SCUOLA DI DOTTORATO
UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Department of **Informatics, Systems and Communication**

Ph. D. program in **Computer Science**, **XXXVIII** cycle

Frictional AI.
Countering Over-Reliance on Clinical Decision Support Systems via Desirable Inefficiencies

Chiara Natali
ID: 897224

Tutor: **Prof. Paolo Napoletano**
Supervisor: **Prof. Federico Cabitza**

Coordinator: **Prof. Leonardo Mariani**

Academic Year **2024/2025**

At present machinery competes against man.
Under proper conditions machinery will serve man.
There is no doubt at all that this is the future of machinery,
and just as trees grow while the country gentleman is asleep,
so while Humanity will be amusing itself, or enjoying cultivated
leisure—which, and not labour, is the aim of man—
or making beautiful things, or reading beautiful things,
or simply contemplating the world with admiration and delight,
machinery will be doing all the necessary and unpleasant work. (. . .)
Is this Utopian? A map of the world that does not include
Utopia is not worth even glancing at, for it leaves out
the one country at which Humanity is always landing.
And when Humanity lands there, it looks out,
and, seeing a better country, sets sail.
Progress is the realisation of Utopias.

— Oscar Wilde, *The Soul of Man Under Socialism*
(Wilde, 1918, pp. 32–34)

ABSTRACT

The introduction of Artificial Intelligence as a seamless support to diagnostic decision-making in medicine holds promise to improve efficiency and patient outcomes. Yet gains in model accuracy do not guarantee clinical benefit. Empirical studies show that AI, once embedded in workflow, can unintentionally weaken the conditions that sustain expert judgement: narrowing attention, suppressing the evaluation of alternatives, and encouraging premature acceptance of algorithmic suggestions. Over time, such patterns risk inappropriate reliance, reduced vigilance, and erosion of diagnostic skills.

This thesis argues that the central challenge is not to improve predictive performance alone, but to actively shape how humans and AI systems interact so as to sustain diagnostic reasoning and preserve professional agency—recognising that this may require rethinking doctrines equating good design with seamlessness and cognitive ease.

To this end, the thesis advances *Frictional AI* as a principled and actionable design paradigm. It develops an interaction protocol perspective and a beyond-accuracy evaluation framework to capture how AI influences decision-making across expertise levels, and consolidates evidence on AI-induced deskilling and upskilling inhibition in medicine. Through controlled studies, it demonstrates that explanation-centred strategies often fail to correct reliance, informing a typology of misleading explanations. The thesis then offers a conceptual synthesis of frictional strategies in AI decision support, and evaluates them empirically through user studies involving radiologists. Results show that structured cognitive friction can reduce inappropriate reliance without unacceptable performance penalties, and is experienced by practitioners as aligned with the interpretive logic of diagnostic work.

The thesis contributes a conceptual foundation, evaluative methodology, and empirical evidence supporting frictional design as a viable path toward responsible and durable human–AI integration in clinical decision-making, reframing friction as a resource for cognitive engagement and skill sustainment.

PUBLICATIONS

Anichini, Giulia, Chiara Natali, and Federico Cabitza (2024). "Invisible to machines: designing AI that supports vision work in radiology." In: *Computer Supported Cooperative Work (CSCW)* 33.4, pp. 993–1036.

Cabitza, Federico, Andrea Campagner, Riccardo Angius, Chiara Natali, and Carlo Reverberi (2023a). "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*, pp. 1–20.

Cabitza, Federico, Andrea Campagner, Lorenzo Famiglini, Chiara Natali, Valerio Caccavella, and Enrico Gallazzi (2023b). "Let me think! investigating the effect of explanations feeding doubts about the AI advice." In: *International cross-domain conference for machine learning and knowledge extraction*. Springer, pp. 155–169.

Cabitza, Federico, Andrea Campagner, Gianclaudio Malgieri, Chiara Natali, David Schneeberger, Karl Stoeger, and Andreas Holzinger (2023c). "Quod erat demonstrandum?-Towards a typology of the concept of explanation for the design of explainable AI." In: *Expert systems with Applications* 213, p. 118888.

Cabitza, Federico, Andrea Campagner, and Chiara Natali (2023). "Decision Support System Quality Assessment Tool." In: *Proceedings of the 15th Biannual Conference of the Italian SIGCHI Chapter*, pp. 1–4.

Cabitza, Federico, Andrea Campagner, Chiara Natali, Enea Parimbelli, Luca Ronzio, and Matteo Cameli (2023d). "Painting the black box white: experimental findings from applying XAI to an ECG reading setting." In: *Machine Learning and Knowledge Extraction* 5.1, pp. 269–286.

Cabitza, Federico, Caterina Fregosi, Andrea Campagner, and Chiara Natali (2024a). "Explanations considered harmful: the impact of misleading explanations on accuracy in hybrid human-ai decision making." In: *World conference on explainable artificial intelligence*. Springer, pp. 255–269.

Cabitza, Federico and Chiara Natali (2022). "Open, multiple, adjunct. decision support at the time of relational AI." In: *Frontiers in Artificial Intelligence and Applications* 354, pp. 243–245.

Cabitza, Federico, Chiara Natali, Lorenzo Famiglini, Andrea Campagner, Valerio Caccavella, and Enrico Gallazzi (2024b). "Never tell me the odds: Investigating pro-hoc explanations in medical decision making." In: *Artificial intelligence in medicine* 150, p. 102819.

Cabitza, Federico, Chiara Natali, Francesco Varanini, and David Gunkel (2025). "Beyond cyborgs: the cybork idea for the de-individuation

of (artificial) intelligence and an emergence-oriented design." In: *AI & SOCIETY* 40.5, pp. 3333–3348.

Facchini, Alessandro, Caterina Fregosi, Chiara Natali, Alberto Termine, and Ben Wilson (2024). "Algorithmic Authority & AI Influence in Decision Settings: Theories and Implications for Design." In: *Proceedings of the 12th International Conference on Human-Agent Interaction*, pp. 472–474.

Fregosi, Caterina, Andrea Campagner, Chiara Natali, and Federico Cabitza (2024). "Assessing appropriate reliance: a framework for evaluating AI influence on user decision-making." In: *AIxHMI@ AI* IA*.

Fregosi, Caterina, Chiara Natali, and Federico Cabitza (2025). "Judicial Protocols in Diagnostic AI: Contrastive Explanations to Preserve Human Agency." In: *Proceedings of the Workshops at the Fourth International Conference on Hybrid Human-Artificial Intelligence (HHAI-WS 2025)*. Pisa, Italy.

Milella, Frida, Chiara Natali, Teresa Scantamburlo, Andrea Campagner, and Federico Cabitza (2023). "The impact of gender and personality in human-AI teaming: The case of collaborative question answering." In: *IFIP Conference on Human-Computer Interaction*. Springer, pp. 329–349.

Natali, Chiara (2023). "Per aspera ad astra, or flourishing via friction: Stimulating cognitive activation by design through frictional decision support systems." In: *CEUR workshop proceedings*. Vol. 3481. CEUR-WS, pp. 15–19.

Natali, Chiara and Federico Cabitza (2025). "Make Some Noise for Ground Truthing! Frictional design against epistemic sclerosis in Decision Support Systems." In.

