feel fully responsible for the diagnoses she approves, or will that sense of responsibility become diluted—shared with the system, or displaced onto it?

1.1. Technical and Ethical Risks in Human-AI Collaboration

This quiet moment of reliance illustrates a growing concern in the integration of AI into high-stakes decision-making: not only how to improve immediate performance, but also how to preserve human agency and responsibility. Clinical Decision Support Systems (CDSSs) that incorporate AI models are increasingly capable of supporting clinicians with diagnostic and treatment decisions, offering the promise of improved efficiency and accuracy. However, these systems also introduce a complex set of risks. At a functional level, there is the danger of *inappropriate reliance*—users may over-trust AI and accept its outputs uncritically, or they may reject beneficial AI support. A central challenge is to foster *appropriate reliance*, the ability to accept AI-generated recommendations when they are correct, and to reject them when they are misleading [1]. Beyond performance-related concerns lie more structural risks that pertain to the quality and ethics of human-AI collaboration itself. As AI systems become increasingly authoritative in their presentation, they may gradually erode clinicians' sense of responsibility, reflective reasoning, and agency [2]. Over time, this can contribute to the erosion of diagnostic capabilities (i.e., deskilling) [3] and the inhibition to develop these skills. Clinicians

Late-breaking work, Demos and Doctoral Consortium, colocated with The 3rd World Conference on eXplainable Artificial Intelligence: July 09–11, 2025, Istanbul, Turkey

caterina.fregosi@unimib.it (C. Fregosi)
0009-0004-7626-8131 (C. Fregosi)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

may lose the motivation or even the capacity to critically engage with complex cases, particularly when AI systems offer confident recommendations with limited scope for deliberation. These risks are particularly salient in medical contexts, where accountability and diagnostic reasoning are integral to both professional identity and clinical accuracy.

1.2. Explainable AI: Promise and Limitations

In response to these unintended consequences of human-AI interaction, Explainable AI (XAI) has emerged as a prominent research direction, aiming to make certain aspects of a system more understandable and interpretable [4].

In this context, we adopt a specific view of explanations as "the (characterizing) output of a XAI system, that is, the output of any computational system aimed at making AI-generated advice more understandable, appropriable and exploitable by their intended users" [5]. In other words, XAI explanations are not necessarily "explanations" in the everyday sense, but rather clarifying outputs about the system's primary recommendation. Nevertheless, growing empirical evidence shows that explanations alone do not necessarily improve decision quality [6]; in fact, explanations can sometimes mislead users in various ways. For a phenomenon known as *white-box paradox* , users may develop overtrust in AI systems simply because the system's internal logic appears intelligible or well-justified, regardless of the actual correctness of the advice. In such cases, the more we see (or we think we do) inside the system, the more we trust it, and persuasive but flawed explanations can mask underlying errors, leading users to accept incorrect recommendations. In our own research, we identified a complementary effect (the *XAI Halo Effect*) in which misleading explanations degraded decision quality even when the AI recommendations themselves were accurate [7]. These effects challenge the assumption that increased explainability inherently leads to improved outcomes in hybrid human-AI collaboration.

Moreover, current systems interaction protocols often assume that users will interpret advices and explanatory cues as intended by designers. In practice, however, users bring their own mental models, domain expertise, and cognitive habits. Thus, what matters is not only the content of the explanation, but also the *interaction protocol*, that is, the way system outputs are framed, structured, and integrated into decision-making workflows. In this light, explanations must be not only intelligible but also behaviorally effective, that is, capable of promoting appropriate reliance, mitigating cognitive biases, and supporting collaborative decision-making. One commonly used explanatory strategy is the display of confidence scores, which serve as indicators of the system's certainty in its own recommendations. However, our recent study, currently under review, demonstrates that such cues are far from neutral [8]. When well-calibrated, confidence scores significantly improved decision accuracy and reduced both automation bias and conservatism bias. In contrast, miscalibrated confidence—for instance, expressing high certainty in incorrect outputs—led users to follow flawed advice or disregard accurate recommendations, ultimately impairing performance. These findings underscore a broader point: effective explainability is not solely a matter of transparency, but of alignment—between what is presented, how it is interpreted, and the behaviors it elicits. Designing explanations that are user-tailored and that support collaborative reasoning, rather than passive acceptance, is essential to minimizing unintended consequences.

