# Judicial Protocols in Diagnostic AI: Contrastive Explanations to Preserve Human Agency

Caterina Fregosi[1,*], Chiara Natali[1,2] and Federico Cabitza[1,3]

[1]University of Milano-Bicocca, Viale Sarca 336, Milano, 20126, Italy
[2]University of Applied Sciences and Arts of Southern Switzerland, Via La Santa 1, Lugano, 6900, Switzerland
[3]IRCCS Ospedale Galeazzi-Sant'Ambrogio, Via Cristina Belgioioso 173, Milano, 20157, Italy

**Abstract**

What if AI in medicine didn't recommend clinicians what to do, but showed them what to consider? Artificial intelligence–based decision support systems (DSS) are increasingly integrated into clinical workflows to enhance diagnostic accuracy and reduce cognitive effort. However, empirical research shows that such systems may unintentionally foster automation bias, reduce critical engagement, and erode clinicians' sense of agency. Within this context, recent work in Human–AI Interaction has emphasized the role of design features that stimulate deliberation, such as cognitive friction and contrastive reasoning. Yet, little is known about how such interaction protocols affect users' perceived decision agency and diagnostic performance. Here we report a study in progress that evaluates the impact of contrastive explanation formats on clinicians' sense of agency, confidence, and diagnostic accuracy. We compare a Traditional DSS that provides a single recommendation with two "Judicial" protocols—Alternative and Antagonist—that introduce competing diagnoses and justifications to promote evaluative reasoning. The study adopts a mixed within- and between-subjects design involving medical students and clinicians, and includes a novel multidimensional HCI Sense of Agency scale tailored to diagnostic tasks. Data collection is currently ongoing.

**Keywords**

Human-AI Interaction, Clinical Decision Support System, Frictional AI

## 1. Introduction

Artificial intelligence-based decision support systems (DSS) are increasingly deployed in high-stakes domains such as medicine, where diagnostic decisions carry significant consequences. These systems promise to improve accuracy while reducing clinicians' cognitive workload. Yet, their integration into clinical practice also raises important concerns. Empirical studies show that AI-assisted decisions can foster detrimental effects such as overreliance, automation bias, and a diminished sense of responsibility and critical reasoning [1, 2, 3, 4]. Lee et al. [5] conducted extensive surveys among knowledge workers, revealing that the use of LLMs notably diminishes perceived cognitive effort involved in critical thinking tasks. Crucially, their findings illustrate that confidence in the tool's capability correlates inversely with independent critical engagement. Workers displaying high confidence in AI outputs reported lower cognitive effort but also reduced independent verification efforts. This observation is critical, as it suggests that extensive reliance on AI systems, although efficient, risks creating cognitive complacency, undermining the cultivation and exercise of critical judgement. In clinical contexts, where accountability and deliberation are central to both professional identity and patient safety, these risks are particularly pressing.

At a functional level, the core challenge is to promote *appropriate reliance*: the ability to accept AI-generated advice when it is correct and to reject it when it is not [6, 7, 8, 9]. Traditional DSSs typically adopt an "Oracular" format [10], offering a single, authoritative recommendation, often accompanied by a persuasive explanation. Although this design may improve short-term performance,

*HHAI-WS 2025: Workshops at the Fourth International Conference on Hybrid Human-Artificial Intelligence (HHAI), June 9–13, 2025, Pisa, Italy*
*Corresponding author.
caterina.fregosi@unimib.it (C. Fregosi); chiara.natali@unimib.it (C. Natali); federico.cabitza@unimib.it (F. Cabitza)
0009-0004-7626-8131 (C. Fregosi); 0000-0002-5171-5239 (C. Natali); 0000-0002-4065-3415 (F. Cabitza)

it can also encourage passive acceptance, reduce deliberation, and gradually erode clinicians' sense of responsibility [11, 12] and even professional skill, in a phenomenon termed *AI-induced deskilling* [13].

Explainable AI (XAI) has emerged as a response to these unintended effects, aiming to make system outputs more interpretable and actionable [14]. However, growing empirical evidence suggests that explanations alone do not guarantee improved outcomes; on the contrary, they may mislead users and degrade decision quality [15, 16]. These findings indicate that effective explainability requires more than transparency: it requires the alignment between what is presented, how it is interpreted, and the behaviors it elicits.