Natali, Chiara, Andrea Campagner, and Federico Cabitza (2024). "Answering the Call to Go Beyond Accuracy: An Online Tool for the Multidimensional Assessment of Decision Support Systems." In: *Proceedings of the 17th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2024)*. Vol. 2. SCITEPRESS, pp. 219–229.

Natali, Chiara, Lorenzo Famiglini, Andrea Campagner, Giovanni Andrea La Maida, Enrico Gallazzi, and Federico Cabitza (2023). "Color shadows 2: Assessing the impact of xai on diagnostic decision-making." In: *World Conference on Explainable Artificial Intelligence*. Springer, pp. 618–629.

Natali, Chiara, Brett Frischmann, and Federico Cabitza (2024). "Stimulating Cognitive Engagement in Hybrid Decision-Making: Friction, Reliance and Biases (preface)." In: *CEUR workshop proceedings*. Vol. 3825. CEUR-WS, pp. 9–15.

Natali, Chiara, Luca Marconi, Leslye Denisse Dias Duran, and Federico Cabitza (2025a). "AI-induced Deskilling in Medicine: A Mixed-

Method Review and Research Agenda for Healthcare and Beyond." In: *Artificial Intelligence Review* 58.11, pp. 1–40.

Natali, Chiara, Luca Marconi, Caterina Fregosi, and Federico Cabitza (2024). "Humans in the Group, Computers in the Coop. Comparison of Individual and Collective Improvement in Cognitive Tasks in Adjunct AI Settings." In: *IFIP Working Conference on Human Work Interaction Design*. Springer, pp. 174–191.

Natali, Chiara, Mohammad Naiseh, Federico Cabitza, and Brett Frischmann (2025b). "Better AI with Designed Friction: Theories, Applications and Research Agenda." In: *Hybrid Human–Artificial Intelligence*. Vol. 408. Frontiers in Artificial Intelligence and Applications. IOS Press, pp. 518–521. DOI: 10.3233/FAIA250680.

Wilson, Ben, Chiara Natali, Matt Roach, Darren Scott, Alma Rahat, David Rawlinson, and Federico Cabitza (2025). "Dimensions of human-machine combination: prompting the development of deployable intelligent decision systems for situated clinical contexts." In: *Computer Supported Cooperative Work (CSCW)*, pp. 1–57.

d'Aiello, Angelo Fabio, Federico Cabitza, Chiara Natali, Sophia Viganò, Paolo Ferrero, Ludovica Bognoni, Giulia Pasqualin, Alessandro Giamberti, and Massimo Chessa (2023). "The effect of holographic heart models and mixed reality for anatomy learning in congenital heart disease: an exploratory study." In: *Journal of Medical Systems* 47.1, p. 64.

ACKNOWLEDGMENTS

Intelligence is situated, and emerges in interaction. This thesis is therefore not just mine, but the product of the hearts and places that welcomed me during the three unforgettable years of my PhD.

My first welcome into this journey came from my supervisor, Prof. Federico Cabitza. I am deeply grateful for his guidance throughout what has been the most intellectually enriching experience of my life, and for being the first to believe in this work. At this moment of pride, fulfilment, and personal growth, I clearly recognise the lessons—intellectual and human—that accompanied it.

I also thank Prof. *Frida Milella* for her support and the lactose-free cappuccinos that marked our shared teaching experience in Interaction Design, and Prof. *Massimo Miglioretti* for teaching me not through an ordinary class, but by example, modelling the professional I aspire to become.

Caterina Fregosi, for her heart of gold and your infectious joy, which shines through during long afternoons in the laboratory as it does in shared artsy flats around Europe while getting ready for conferences, workshops, summer schools. As you said, it is *reciprocal support from day one*. Thank you for inspiring me in the big and the little things alike. *Lorenzo Rovida*, for being my first friend in Bicocca: absolute wholesomeness extract, which is all the more impressive considering he is running just on a macchiato decaf. *Andrea Campagner*, for the mentoring and support, as I think a part of me will always be the novice in the lab who looks up to him. To *Alessia Papale* and *Gloria Lopiano* — although our time in the lab overlapped briefly, you made the final stretch brighter.

My year in Lugano at IDSIA SUPSI will remain a happy place to revisit during more than a few daydreams. Thanks to Prof. *Alessandro Facchini*, Prof. *Elisa Rubegni*, for their invaluable mentorship. *Stefano Damato*, *Franca Corradini*, for their invaluable friendship. *Alberto Termine*, for both. *Maria Milheiro*, for nudging me to plunge into the Lugano lake to wash off all academic worries, and teaching me *sueca*.

My research was made all the more precious and meaningful thanks to the engagement with the *Hybrid Human Artificial Intelligence* (HHAI) community. Thank you especially to Prof. *Ilaria Tiddi* and Prof. *Stefan Schlobach*, for opening my eyes and my heart to such a lively schol-

arly environment, and for opening their arms in welcoming me into it. Thank you to *Lorenzo Valerio* for the support as co-Proceedings & Publicity Chair during HHAI 2025.

To Prof. *Brett Frischmann* and the Frictional AI community. Thank you for having made this journey more meaningful–and fun. My warmest thanks for their friendship and support to *Simon B. Fischer*, *Leslye Denisse Dias Duran*, *Regina De Brito Duarte*, *Ben Wilson*.

And for all the friends from other disciplines I shared the PhD journey with, each with their wisdom, perspective, and irreplaceable spot in my heart: *Ylenia Baldanza*, *Lucia Cucchi*, *Vivian Grillo*, *Ilaria Iannuccilli*, *Vanessa Nardini*, *Camilla Pedrelli*. And all my friends from beyond academia, who gave me all their love and support even when my academic worries made little sense to them – often they were right to think so, and I am grateful for that.

I would not be me, and half of my papers would not have seen the light of the day, without cafés where to write, gluten free bakeries where to recharge at, and jazz clubs where to forget about it all – There is a bit of *GluFree Bakery*, *ZeroPensieri*, *Yellowsquare* and *Corte dei Miracoli* in each of these pages.

To *Corte dei Miracoli*, discovered the same day I started my doctorate, for introducing into my life more music, more spontaneity–a place where friends are made in a second and stay for life, at the small cost of sleep-deprived Monday mornings in the lab with ears still ringing from the Sunday jazz jam. What a marvelous full circle to raise a toast for my doctorate there!

Finally, this thesis was written for the most part in a condition of voluntary nomadism. A heartfelt thank you to the *Bavetta* family (and *Shimmi*) in Enfield and to *Marta Cignetti* in Paris for supporting me during the very last stretch.

To my family, who is my privilege. In this writing effort, I remembered what you would say when we went up the mountains – do not look how far away the destination is, just walk. Even if the knees are shaking a bit.

To *Jacob*. Thank you for being my rock, and all that jazz. ♪

FUNDING ACKNOWLEDGMENTS: I gratefully acknowledge 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 2024–2025.