1.3. Rethinking Human-AI Interaction: Beyond the Oracular Model

These concerns call for a shift from explanation as output to explanation as *interaction design*, able to foster critical reflection and to preserve user agency and responsibility.

Nevertheless, most CDSSs are still designed according to an "Oracular" model, in which a single recommendation is presented with the intent to persuade rather than promote dialogue [9]. This approach can lead users to accept AI-generated recommendations passively. As a response, researchers have begun to explore alternative explanation paradigms [10, 9, 11], designed to prompt users to actively engage with the decision space, weighing competing arguments rather than passively accepting AI outputs. Building on this line of research, we propose a protocol, *Judicial*, which draws inspiration from deliberative practices in judicial settings. Rather than offering a definitive recommendation, the system presents two contrastive explanations, each supporting a different decision outcome. The aim is to re-engage the user's discriminative capacities by requiring an active choice between alternatives. This approach motivates my broader research objective: not merely to improve diagnostic accuracy, but to investigate which forms of explanation—and how their content, structure, and framing influence users' engagement with AI systems. Specifically, I aim to examine how different explanatory strategies affect critical reflection, perceived agency, and responsibility, and how these effects may vary depending on individual expertise, cognitive style, and decision-making preferences.

2. Background and Related Work

There is a growing consensus that Human-AI collaboration should not be conceived in terms of a single optimal design, but rather as a socio-technical intervention whose impact depends on how the system is embedded into users' workflows and cognitive routines. Our recent work, currently under review, on protocol-driven design in hybrid intelligence, emphasizes that different interaction protocols influence not only decision accuracy, but also patterns of user reliance, the degree of dependence on AI, and the long-term risk of professional deskilling [12]. This shift reflects a broader move away from one-size-fits-all systems that deliver authoritative recommendations, toward adaptive and reflective interaction models tailored to users' roles, expertise, and decision-making needs [13]. In this view, supporting agency and responsibility in AI-supported decision-making requires not just technical robustness or explainability, but also careful attention to the mode and timing of human-AI interaction.

2.1. Agency

Human *agency*, understood as the subjective experience of authorship and responsibility over one's decisions, involves more than the capacity to choose—it entails experiencing those choices as authentically one's own. In decision-making contexts supported by AI, this sense of experiential ownership can be undermined. Specifically, Explainable AI systems that provide a single recommendation and a persuasive explanation in an authoritative or overly confident manner may lead users to gradually disengage from the decision-making process, thereby reducing opportunities for reflective thinking [14].

Over time, this erosion of agency may give rise to a phenomenon akin to *deresponsibilization*, in which users come to view themselves less as accountable decision-makers and more as passive executors of algorithmic outputs [15]. These risks are particularly pronounced in high-stakes domains such as clinical diagnostics, where active user engagement with AI outputs is critical for ensuring both safety and the quality of decision-making. Addressing these challenges requires the development of interaction models that go beyond mere explanation and instead support the user's role as a responsible agent within the decision-making loop.

2.2. Frictional AI and the Judicial Protocol

While decision support systems have demonstrated clear benefits in improving diagnostic accuracy, their widespread adoption has raised significant concerns regarding uncritical reliance [16] and the gradual erosion of sense of agency, responsibility, and diagnostic skills [3]. These risks are particularly evident in systems that follow an oracular interaction model, in which a single, confident recommendation is presented [9]. In response, recent research has emphasized the importance of interaction protocols that actively support users' cognitive engagement and promote reflective, accountable decision-making. Such protocols include mechanisms that prompt users to interrogate both their own reasoning and the system's recommendations. Building on the concept of *cognitive friction*, some design strategies deliberately introduce "positive friction" to foster human reflection and reduce uncritical acceptance [17].