From this perspective, explanations should be understood not merely as outputs, but as components of an *interaction protocol* [17]. Current systems often overlook the fact that users bring their own expertise, mental models, and cognitive styles. Therefore, explanatory strategies must be both intelligible and behaviorally effective—capable of promoting reflection, mitigating bias, and supporting a sense of agency. Recent research has proposed the introduction of *cognitive friction*—deliberate design features that challenge users' reasoning to counteract passive reliance and encourage mindful engagement [18, 19, 20]. Within this line of work, the paradigm of *Frictional AI* encompasses methods that intentionally introduce cognitive challenges to stimulate critical reflection in human–AI collaboration [21, 22, 23].

## 2. Beyond Accuracy and Trust

Most research on human–AI collaboration in clinical decision-making has focused on performance metrics such as diagnostic accuracy or relational constructs such as trust and reliance [24]. While these are critical outcomes, they overlook an equally important dimension: whether users remain active and accountable participants in the decision-making process. This dimension is often described as the *sense of agency*—the experiential perception of being the originator and owner of one's actions and their consequences [12]. In clinical contexts, agency is not merely a psychological state but a professional requirement: clinicians must justify their diagnostic reasoning, assume responsibility for their decisions, and maintain their diagnostic competence over time. If AI systems erode the sense that decisions are genuinely one's own, they risk undermining accountability and contributing to professional deskilling over time [25, 13].

For this reason, preserving users' ability to experience decisions as authentically theirs has recently been highlighted as a key outcome in the design of decision support systems. Several strands of research have begun to propose alternative design paradigms that aim not only to inform users, but also to preserve their critical reasoning and experiential sense of authorship. These approaches shift the focus from explanations as tools of persuasion or transparency, toward interaction protocols that deliberately foster accountability and user engagement. Hildebrandt [26] introduces the idea of *agonistic machine learning*, where systems expose users to disagreement in order to prevent overconformity to algorithmic advice.

Miller [11] proposes a paradigm shift from recommendation-driven to hypothesis-driven decision support that helps users evaluate competing alternatives. These approaches redefine the role of the system from authoritative oracle to dialogical partner, foregrounding deliberation over persuasion.

A second strand emphasizes the value of deliberate *friction*. Research on Frictional AI [21, 22, 23] shows how cognitive effort can be purposefully introduced into the interaction to disrupt unreflective reliance and stimulate critical engagement. Related work on *cognitive forcing functions* [2] similarly demonstrates that interventions designed to slow down or challenge the decision process can reduce automation bias and foster more mindful judgment. While these strategies may temporarily increase cognitive workload, they aim to preserve long-term competence and accountability in high-stakes contexts.

Other proposals explicitly focus on surfacing dissent. Haselager et al. [27] describe *Reflection Machines*, systems that create deliberate tension by highlighting counterarguments, prompting users to reconsider initial intuitions. Reingold et al. [28] propose *Dissenting Explanations*, where models are trained to generate adversarial viewpoints rather than converge toward consensus. Similarly, Sarkar argues for

systems that occasionally challenge the user [29]. He presents an intriguing conceptual shift from the traditional view of AI as an assistant toward viewing AI as a provocateur. In contrast to AI simply fulfilling user-directed tasks efficiently, a provocateur AI purposefully surfaces counter-arguments, fringe cases, or alternative framings, forcing the user to articulate, defend, and possibly revise their stance. This reframing aligns AI systems with pedagogical objectives, positioning them as tools that facilitate deeper cognitive engagement rather than merely augmenting productivity. These lines of work share the assumption that disagreement, when well-structured, can protect against overreliance and preserve users' engagement as decision-makers.