I gratefully acknowledge the PhD grant awarded by the Fondazione Fratelli Confalonieri for the academic year 2023–2024, which has been instrumental in facilitating my research pursuits.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

AB Automation Bias

AI Artificial Intelligence

AIER AI Error Rate

AIM Breast cancer detection system

AIT Pneumothorax detection system

AM Activation Maps

CASA Computers Are Social Actors

CB Conservatism Bias

CSCW Computer Supported Cooperative Work

DSS Decision Support Systems

ECG Electrocardiogram

EHR Electronic Health Record

FHD Final Human Decision

GPT Generative Pre-trained Transformer

HAI Human-AI Interaction

HAI-IPs Human-AI Interaction Protocols

HCAI Human-Centred Artificial Intelligence

HCI Human-Computer Interaction

HD1 First Human Decision

HE Halo Effect

HHAI Hybrid Human Artificial Intelligence

HIC Human-in-Command

HITL Human-in-the-Loop

HOTL Human-on-the-Loop

IxD Interaction Design

LLMs Large Language Models

MRI Magnetic Resonance Imaging

STS Science, Technology and Society

WBP White Box Paradox

XAI eXplainable AI

1

INTRODUCTION

> *Increasing the effectiveness of the individual's use of his basic capabilities is a problem in redesigning the changeable parts of a system. [. . .] To redesign a structure, we must learn as much as we can of what is known about the basic materials and components as they are utilized within the structure; beyond that, we must learn how to view, to measure, to analyze, and to evaluate in terms of the functional whole and its purpose. In this particular case, no existing analytic theory is by itself adequate for the purpose of analyzing and evaluating overall system performance; pursuit of an improved system thus demands the use of experimental methods.*
>
> — Douglas C. Engelbart, *Augmenting Human Intellect*
>
> (Engelbart, 2023, p. 23)

1.1 RESEARCH MOTIVATION

In current design paradigms, AI systems tend to function as oracles (Miller and Masarie Jr, 1990) – they output a prediction or decision which the human is expected to trust, largely without deeper interaction. This one-directional, automation-centric approach can cause over-reliance on algorithmic outputs: through *automation bias*, users may default to the AI's answer even when they have reservations, effectively abdicating active decision-making (Skitka, Mosier, and Burdick, 1999).

One major consequence is deskilling. When professionals routinely defer to automated recommendations, their own skills and situational awareness may atrophy over time, following the neurologic principle of "use it or lose it" (Shors et al., 2012). For example, physicians who rely on AI may lose the habit of thoroughly examining patients or forming independent differentials, and junior doctors get fewer opportunities to practice core skills. A vivid illustration comes from surgical training with robotic assistants: senior surgeons using surgical robots can perform operations with less need for assistance, meaning trainees are no longer invited to "see one, do one, teach one" as before (Beane, 2019). This points to a systemic issue: automation can interrupt the pipeline of skill transmission. Over time, an organization risks losing its reservoir of expertise; seasoned workers may no longer be capable of mentoring novices in tasks they themselves have ceded to AI. Deskilling is not just an individual problem but an organizational

and societal one, weakening the resilience and adaptive capacity of entire professions.

Compounding the issue is what Cabitza (2021a) calls *epistemic sclerosis*, a hardening of knowledge caused by AI's tendency to perpetuate historical data patterns. As AI oracles deliver answers based on past training data, they can ossify current practice – making it less likely that humans will question prevailing assumptions or explore novel solutions. Clinicians might accept an AI's diagnosis suggestion as definitive, rather than considering an atypical condition not in the algorithm's model. Over time, the breadth of human expertise narrows, and innovation or creative problem-solving is stifled by an overconfidence in the algorithm. Epistemic sclerosis thus goes hand-in-hand with deskilling: not only are skills lost, but the very knowledge landscape freezes in place. In fields like medicine that evolve with new research and rare cases, this rigidity is dangerous.

All these factors culminate in precarious situations often described through the metaphor of *moral crumple zones*. As Elish (2019) observed, when highly automated systems fail, it is the human operator – now out-of-the-loop and perhaps deskilled – who absorbs the blame and consequences. This phenomenon underscores to an irony of automation (Bainbridge, 1983)a: the human operator is expected to take a monitoring role to AI, and to intervene heroically in emergencies, exactly when their skills and awareness have been dulled by sustained automation use. This arrangement creates a "responsibility gap" (Sio and Hoven, 2018) in which the human is formally accountable but not truly empowered to ensure good outcomes.

Finally, it is increasingly clear that standard fixes like eXplainable AI (XAI) are not panaceas for these deep-rooted problems. While explanations of AI outputs are often touted as a solution for users will then properly calibrate their trust [REF], evidence shows that this assumption is tenuous: explanations alone do not reliably prevent over-trust or misuse (Bertrand et al., 2022). In some cases, adding an explanation can create a mere illusion of understanding (Kaur et al., 2024) that includes the case of plausible-sounding (or -looking) *placebic* explanations (Eiband et al., 2019). For example, a convincingly highlighted feature (say, a medical image region) might persuade a doctor that the AI's prediction is grounded, when in reality the explanation method is imperfect and the model is wrong. The net effect is that XAI, in its current form, can at times fail to resolve the core issue of mis-calibrated trust and may may reinforce misplaced trust rather than promote skep ticism and verification" (Brdnik, Colakovic, and Karakatič, 2025, p. 185). It addresses symptoms (opacity) but not the underlying dynamics: the user's reduced engagement, the loss of skill, and the structural reliance on the AI's outputs.

In fact, focusing on transparency alone can distract from more impactful design interventions. What is needed is a broader rethinking of the human–AI collaboration protocol (Cabitza et al., 2023e): explanation is not sufficient if the workflow still casts the human as a relatively passive checker of the AI's work. Meaningful solutions must go further to ensure the human remains an active, critical participant throughout.

In summary, the prevailing approach of deploying AI as oracle-like decision aids – even when augmented with basic explainability – tends to engender a cascade of negative outcomes: uncritical use of the AI, human deskilling and upskilling inhibition, epistemic stagnation in the domain, and brittle human oversight that can collapse in crisis, creating "moral crumple zones." These interlocking problems highlight why a more radical shift is needed in how we design AI for high-stakes domains.

This thesis takes the field of medical AI as an empirical testbed to probe questions on the impact of AI on human judgment in high-risk, high-knowledge settings. While Artificial Intelligence is entering clinical practice under the promise of augmenting decision-making and improving patient outcomes, evidence from deployed decision support suggests a persistent mismatch between *technical performance* and *clinical impact* (Aristidou, Jena, and Topol, 2022). Models that perform well in isolation can, once embedded in everyday workflows, alter how clinicians attend to evidence, handle uncertainty, and distribute responsibility. In high-stakes settings, these interaction effects are not peripheral: they can shape diagnostic vigilance, patterns of reliance, and the maintenance of professional competence.

This thesis is motivated by a growing body of empirical evidence that AI can reshape cognition and practice in ways that are not visible through accuracy metrics alone. Even highly performing systems can induce inappropriate reliance and automation bias, redistribute competence unevenly across expertise levels, and weaken clinicians' capacity to operate independently of technology. These risks are especially consequential in medicine, where uncertainty is endemic, categories are negotiated in practice, and responsibility cannot be delegated to an artefact. Prevailing responses often appeal to increasing system transparency through eXplainable AI: however, explanations can become persuasive artefacts that inflate confidence without improving calibration.