To this end, our research group (MUDI Lab, University of Milano-Bicocca) introduced the term *Frictional AI* as an umbrella concept for a variety of methods aimed at encouraging reflection in human-AI decision making processes by intentionally introducing cognitive friction.
One such method is the Judicial protocol, inspired by the reasoning practices of judges. Rather than

providing a single recommendation, this protocol presents multiple, contrasting diagnostic alternatives, each supported by persuasive, yet potentially fallible, justifications. These justifications are not necessarily factual; instead, they are constructed to argue persuasively in favor of one classification over another. For this reason, we refer to them as *perorative explanations*. This interaction model aligns with approaches such as Evaluative AI [9], the Reflection Machine [10], and the concept of Dissenting Explanations [11]. These approaches support a model of human-AI collaboration in which users remain central, accountable agents. A design shift increasingly recognized in both HCI and XAI communities as critical for preserving human agency, professional judgment, and long-term competence. These theoretical perspectives collectively inform the present research project, which operationalizes the principles of Frictional AI through a novel interaction protocol. The following sections present the Judicial paradigm and its empirical evaluation in a clinical decision-making context.

3. Research question/s, Hypothesis and Objectives

To assess the practical implications of Judicial approach, this study investigates the impact of three different explanation conditions in a diagnostic task: (1) a **Traditional** protocol, offering a single recommendation with explanation; (2) an **Alternative Judicial** protocol, where a single AI system provides two contrasting diagnoses with justifications; and (3) an **Antagonist Judicial** protocol, in which two separate systems each advocate for a different diagnosis with distinct explanatory arguments. By stratifying the sample of medical students and experienced clinicians based on their level of clinical expertise, the study will evaluate the impact of the three explanation protocols on users' diagnostic accuracy, confidence, sense of agency, sense of responsibility, perceived influence, and perceived utility of the system. By exploring these variables, the study seeks to contribute to the design of DSS that better support critical decision-making in clinical practice.

However, before evaluating these dimensions, it is essential to determine whether the cognitive demands imposed by the Judicial protocol might detrimentally affect diagnostic performance. Indeed, in high-stakes contexts such as healthcare, accuracy remains a priority. For this reason, our exploratory study [18], detailed in the Preliminary Results section, was designed to first assess whether the Judicial protocol could preserve or even enhance diagnostic accuracy, rather than impair it due to increased cognitive load, yielding encouraging preliminary findings.

3.1. Research questions

We aim to address the following research questions:

1. **RQ1:** Are there significant differences in users' perceived sense of agency and responsibility between the Traditional DSS and the Judicial explanation protocols?
2. **RQ2:** Are there significant differences in diagnostic accuracy and user confidence between the Traditional DSS and the Judicial explanation format?
3. **RQ3:** Are there significant differences in diagnostic accuracy and user confidence between the Antagonist and Alternative Judicial conditions?
4. **RQ4:** Are there significant differences in the perceived influence and perceived utility of the AI system between the Traditional DSS and the Judicial explanation formats?
5. **RQ5:** Are there significant differences in the perceived utility, influence, sense of agency and sense of responsibility between the Antagonist and Alternative Judicial protocols?

To further explore potential moderating effects, the sample will be stratified by level of clinical experience, allowing us to assess whether these variables vary as a function of users' expertise.

3.2. Hypotheses

We formulate the following hypotheses:

- H1: Participants are expected to report a stronger sense of agency and responsibility in response to Judicial protocol cases compared to Traditional DSS cases. The Judicial formats are specifically designed to preserve agency by promoting critical thinking and encouraging users to adjudicate between diagnostic alternatives. In contrast, the Traditional protocol positions the user as a passive recipient of advice, which may diminish the perceived sense of agency.
- H2: We hypothesize that Traditional protocol cases will be associated with higher diagnostic accuracy, due to the presentation of a single, high-confidence recommendation that minimizes ambiguity and cognitive load. However, Judicial protocol cases (Alternative or Antagonist)—by encouraging deliberation and critical evaluation through contrastive explanations—may foster higher post-decision confidence, particularly when users reach a conclusion after resolving conflicting information.
- H3: A significant difference in diagnostic accuracy and user confidence is expected between the Antagonist and Alternative Judicial protocols. Receiving two opposing recommendations from distinct systems (Antagonist) may be perceived as more epistemically legitimate, as it mirrors inter-expert disagreement. In contrast, encountering contradictory outputs from a single system (Alternative) may elicit skepticism regarding the system's internal coherence. These differences in perceived epistemic authority may influence both user confidence and diagnostic decisions, depending on how participants interpret the source and implications of the disagreement.
- H4: Perceived influence and utility of the AI system on users' decisions are expected to differ significantly between the Traditional and Judicial protocols. The Traditional format presents a single recommendation with a persuasive explanation, which may lead to a stronger subjective sense of system influence—users may feel they are merely following the AI's advice. In contrast, Judicial protocols do not provide ad advice, potentially diminishing the perception of being directly "guided" by the system. However, the Judicial formats may be perceived as more useful by users who value deliberation and autonomy, as they offer richer informational input and promote active diagnostic reasoning. Thus, while influence may be lower, perceived utility could be equal or even higher, particularly among experienced users.
- H5: Perceived utility, influence, agency, and responsibility are expected to differ between the Alternative and Antagonist Judicial protocols. When conflicting explanations are presented by a single system (Alternative), users may question the system's credibility, potentially reducing perceived utility and trust. In contrast, when divergent recommendations are attributed to distinct systems (Antagonist), the disagreement may be interpreted as reflective of epistemic complexity rather than internal inconsistency, potentially enhancing user engagement and the perceived value of the AI as a decision support tool.