Despite these advances, empirical studies rarely assess whether such designs truly preserve users' *agency*. Existing work typically relies on proxies such as trust, reliance, or satisfaction, which do not capture whether decisions are still experienced as authentically one's own. To address this gap we propose the *Judicial protocol*, which operationalizes evaluative and frictional principles within clinical DSSs. Unlike conventional DSSs that provide a single authoritative output, Judicial protocols present two contrastive diagnostic alternatives, each supported by persuasive (yet fallible) justifications. The goal is to re-engage users' discriminative capacities by requiring them to actively adjudicate between competing arguments. We implement this paradigm in two variants: the **Alternative Judicial protocol**, in which a single system provides two alternative diagnoses with corresponding justifications, and the **Antagonist Judicial protocol**, in which two separate systems each advocate for a different diagnosis. By comparing these designs to the **Traditional protocol**, which offers a single recommendation with explanation, we aim to investigate how contrastive explanations affect not only diagnostic accuracy and confidence, but also users' perceived agency, responsibility, and the perceived role of AI in the decision-making process. In parallel, we introduce a novel *HCI Sense of Agency scale*, tailored to the clinical decision-making context, which decomposes agency into three dimensions: influence, ownership, and responsibility. This contribution allows us to move beyond accuracy and trust, and to empirically assess whether alternative interaction protocols can sustain clinicians' agency in AI-supported diagnostic practice.

## 3. 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.

## 4. Methods

### 4.1. Participants

To test these hypotheses, a between- and within-subjects experimental design will be implemented (see Figure 1). Participants, comprising medical students and clinicians from the University of Milan, will be randomly assigned to one of two groups: **Alternative Judicial** or **Antagonist Judicial**. Responses to the online survey will be collected anonymously.

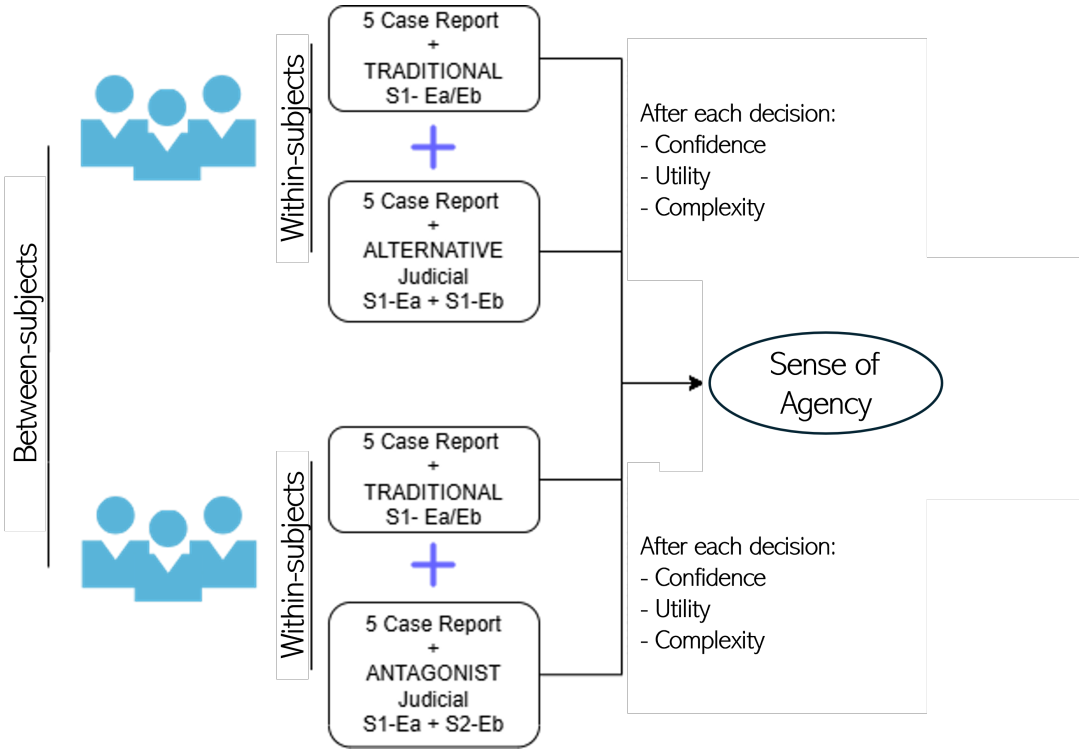


**Figure 1:** Experimental design.

## 4.2. Procedure

The experiment was conducted through an online interface built using LimeSurvey[1]. Each participant evaluated a total of 10 clinical cases, presented in a fixed and predefined order. In the first phase, participants completed 5 cases supported by the *Traditional DSS*, which provided a single diagnostic recommendation accompanied by an explanation. After each case, participants recorded their diagnostic choice, rated their confidence in the decision, assessed the perceived utility of the system, and evaluated the perceived complexity of the case on a 4-point ordinal scale. Upon completion of the five Traditional cases, they filled in the Sense of Agency scale, which included the three constructs of *influence*, *ownership*, and *responsibility*.