What is missing is a design and evaluation vocabulary that treats human–AI collaboration as an emergent system, enabling for trade-off considerations that balance the imperative of accuracy with other, human-centred considerations: the maintenance of skill, the promotion of reflective judgment, and responsibility. Therefore, the motivation

of this thesis is addressing these failure modes at their root, fostering systems that truly support human expertise rather than supplant or dull it.

Accordingly, the thesis asks:

> *How can Human–AI interaction protocols be designed and evaluated to sustain diagnostic reasoning, prevent inappropriate reliance, and enable robust hybrid intelligence in situated clinical work?*

To address this question, the thesis develops and tests the claim that *interaction protocols are a primary lever of clinical AI impact*. On this basis, the thesis develops *Frictional AI*: an interaction-centred approach to decision support that treats well-calibrated cognitive friction as a resource for sustaining diagnostic reasoning and appropriate reliance. Rather than eliminating effort, frictional protocols introduce structured moments of engagement (e.g., commitment, contestation, comparison) designed to mitigate interaction-induced bias and reduce the long-term risk of deskilling.

1.2 INTERDISCIPLINARY POSITIONING

The positioning of this thesis within Human-AI Interaction is necessarily interdisciplinary, because the phenomenon under investigation—clinical AI in practice—is irreducibly socio-technical. Clinical AI is simultaneously a computational artefact that produces statistical outputs, an interactional system that shapes human judgement, and an organisational intervention that redistributes responsibility, authority, and skill. No single disciplinary lens adequately captures these coupled dynamics.

This thesis is grounded on theories and approaches of Computer-Supported Cooperative Work (CSCW) and Hybrid Intelligence (HI). CSCW, in has shown that professional work is adaptive, situated, and negotiated, often exceeding what formalised technical systems anticipate.

Cognitive science provides additional tools for analysing how these reconfigurations affect human reasoning. Dual-process accounts clarify how AI support can differentially engage fast, intuitive judgements and slower, analytic reasoning, while research on cognitive biases offers empirically tractable constructs for assessing over-reliance, under-reliance, and shifts in vigilance. Distributed cognition extends the unit of analysis beyond the individual, framing clinical decision-making as an achievement of human–AI configurations in which reasoning is distributed across people, artefacts, representations, and routines.

Hybrid Intelligence integrates these strands by explicitly framing AI as an augmenting component of joint cognitive systems rather than as a replacement for human expertise. Adopting an HI lens shifts

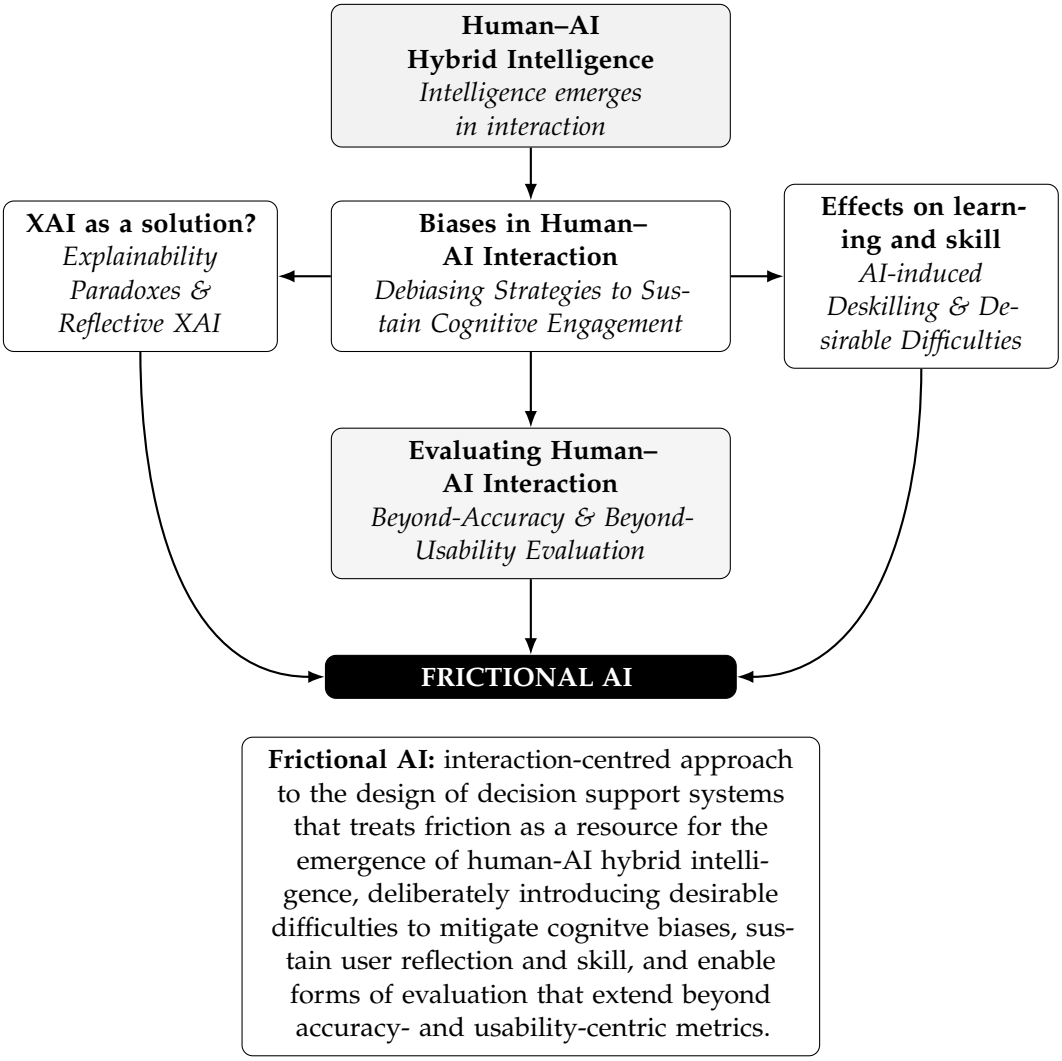


Figure 1.1: Conceptual synthesis of theoretical and evaluative foundations leading to *Frictional AI*.

both design and evaluation away from automated verdict generation towards collaborative problem-solving, where effectiveness depends on how human and machine contributions are orchestrated over time. This orientation also motivates evaluation approaches that go beyond individual accuracy to examine joint outcomes, learning effects, and longer-term impacts on skill and responsibility.