4. Methods

To test these hypotheses, a between- and within-subjects experimental design will be implemented (see Figure 1.

Participants Participants—comprising clinicians and medical students from the University of Milano Statale—will be randomly assigned to one of two experimental conditions: **Alternative Judicial** or **Antagonist Judicial**. We aim to recruit a minimum of 50 participants, all of whom will participate on a voluntary basis. Responses to the online survey will be collected anonymously.

Procedure During the experimental session, each participant will evaluate a total of 10 clinical cases, presented in a fixed, predefined order. These cases comprise 5 supported by the *Traditional DSS*, which provides a single diagnostic recommendation accompanied by an explanatory rationale, and 5 supported by either the *Alternative Judicial* or *Antagonist Judicial* format, depending on the participant's group assignment. In the *Alternative Judicial* group, a single DSS presents two alternative diagnoses, each with a corresponding explanation. In the *Antagonist Judicial* group, two distinct DSSs each advocate for a different diagnosis, accompanied by their respective justifications.

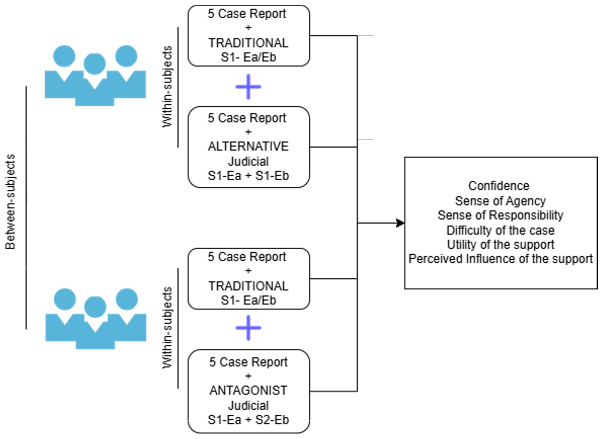

Figure 1: Experimental design.

After each AI-assisted decision, participants will be asked to rate their confidence in the decision, their perception of the case's complexity, and the utility of the support system This experimental design enables both within-subject comparisons (Traditional vs. Judicial cases) and between-subject comparisons (Alternative vs. Antagonist Judicial conditions). All AI recommendations used in the study are simulated to ensure consistency across participants and to maintain full control over the diagnostic content. Clinical cases are adapted from the *New England Journal of Medicine*[1] and include symptomatology, medical history, and laboratory results, together forming realistic complex diagnostic scenarios. Each case includes multiple differential diagnoses, with the correct diagnosis explicitly identified in the clinical discussion sections of the source articles. For the Judicial conditions, the correct and alternative diagnoses were selected with the assistance of a clinician among the study's authors, who identified the most plausible alternative diagnosis for each case based on clinical reasoning. In all conditions, AI explanations are designed to be persuasive and are grounded in the clinical features of the respective case. To simulate a realistic yet imperfect decision support system, the AI accuracy in Traditional cases was fixed at 80%, with incorrect recommendations introduced in a controlled and systematic manner. Participants interact with the AI via an online interface developed using LimeSurvey[2]. For each clinical case, they first view the AI output and then choose between two diagnostic options, followed by a self-assessed confidence rating on a 4-point ordinal scale. After each condition (i.e., after completing the 5 Traditional cases and again after the 5 Judicial cases), participants also evaluate the sense of agency (AGC), responsibility (RESP), and the influence of the AI system on their decision (INF) (see Figure 1). These measures aim to assess participants' sense of agency under each explanation condition. After the experimental session, participants will complete two standardized psychometric instruments to assess stable individual differences relevant to decision-making. Specifically, they will complete the Italian version of the Short Big Five Inventory (BFI) to measure personality traits according to the five-factor model, and the Italian adaptation of the Decision Styles Scale (DSS), which focuses on the rational decision style and the intuitive decision style. These measures will enable the exploration of potential interactions between dispositional traits and responses to different AI explanation protocols.