In the second phase, participants completed 5 cases supported by one of the Judicial protocols, depending on group assignment: in the *Alternative Judicial* condition, a single DSS presented two alternative diagnoses with corresponding justifications, while in the *Antagonist Judicial* condition two distinct DSSs each advocated for a different diagnosis with their own explanatory arguments. As in the first phase, after each case participants provided their diagnostic decision, confidence rating, system utility rating, and case complexity rating. At the end of this block, they again completed the Sense of Agency scale.

Finally, after completing all 10 cases, participants filled out two standardized psychometric instruments: the short version of the Big Five Inventory (BFI) to measure personality traits, and an adapted version of the Decision Styles Scale (DSS) tailored to the diagnostic context, focusing on rational and intuitive decision-making styles. The analyses aim to reveal both main effects and interaction effects between the DSS format (Traditional vs. Judicial) and the explanation style (Alternative vs. Antagonist) on key outcome variables.

All AI recommendations in the study are simulated to ensure consistency across participants and allow for full control over the diagnostic content. Clinical cases are adapted by an expert clinician from *The New England Journal of Medicine*[2] and include symptomatology, medical history, and lab results

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

that together form a realistic diagnostic scenario. AI explanations in the Judicial conditions are crafted to be persuasive and grounded in the clinical features of each case.

### 4.3. Measures

The results of this study are evaluated through three straightforward measures—diagnostic accuracy, self-reported confidence in one's decision, and self-reported utility of the AI system—alongside a novel instrument designed to capture participants' sense of agency.

Diagnostic accuracy provides the most direct and objective outcome, recorded dichotomously as either correct (1) or incorrect (0), based on the reference diagnosis reported in the source medical literature. This measure establishes a clear baseline against which the influence of different interaction protocols can be assessed.

Complementing this measure, participants also report their confidence in each decision and their perceived utility of the AI system. Both constructs are measured using 4-point ordinal scales specifically designed to reduce central tendency bias, thereby ensuring more discriminative responses. Confidence offers insight into the degree of certainty with which participants endorsed their diagnostic choices, while perceived utility reflects the extent to which the system was judged to be supportive in practice.

Since to our knowledge, there are currently no validated instruments in HCI or clinical decision-making research to directly measure the *sense of agency*, we developed an ad hoc scale for this study.

Building on prior work in psychology, the scale operationalizes agency across three complementary constructs: **Influence** (the extent to which users felt steered or constrained by the system), **Ownership** (the degree to which users experienced the decision as authentically their own), and **Responsibility** (the extent to which users perceived themselves as accountable for the outcomes).

The scale comprises 22 items, presented on a 4-point ordinal scale (1 = strongly disagree, 4 = strongly agree), with an additional *Not applicable* option to account for cases where participants felt an item did not reflect their experience. Negatively worded items (indicated as INV) were included to control for acquiescence bias. To avoid priming effects, items from the three constructs were presented in a randomized order rather than grouped by construct.

## 5. Limitations

This study has several limitations that should be acknowledged. First, the study design may be subject to learning and anchoring effects, as all participants completed the Traditional DSS block prior to the Judicial block. This fixed ordering could artificially inflate or suppress performance and perceived agency in the second block. Future iterations should incorporate counterbalancing or explicitly model trial index to disentangle genuine protocol effects from order-related influences.

Second, the generalizability of findings is constrained by the use of a single-institution convenience sample consisting of medical students and clinicians drawn from one context. Moreover, the decision support system was simulated rather than integrated into a real-world clinical workflow, which may limit ecological validity and the applicability of results to actual practice environments.

Third, the repeated administration of the agency scale after each block may introduce scale reactivity. By repeatedly prompting participants to reflect on their sense of influence, ownership, and responsibility, the measure itself may shape the construct under study. Future work should examine alternate forms of the instrument, reduce measurement frequency, or introduce spacing strategies to minimize such reactivity.