These traditions are not trivially compatible. The thesis integrates them through an emergence-oriented stance: intelligence, competence, and accountability are analysed at the level of situated human–AI configurations rather than isolated agents or models (Figure 1.2). Methodologically, this entails combining interaction design, controlled evaluation, and practice-sensitive interpretation. Substantively, it motivates the thesis's central proposal that beneficial friction—rather than frictionless optimisation—can serve as a principled design strategy for

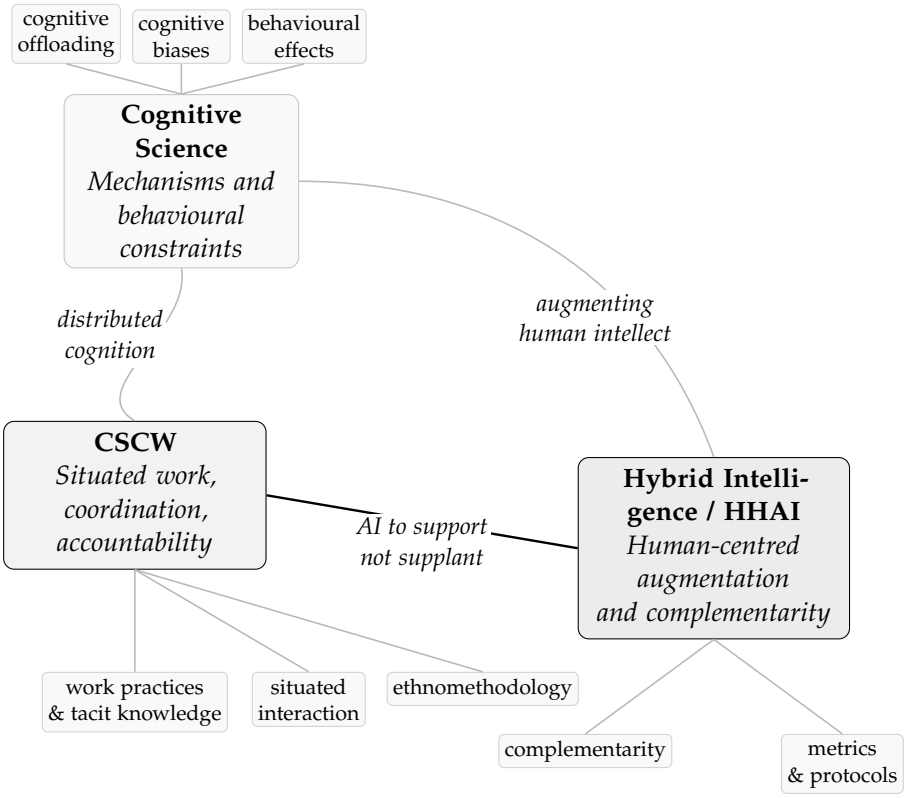


Figure 1.2: Disciplines bridged by the thesis in its conceptual framing. CSCW and Hybrid Intelligence/HHAI form the primary axis, while Cognitive Science provides auxiliary constraints and mechanisms that inform augmentation and its evaluation.

sustaining human judgement, skill, and organisational resilience in high-stakes decision support.

1.3 CONTRIBUTIONS

This thesis makes three classes of contribution to Human–AI interaction in high-stakes decision support, foregrounding the introduction of *Frictional AI* as illustrated in Figure 1.3.

CONCEPTUAL CONTRIBUTION: COLLABORATION PROTOCOLS AS THE UNIT OF HYBRID INTELLIGENCE. The thesis advances an emergence-oriented account of human–AI collaboration in which intelligence and responsibility are properties of socio-technical configurations. It reframes the design object from the model to the *interaction protocol*, providing conceptual vocabulary for analysing how sequencing, authority, and contestation shape judgment and learning in clinical work.

METHODOLOGICAL CONTRIBUTION: BEYOND-ACCURACY EVALUATION OF HUMAN–AI SYSTEMS. The thesis develops and oper-

ationalises evaluation approaches that go beyond predictive performance to capture reliance patterns, dominance and influence effects, and distributive consequences across expertise levels. These measures provide a way to evaluate whether apparent gains reflect genuine synergy or risky deference.

EMPIRICAL AND DESIGN CONTRIBUTION: FRICTIONAL AI PROTOCOLS IN CLINICAL DECISION-MAKING. Through empirical studies in radiology and complementary qualitative evidence, the thesis designs and evaluates frictional interaction protocols. The results show how structured friction can mitigate inappropriate reliance while remaining compatible with clinical reasoning and professional expectations.

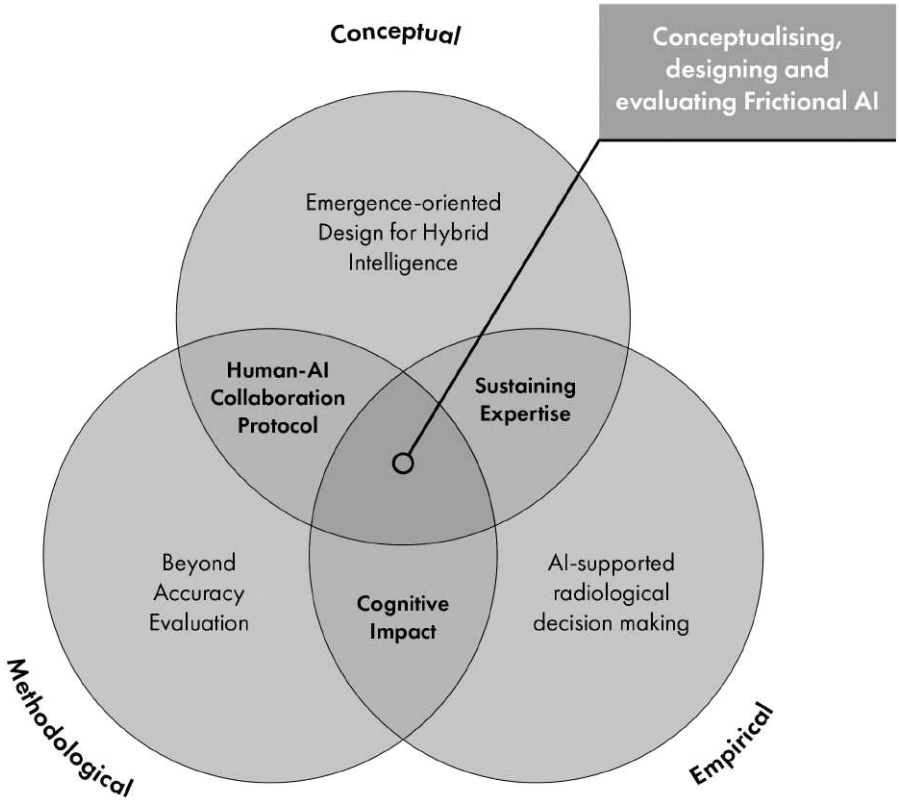


Figure 1.3: Summary of research contributions across the conceptual, methodological and empirical levels. The thesis advances an interaction-centred account of hybrid intelligence, introduces beyond-accuracy methods for evaluating human–AI interaction, and provides empirical evidence from AI-supported radiological decision-making. Their integration grounds Frictional AI as a design paradigm based on interaction protocols that sustain human judgement and professional expertise in high-stakes clinical settings.

The empirical and conceptual work presented in this thesis has been conducted within collaborative research programmes at MUDI Lab and with international partners. In Section 7.1 I clarify my specific contributions to each co-authored publication underpinning the thesis.

Beyond the individual studies, the thesis makes an additional integrative contribution by unifying conceptual, methodological, and empirical strands into a coherent paradigm of *Frictional AI* and by articulating its implications for the design, evaluation, and governance of clinical decision support. This integrates the conceptual, methodological, and empirical strands into a unified framework for frictional, interaction-centred Hybrid Intelligence, as elaborated in Chapter 11.