Sense of Agency: Participants' sense of agency is assessed through a set of post-decision items measuring perceived influence, ownership over decisions, and sense of responsibility in each decision-making task. These items are inspired by constructs from the Sense of Agency literature and were adapted to fit the clinical decision-making context.

Accuracy: Diagnostic accuracy is coded dichotomously as correct (1) or incorrect (0), based on the reference diagnosis reported in the source medical literature.

Confidence, Utility: These variables are measured using self-reported 4-point ordinal scales, designed to minimize central tendency bias.

[1] https://www.nejm.org/

[2] https://www.limesurvey.org/it

5. Preliminary Results

We conducted an exploratory user study in a controlled setting [18]. Sixteen medical professionals (8 spine surgeons and 8 musculoskeletal radiologists) were recruited to participate. The task involved assessing 18 vertebral X-ray images for the presence of fractures. Participants first examined each image independently and recorded their diagnosis and confidence. In a second phase, they were exposed to the Judicial AI protocol, which presented two activation maps offering contrasting, perorative explanations supporting either a positive (fracture) or negative (no fracture) classification. After reviewing these maps, participants could revise their diagnosis and re-rate their confidence. This human-first protocol ensured that participants' initial judgments were uninfluenced by AI outputs. The non-inferiority test showed that overall diagnostic accuracy with Judicial AI support was not inferior to unaided decision-making (Z = 3.94, $p < .001$), and even improved significantly among experienced clinicians (Glass's Delta = 0.99, 95% CI [0.50, 1.47], p = .045). Regarding confidence, the Wilcoxon signed-rank test confirmed non-inferiority overall ($p < .001$), and confidence gains were more pronounced in complex cases (Cliff's Delta = 0.296, p = .034). Less experienced users exhibited a modest improvement in confidence but no significant accuracy benefit. The perceived utility of Judicial support was rated positively by 57% of participants, significantly above chance (Binomial test, p = .015), with no significant differences across expertise levels (Mann-Whitney, p = .32). These preliminary findings support the feasibility of the Judicial AI in clinical diagnostic tasks. The protocol maintained or improved diagnostic performance overall, with particularly strong effects for experienced clinicians and in more complex cases, contexts where interpretive support is arguably most needed. The system also enhanced users' confidence, suggesting its value as a reflective support tool. The modest impact on less experienced users may be due to the increased cognitive load associated with interpreting two explanations. This suggests the need for adaptive support tailored on users' interaction needs.

6. Next step /Future Works

The next phase of this research involves the implementation of the experimental protocol described above. This study will assess whether contrastive explanation formats, as instantiated in the Judicial paradigm, can enhance not only diagnostic accuracy but also users' sense of agency and responsibility. A central objective is to determine whether the deliberative framing promoted by Judicial protocols meaningfully reinforces users' experiential ownership of the decision-making process. In addition, the study will examine whether the effects of different explanation protocols are moderated by clinical experience, as suggested by preliminary findings. This will help determine whether more experienced clinicians are better equipped to engage with contrastive information, while less experienced users may benefit from more structured or guided support. Such insights are critical for designing adaptive decision support systems that tailor explanatory strategies to users' expertise, cognitive style, and interaction needs. A further line of research concerns the role of framing within the Judicial paradigm itself. By comparing the Antagonist and Alternative versions, where conflicting explanations are attributed to two distinct systems or to a single source, respectively, we aim to clarify whether the origin of disagreement influences perceptions of trust, responsibility, and reliance. If framing effects are substantial, future research will need to investigate how users interpret epistemic authority under different configurations, and how these interpretations shape collaborative dynamics. Taken together, these research directions contribute to a broader goal: the development of human–AI interaction protocols that move beyond accuracy and transparency to actively support accountable, and user-centered collaboration.