Finally, the study's reliance on binary accuracy scoring simplifies diagnostic reasoning to a dichotomous correct/incorrect outcome. This approach may overlook partial correctness, reasonable differential diagnoses, or clinically meaningful prioritization of possibilities. Future research could adopt graded scoring schemes or use expert panel adjudication to capture the nuance of diagnostic quality.

**Table 1**
Sense of Agency Scale items, organized by construct (not presentation order).

| Construct | Item |
|---|---|
| Influence | INF1: The system deliberately tried to steer me toward particular decisions.<br>INF2: I felt that my final choices were manipulated by the system's recommendations.<br>INF3: The way information was presented during decision-making exerted undue influence on my decisions.<br>INF4: I believe the system's design restricted my freedom to choose freely in my decision-making.<br>INF5 (INV): I felt the system did not push me toward biased decisions.<br>INF6: The system restricted my ability to decide as I wanted.<br>INF7: The system's recommendations made me question my own judgment or intuition several times.<br>INF8: The information presented by the system influenced my decision-making process.<br>INF9 (INV): I found it easy to ignore the system's advice when I wanted. |
| Responsibility | RES1: I consider myself fully responsible for the outcome of the decisions I made using the system.<br>RES2: If the decisions lead to negative consequences, I expect that others will hold me, not the system, accountable.<br>RES3: I am prepared to justify and defend my decisions to an external reviewer or authority.<br>RES4: Any legal or financial liability arising from my decisions ultimately rests with me.<br>RES5 (INV): Because I followed the system's advice, I feel less responsible for the consequences of my decisions.<br>RES6: I consider it my responsibility to critically evaluate the system's recommendations before deciding. |
| Ownership | OWN1: I felt free to make my own choice even when the system offered suggestions.<br>OWN2: The final decisions reflect my own preferences or knowledge rather than the system's.<br>OWN3: At every stage, I felt fully in control of the decision-making process.<br>OWN4: Even without the system's recommendations, I would have reached essentially the same decisions.<br>OWN5: I regard the decisions I reached through the system as entirely my own.<br>OWN6 (INV): The decisions felt as though they belong more to the system than to me.<br>OWN7: I felt confident that the decisions were based primarily on my judgment rather than on the system's influence. |

## 6. Conclusions

The growing discourse on frictional AI has emphasised the need for interaction protocols that move beyond accuracy and trust as the only measures of success. Clinical decision support systems must not only inform but also sustain clinicians' capacity to deliberate, justify, and ultimately own their decisions. Yet, empirical tools to assess such beyond-utility outcomes remain scarce. This study takes a step in that direction by proposing a dual contribution: first, the introduction of judicial AI protocols that operationalise contrastive and agonistic principles through Alternative and Antagonist designs; and second, the development of a multidimensional Sense of Agency scale that captures clinicians' perceived influence, ownership, and responsibility in diagnostic tasks.

Together, these contributions set the ground for a systematic evaluation of whether deliberately frictional interaction protocols can preserve professional agency without compromising diagnostic performance. By reframing decision support from oracular recommendation to judicial adjudication, our work foregrounds the clinician as an active arbiter rather than a passive recipient of algorithmic

advice. The proposed scale, in turn, offers a means to empirically test this claim and to anchor design choices in robust measures of agency.

Looking ahead, the framework outlined here aspires to shift both research and design practice: from interventions judged solely by short-term accuracy gains, towards systems that also safeguard the longer-term values of accountability, responsibility, and professional competence. In doing so, we hope to contribute to a broader reorientation of Human–AI Interaction—one in which diagnostic AI is not only evaluated for what it adds, but also for what it preserves.

## Acknowledgments

C. Fregosi and F. Cabitza acknowledge funding support provided by the Italian project PRIN PNRR 2022 InXAID - Interaction with eXplainable Artificial Intelligence in (medical) Decision making. CUP: H53D23008090001 funded by the European Union - Next Generation EU.

C. Natali acknowledges the financial support provided by the Federal Commission for Scholarships for Foreign Students in the form of the Swiss Government Excellence Scholarship (ESKAS No. 2024.0002) for the academic year 2024-25.

## Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT-5 to improve readability (grammar, style, and clarity). The tool was not used to generate ideas or substantive content, and all text was reviewed and edited by the authors.

## References

[1] T. Kliegr, Š. Bahník, J. Fürnkranz, A review of possible effects of cognitive biases on interpretation of rule-based machine learning models, Artificial Intelligence 295 (2021) 103458.

[2] 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.

[3] F. Cabitza, A. Campagner, R. Angius, C. Natali, C. Reverberi, Ai shall have no dominion: on how to measure technology dominance in ai-supported human decision-making, in: Proceedings of the 2023 CHI conference on human factors in computing systems, 2023, pp. 1–20.

[4] M. Vered, T. Livni, P. D. L. Howe, T. Miller, L. Sonenberg, The effects of explanations on automation bias, Artificial Intelligence 322 (2023) 103952.

[5] H.-P. H. Lee, A. Sarkar, L. Tankelevitch, I. Drosos, S. Rintel, R. Banks, N. Wilson, The impact of generative ai on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers (2025).

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

[7] M. Schemmer, N. Kuehl, C. Benz, A. Bartos, G. Satzger, Appropriate reliance on ai advice: Conceptualization and the effect of explanations, in: Proceedings of the 28th International Conference on Intelligent User Interfaces, 2023, pp. 410–422.

[8] A. Schmitt, T. Wambsganß, M. Söllner, A. Janson, Towards a trust reliance paradox? exploring the gap between perceived trust in and reliance on algorithmic advice, in: Proceedings of the International Conference on Information Systems (ICIS) 2021, 2021.

[9] Z. Guo, Y. Wu, J. D. Hartline, J. Hullman, A decision theoretic framework for measuring ai reliance, in: The 2024 ACM Conference on Fairness, Accountability, and Transparency, 2024, pp. 221–236.

[10] R. A. Miller, F. Masarie Jr, The demise of the "greek oracle" model for medical diagnostic systems, Methods of information in medicine 29 (1990) 1–2.

[11] 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.
[12] J. W. Moore, What is the sense of agency and why does it matter?, Frontiers in psychology 7 (2016) 1272.
[13] C. Natali, L. Marconi, L. D. Dias Duran, F. Cabitza, Ai-induced deskilling in medicine: A mixed-method review and research agenda for healthcare and beyond, Artificial Intelligence Review 58 (2025) 1–40.
[14] 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.
[15] 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.
[16] 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.
[17] 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.
[18] A. Cooper, The inmates are running the asylum, Springer, 1999.
[19] A. L. Cox, S. J. Gould, M. E. Cecchinato, I. Iacovides, I. Renfree, Design frictions for mindful interactions: The case for microboundaries, in: Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems, 2016, pp. 1389–1397.
[20] 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.
[21] F. Cabitza, A. Campagner, D. Ciucci, A. Seveso, Programmed inefficiencies in dss-supported human decision making, in: Modeling Decisions for Artificial Intelligence: 16th International Conference, MDAI 2019, Milan, Italy, September 4–6, 2019, Proceedings 16, Springer, 2019, pp. 201–212.
[22] C. Natali, et al., Per aspera ad astra, or flourishing via friction: Stimulating cognitive activation by design through frictional decision support systems, in: CEUR workshop proceedings, volume 3481, CEUR-WS, 2023, pp. 15–19.
[23] F. Cabitza, C. Natali, L. Famiglini, A. Campagner, V. Caccavella, E. Gallazzi, Never tell me the odds: Investigating pro-hoc explanations in medical decision making, Artificial intelligence in medicine 150 (2024) 102819.
[24] C. Natali, A. Campagner, F. Cabitza, Answering the call to go beyond accuracy: An online tool for the multidimensional assessment of decision support systems., in: BIOSTEC (2), 2024, pp. 219–229.
[25] 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.
[26] M. Hildebrandt, Privacy as protection of the incomputable self: From agnostic to agonistic machine learning, Theoretical Inquiries in Law 20 (2019) 83–121.
[27] 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.
[28] 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.
[29] A. Sarkar, Ai should challenge, not obey, Communications of the ACM 67 (2024) 18–21.