1.4 ROADMAP

The thesis is organised in three parts.

PART I: CONCEPTUAL AND CRITICAL FOUNDATIONS. Chapter 2 situates AI decision support systems within the realities of diagnostic practice, focusing on their reconfiguration of professional expertise and the cognitive biases they may introduce or amplify. Chapter 3 characterises *emergence-oriented design* as an interdisciplinary perspective in Human-AI interaction and beyond, and develops the notion of the Human–AI Collaboration Protocol as a unit of design and analysis. Chapter 4 examines the limits of explainability as a general safeguard against inappropriate reliance, arguing that explanations must be assessed in terms of their behavioural effects rather than transparency alone, and discusses examples of *Reflective XAI* aimed at fostering user reflection. Chapter 10 situates this work within the broader discourse on reflection-inducing design in Human–AI interaction, with particular attention to strategies based on deliberately introduced interactional friction or *cognitive forcing*. Finally, Chapter 6 articulates the scientific gap and research programme that motivate the empirical studies presented in this thesis.

PART II: FINDINGS Following an overview of the featured studies and clarification of my authorial contributions (Chapter 7), Part II presents the core findings of the thesis across conceptual, methodological, and empirical dimensions. Chapter 8 develops an emergence-oriented account of Human–AI interaction, introducing Human–AI Collaboration Protocols as configurations through which intelligence, responsibility, and risk arise in practice, and examining AI-induced deskilling as an emergent structural concern. Chapter 9 introduces and applies a beyond-accuracy framework for evaluating Human–AI interaction, reporting empirical evidence on AI's impact on human judgement and analysing the behavioural risks associated with ex-

plainability, including the explainability paradoxes. Chapter 10 reports findings from the empirical programme conducted in radiology, combining controlled studies of frictional interaction designs with qualitative investigations of radiologists' preferences and workflow logic.

PART III: SYNTHESIS AND DISCUSSION. The final part integrates insights from all contributions in the *Discussion* (Chapter 11), showing how deliberately introduced cognitive friction—operationalised through interaction protocols and evaluative frameworks—constitutes a principled approach to responsible AI-supported decision-making in clinical practice. The *Conclusion* (Chapter 11) summarises the thesis contributions, outlines limitations and directions for future work.

Part I

BACKGROUND

2

CLINICAL DECISION-MAKING SUPPORTED BY AI

> *It is clear that technical problem solving is a radically incomplete description of what [...] physicians actually do. [...] When practitioners choose to address new or unique problems which do not fit known categories, their inquiry is not a threefold mapping [...] but a design process artistic in nature.*
> — Donald A. Schön, *The Reflective Practitioner* (Schön, 2017)

2.1 THE PROMISES OF MEDICAL AI

Clinical decision-making is an exemplary domain for studying the impact of AI on human decision-making. AI promises to revolutionise healthcare by enhancing diagnostic accuracy, speed, and coverage of medical shortages (Bienefeld, Keller, and Grote, 2025), especially in underserved rural areas (Guo and Li, 2018). Within these settings, AI decision support systems are typically introduced to improve efficiency and consistency, with claims that AI-supported radiology would be moving "from a subjective perceptual skill to a more objective science"(Pesapane, Codari, and Sardanelli, 2018, p. 5).

Radiology, in particular, has repeatedly been identified as a specialty at risk of partial or even full automation, with predictions of professional displacement accompanying successive advances in image analysis and machine learning (Langlotz, 2019). These claims draw on a broader view of professional work as amenable to formalisation and optimisation. Under a model of technical rationality, clinical practice is framed as a form of problem-solving in which physicians select and apply appropriate diagnostic and therapeutic techniques to maximise an objective function—typically accuracy or efficiency—within known constraints (Schön, 2017). From this perspective, medical judgement appears decomposable into large but ultimately enumerable sets of rules, heuristics, and inferential steps, rendering it a plausible target for computational replication.

Relatedly, influential accounts of technological disruption in the professions argue that many components of expert work are routine and process-based, and therefore do not require judgement, creativity, or empathy (Susskind and Susskind, 2016). Once professional tasks are disaggregated into constituent activities, such accounts suggest, they can increasingly be redistributed to less-specialised workers or automated systems, with high-performing technologies assuming a

growing share of decision-making responsibility.

Tacit knowledge in clinical work

However, clinical practice poses distinctive challenges for effective Human–AI interaction that are insufficiently captured by these optimisation-oriented framings. Medical work is uncertainty-laden and deeply dependent on tacit knowledge—the idea that practitioners, particularly experts, "know more than they can tell" (Polanyi, 1966, p. 4). Diagnostic reasoning often relies on experiential judgement, situational awareness, and the ability to integrate heterogeneous and partially articulated cues, rather than on the application of explicit rules alone.

Ethnomethodological and CSCW accounts of clinical practice in radiology, pathology, and endoscopy, show that interpretation hinges on cues that are social, experiential and contextual, and on invisible work (Star and Strauss, 1999) that prepares, translates, and stabilises evidence.

Radiology-specific issues

Anichini, Natali, and Cabitza (2024) delineated three radiology-specific dimensions that make AI support in this areas particularly tricky: (i) the *undatafiable* layer of contextual and experiential cues that resist neat labelling; (ii) *mediated communication* via written reports, where omission, emphasis, and hedge language perform accountability and reputation work; and (iii) the risk of entrenching *supernormality*, i.e., reifying narrow normal/abnormal templates that over-standardise perception.

For example, the act of writing a radiological report is not a direct textual translation of the information conveyed by imaging data: reports articulate what is *significant* for intended clinical action. This includes strategically downplaying harmless anomalies, staging information to guide the referrer, and anticipating the consequences of one's wording. Such selection is expertise-laden and socially situated: reporting is shaped by daily, individual and collective refinement of judgement—a practice that anomaly-detection tools can disrupt by shifting what becomes salient and how it is framed (Anichini, Natali, and Cabitza, 2024).

Detection pipelines frequently encode criteria of "normality" that diverge from everyday clinical sense-making. Post-operative changes, devices, or therapy effects are recurrently mislabelled as pathological—e.g., scar tissue in mammography flagged with high malignancy scores—inviting alarm fatigue and automation bias over time (Anichini, Natali, and Cabitza, 2024). This reflects *super-normality* (Beaulieu, 2001): a data-driven conception of the normal body that sidelines common, non-pathological alterations, thereby narrowing the field of permissible interpretation (Anichini, Natali, and Cabitza, 2024).

the AI chasm between promise and reality

It is then unsurprising that an *AI chasm* persists between techno-optimistic promises and clinical reality: Keane and Topol (2018) point

to "the gulf between developing a scientifically sound algorithm and its use in any meaningful real-world applications", which underlies a clash of cultures between computer scientists and clinicians (McCradden et al., 2022). Borrowing an expression by Suchman (2007), in their practice radiologists have a "presence to the unfolding situation of interaction not available to the machine" (Suchman, 2007, p. 12), including the tacit knowledge that guides their diagnostic work (Collins, 2005). Often, it is exactly this contextual and sensorial information–such as tactile examination–that guides diagnosis.