Declaration on Generative AI

The author has not employed any Generative AI tools.

References

[1] J. D. Lee, K. A. See, Trust in automation: Designing for appropriate reliance, Human factors 46 (2004) 50–80.

[2] J. W. Moore, What is the sense of agency and why does it matter?, Frontiers in psychology 7 (2016) 1272.

[3] Y. S. J. Aquino, W. A. Rogers, A. Braunack-Mayer, H. Frazer, K. T. Win, N. Houssami, C. Degeling, C. Semsarian, S. M. Carter, Utopia versus dystopia: professional perspectives on the impact of healthcare artificial intelligence on clinical roles and skills, International Journal of Medical Informatics 169 (2023) 104903.

[4] L. Longo, M. Brcic, F. Cabitza, J. Choi, R. Confalonieri, J. Del Ser, R. Guidotti, Y. Hayashi, F. Herrera, A. Holzinger, et al., Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions, Information Fusion (2024) 102301.

[5] F. Cabitza, A. Campagner, G. Malgieri, C. Natali, D. Schneeberger, K. Stoeger, A. Holzinger, Quod erat demonstrandum?-towards a typology of the concept of explanation for the design of explainable ai, Expert systems with Applications 213 (2023) 118888.

[6] G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, D. Weld, Does the whole exceed its parts? the effect of ai explanations on complementary team performance, in: Proceedings of the 2021 CHI conference on human factors in computing systems, 2021, pp. 1–16.

[7] F. Cabitza, C. Fregosi, A. Campagner, C. Natali, Explanations considered harmful: The impact of misleading explanations on accuracy in hybrid human-ai decision making, in: World Conference on Explainable Artificial Intelligence, Springer, 2024, pp. 255–269.

[8] C. Fregosi, L. Vicente, A. Campagner, F. Cabitza, Too sure for our own good: A user study on ai confidence and human reliance, Submitted to the 41st Conference on Uncertainty in Artificial Intelligence (UAI), 2025. Manuscript under review.

[9] T. Miller, Explainable ai is dead, long live explainable ai! hypothesis-driven decision support using evaluative ai, in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, 2023, pp. 333–342.

[10] P. Haselager, H. Schraffenberger, S. Thill, S. Fischer, P. Lanillos, S. Van De Groes, M. Van Hooff, Reflection machines: Supporting effective human oversight over medical decision support systems, Cambridge Quarterly of Healthcare Ethics 33 (2024) 380–389.

[11] O. Reingold, J. H. Shen, A. Talati, Dissenting explanations: Leveraging disagreement to reduce model overreliance, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, 2024, pp. 21537–21544.

[12] A. Campagner, C. Fregosi, F. Cabitza, Five degrees of separation: Investigating the unexpected potential of displaced human-ai collaboration protocols for apter ai support, 2025. Submitted to the 28th CSCW. Manuscript under review.

[13] Q. Yang, A. Steinfeld, C. Rosé, J. Zimmerman, Re-examining whether, why, and how human-ai interaction is uniquely difficult to design, in: Proceedings of the 2020 chi conference on human factors in computing systems, 2020, pp. 1–13.

[14] R. Legaspi, W. Xu, T. Konishi, S. Wada, N. Kobayashi, Y. Naruse, Y. Ishikawa, The sense of agency in human–ai interactions, Knowledge-Based Systems 286 (2024) 111298.

[15] C. Sureau, Medical deresponsibilization, Journal of assisted reproduction and genetics 12 (1995).

[16] Z. Buçinca, M. B. Malaya, K. Z. Gajos, 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 (2021) 1–21.

[17] Z. Chen, R. Schmidt, Exploring a behavioral model of "positive friction" in human-ai interaction, in: International Conference on Human-Computer Interaction, Springer, 2024, pp. 3–22.

[18] F. Cabitza, L. Famiglini, C. Fregosi, S. Pe, E. Parimbelli, G. A. La Maida, E. Gallazzi, From oracular to judicial: Enhancing clinical decision making through contrasting explanations and a novel interaction protocol, in: Proceedings of the 30th International Conference on Intelligent User Interfaces, 2025, pp. 745–754.