2.2 RECONFIGURING WORK, RESPONSIBILITY, AND SKILL

integration of AI reconfigures workflows

The integration of AI reconfigures workflows (*cfr.* the concept of *algorithmic refraction* by Christin 2020), authority, and even responsibility in unintended ways. Radiological AI decision support provides diagnostic advice based on deep learning algorithms, which are inherently opaque, while clinicians remain legally and morally accountable. This asymmetry introduces new challenges around liability, control, and confidence, placing clinicians in a position where responsibility is retained even as decisional authority is partially displaced (Graziani et al., 2023). In such configurations, practitioners risk becoming caught in a *moral crumple zone*, absorbing the consequences of system failures without having meaningful control over their operation (Hao, 2022).

Jussupow et al. (2021) suggest that this web of complexity and consequence in medical decisions is the reason why AI is unlikely to substitute clinicians in diagnostic work. The consensus among both physicians and developers is that the function of AI in clinical workflows is that of supporting medical experts drawing their own informed conclusion, noting that "it is difficult, yet critical, that physicians supervise CAID *[computer-aided intelligent diagnosis]* systems and do not follow AI advice without scrutiny" (Jussupow et al., 2021, p. 3). This supervisory ideal, however, is fragile in practice.

Ironies of automation and deskilling

When clinicians are repositioned primarily as auditors of AI output, their ability to intervene effectively when systems err may gradually erode. This dynamic exemplifies a classic *irony of automation*: as automation increases, opportunities to exercise and maintain critical skills diminish, precisely when those skills are most needed in cases of system failure (Bainbridge, 1983). Over time, reduced engagement in diagnostic reasoning can weaken clinicians' capacity to detect anomalies, challenge recommendations, or reconstruct the rationale behind a decision.

These effects are not evenly distributed across career stages. Trainees and residents, in particular, express concern that automated classification systems shape what they attend to, how they justify findings, and

when they commit to a diagnosis, potentially constraining learning before experiential knowledge has fully consolidated (Anichini, Natali, and Cabitza, 2024, p. 1014–1015). The institutional risk is a pipeline of complacent practitioners, optimised for tool-following rather than diagnostic reasoning. As Simone and Schmidt (1993) warned, what is needed instead are tools that remain heuristic and open to interpretation, enabling situated judgement and, at times, intelligent improvisation.

Recent work on skill formation reinforces these concerns. Beane (2024) argue that mastery depends on a combination of challenge, complexity, and trust-based relationships between experts and novices. When intelligent technologies are deployed in ways that remove challenge or simplify decision-making prematurely, they risk subtly degrading human capability rather than supporting its development. In the pursuit of productivity optimisation, organisations may inadvertently disrupt the very conditions required for skill acquisition, hollowing out expertise and reducing adaptive capacity.

Over-reliance on decision support can erode professional judgement: when AI offers the path of least resistance, clinicians risk loss of skill (Natali et al., 2025) and even cognitive atrophy (Carr, 2015). Such dependence leaves organisations exposed when support is unavailable or fails (Hernández-Orallo and Vold, 2019). By contrast, interaction designs that cultivate scrutiny and sustain professional judgement move the socio-technical system toward *antifragility*, improving under stressors rather than degrading (Taleb, 2012). The institutional stake is clear: without intentional preservation and training of human skill, clinical infrastructures become fragile—procedurally efficient yet epistemically brittle.

2.3 DETERMINANTS OF HUMAN BIASES IN AI-SUPPORTED DECISION-MAKING

The phenomenon of over-reliance on AI decision aids in radiology is well-documented (Li and Little, 2023), while lower-expertise decision-makers are more susceptible to accepting an erroneous AI suggestion (Gaube et al., 2021). Because of these complex reliance dynamics, decontextualised human–machine performance comparisons—abstracted from radiologists' multi-layered interpretive work—offer little purchase on how advice will actually be taken up, trusted, or resisted (Anichini, Natali, and Cabitza, 2024). Plass et al. (2022) argue that AI systems should be evaluated by their alignment with human diagnostic practices, instead of technical metrics only.

A key driver of over-reliance is our tendency toward cognitive miserliness (Fiske and Taylor, 2020). Under time pressure and workload, people select the lowest-effort path and allow technology to steer the course (Fiske and Taylor, 2020; Grissinger, 2019, p. 321). Mitigation

therefore has two fronts: improve the reliability of the tools *and* help clinicians more accurately appraise that reliability so they can apply appropriate monitoring and verification strategies (Grissinger, 2019, p. 321). This is especially salient amid rapid model churn: short AI lifecycles and frequent re-releases (sometimes with marginal performance gains) outpace users' ability to build calibrated trust (Graziani et al., 2023) ensure human oversight – which, according to Art. 14 of the EU AI Act, *shall aim to prevent or minimise the risks to health, safety or fundamental rights that may emerge when a high-risk AI system is used* (European Parliament and the Council of the European Union, 2024).

In a study of sentiment analysis and question-answering tasks, Bansal et al. (2021) have shown that human–AI "teams" often do not achieve the hoped-for complementary performance; instead, a human with AI can under-perform the AI alone if the human routinely defaults to the AI's judgment even in those instances where the AI is mistaken and the human could have been correct.

Automation bias

This is the widely documented phenomenon of automation bias (Mosier and Skitka, 1999; Skitka, Mosier, and Burdick, 1999), or "automation-included complacency" (Lyell and Coiera, 2017), which is codified in international standards as the "type of human cognitive bias due to over-reliance on the recommendations of an AI system"(ISO/IEC, 2021). Automation bias has been documented in domains from aviation to medicine: when an automated system is present, people tend to discount their own judgment even if the automation is wrong, leading to errors they might not have made working alone. In medical diagnosis, for example, physicians assisted by AI may fail to double-check a flawed algorithmic assessment, resulting in misdiagnosis. In its more passive form, this becomes *algorithmic loafing*—accepting machine output without exerting any active verification efforts (Inuwa-Dutse et al., 2023). For example, "People may reduce their effort or shed responsibility while carrying out a task if an automated system performs the same function." (Grissinger, 2019, p. 321)

Algorithmic aversion

Evidence suggests a systematic asymmetry: while automation bias is particularly common among novices and less-experienced users (Lebiere et al., 2021), whereas experienced users more often exhibit algorithmic aversion (Dietvorst, Simmons, and Massey, 2014) or, in Human-first decision-making settings, conservatism bias—a form of self-anchoring and reluctance to update one's initial decision due to undue scepticism of AI's correcting advice (Cabitza et al., 2023e).

Several well-established cognitive mechanisms help explain these effects–over a hundred of which have been observed to affect clinical decision-making (Croskerry, 2013). People are *cognitive misers* (Fiske

and Taylor, 2020) who often prefer minimal-effort strategies; heuristic reasoning supplies quick, automatic responses without effortful processing (Tversky and Kahneman, 1974). Dual-Process Theory distinguishes fast, intuitive System 1 thinking from slow, analytic System 2 reflection (Kahneman, 2013). Chiriatti et al. (2024) hypothesised that the outsourcing of certain cognitive tasks to AI coils entails a sort of System 0. In practice, *cognitive offloading* shifts internal processing to external artefacts, including AI systems (Diaz Alfaro, Fiore, and Oden, 2024; Roth et al., 2022). Increased deference towards "intelligent" systems can be motivated by preoccupations regarding the bounded rationality of humans (Dellermann et al., 2019) and the "noise" in their judgment (Kahneman, 2013), leading to "algorithm appreciation", when "people consistently give more weight to equivalent advice when it is labeled as coming from an algorithmic versus human source" (Logg, Minson, and Moore, 2019). Users may also overestimate technological capability, relax vigilance, and shed responsibility—*automation complacency*—especially after repeated exposure to seemingly correct outputs (Grissinger, 2019; Inuwa-Dutse et al., 2023). Related biases in AI-assisted judgement include selective adherence, priming, framing, and anchoring (Alon-Barkat and Busuioc, 2023; Rastogi et al., 2022; Riva, Aureli, and Silvestrini, 2022).

Importantly, bias in human–AI interaction does not stem solely from "technical flaws" or "human errors"; it emerges dynamically from the interplay of human psychology and AI behaviour (Glickman and Sharot, 2025; Hinduja et al., 2025). For example, process timing matters. If AI advice frames the case at the outset, decision makers may experience "AI-based confirmation," failing to conduct an independent assessment and overlooking contradictions: "the AI advice can influence physicians at the beginning of their decision making process. Thus, they may fail to conduct their own assessment of the data independently." (Jussupow et al., 2021). The problem, then, is not whether AI should assist, but *how* to structure that assistance so that human reasoning remains engaged.

3

DESIGNING FOR EMERGENCE IN HUMAN-AI INTERACTION

> *All new technologies develop within the background of a tacit understanding of human nature and human work. The use of technology in turn leads to fundamental challenges in what we do, and ultimately in what it is to be human. We encounter the deep questions of design when we recognize that in designing tools we are designing ways of being.*
>
> — Terry Winograd and Fernando Flores, *Understanding Computers and Cognition* (Winograd and Flores, 1986)

3.1 EMERGENCE ACROSS CSCW, DISTRIBUTED COGNITION, AND HYBRID INTELLIGENCE

The idea that complex wholes exhibit properties that cannot be reduced to their parts has a long intellectual genealogy. In nineteenth-century biology, disputes over *vitalism* and the nature of "the living" fostered a strand of holistic thinking in which the crucial explanatory factor was not a non-material essence but the *degree of organisation* of an organism. In the early twentieth century, this sensibility was sharpened by so-called organismic biologists, for whom the organism—rather than isolated mechanisms—was the unit of analysis, as exemplified by Woodger's *Biological Principles* (1929) (Checkland, 1999; Woodger, 2014).

It is in this context that Ludwig von Bertalanffy advanced the founding ideas of General Systems Theory, arguing from the late 1940s onwards that the conceptual apparatus developed for organisms could be extended to *complex wholes of any kind*—that is, to *systems* (Bertalanffy, 1968).

A central move in this tradition is to define a system not merely as an aggregation of components, but as a *whole* whose unity is analytically warranted by its properties. As Checkland (1999) puts it, for an observer to treat some complex entity as a whole separable from its environment, it must display properties that are attributable to it *as a single entity*; these are its *emergent properties*, the properties that make the whole "more than the sum of its parts" (Checkland, 1999, p. 50).

When interactive computing emerged as a practical and theoretical project, similar questions were reformulated in terms of *human–machine coupling*. As early as 1960, J. C. R. Licklider's vision of "man–computer symbiosis" framed computing as an active partnership between person and machine aimed at "tight coupling" of complementary capaci-

ties (Licklider, 2008). Importantly, this displaces the design challenge: rather than improving computers in isolation, one must engineer the *relation* between human cognition and computational artefacts. Rheingold's reading of Licklider makes the point explicit: the key problems are only partly about better brains or better machines, and primarily about the way they are coupled (Rheingold and Toms, 1991, p. 141). On this view, effective intelligence is not reducible to either component; it arises in the dynamics of coordination—timing, representation, interaction, and feedback.

This relational framing continues in Engelbart's augmentation programme. As Rheingold and Toms (1991, p. 184) notes, Engelbart's aim was not simply automation, but a transformation of the overall cognitive system: computers as instruments for thinking "in a more effective, wider-ranging, more articulate" way. Engelbart's own formulation is design-theoretically significant: the intellectual worker must know the capabilities of their tools and develop methods and heuristics for using them; process capabilities are distributed across person and artefact (Engelbart, 1963). These early accounts already point toward a non-reductionist view in which performance and responsibility depend on how capacities are composed in practice – with implications for work practices, distributed cognition, and human–AI complementarity.

EMERGENCE IN CSCW Traditional human-factors approaches attempted to pre-define tasks and constrain deviation, yet evidence from real-world use repeatedly showed that successful technologies are frequently those that users can appropriate and adapt beyond their original intent. In this context, Alter (2010) argued that HCI should explicitly investigate "designing for emergence", that is, recognising and leveraging emergent behaviours in socio-technical systems. A related response is the idea of *meta-design*: designing not only artefacts, but conditions under which users can extend, reconfigure, and continue the design work (Fischer, 2003; Fischer and Herrmann, 2011).

In parallel, research on appropriation examined how to support unanticipated uses without treating them as failures. Dix captures the relevant stance succinctly: "You may not be able to design for the unexpected, but you can design to allow the unexpected" (Dix, 2007, p. 1). The practical guidelines proposed in this line of work (*allow interpretations*, *provide visibility*, *expose intentions*, *support not control*, *plugability and configuration*, *encourage sharing*, *learn from appropriation*) are united by an ethic of openness: designers cannot foresee all contexts of use and must therefore enable interpretive flexibility. As Dix emphasises, the thread uniting these guidelines is a form of humility—acknowledging that one does not fully understand what will happen in real use, regardless of how thorough a user-centred process

appears (Dix, 2007, p. 3). In an emergence-oriented view, "deviations" can be adaptive responses and sources of innovation.

CSCW theory provides some of the clearest arguments against a control-oriented design imaginary. In Suchman (1987), Suchman shows that human action cannot be fully pre-scripted because people continually adapt to local contingencies; plans do not determine action so much as serve as resources that are interpreted in situ. Building on situated perspectives, Dourish's account of embodied interaction shifts the analytical focus from interfaces to interaction as a practical accomplishment (Dourish, 2001). The effectiveness of a system cannot be assessed solely by its formal specification; instead, it requires understanding how it becomes "part and parcel of a set of working practices" in real settings (Dourish, 2001, p. 62). Dourish's distinction between *work processes* (formalised procedures) and *work practices* (the informal, routine mechanisms through which procedures are actually enacted) makes the point concrete: "getting things done" involves approximation, improvisation, and ad hoc repair in response to everyday contingencies (Dourish, 2001, p. 63). Consequently, the duality of process and practice is inevitable; practice is dynamic and mediates between process descriptions and the circumstances of enactment (Dourish, 2001, p. 62).

In this account, interaction is "intimately connected with the settings in which it occurs", and computation must be analysed in terms of how it fits with those settings (Dourish, 2001, p. 19). In short, meaning is not located in outputs alone; it is produced through engaged interaction with artefacts in physical and social worlds (Dourish, 2001). Relatedly, Nonaka's notion of *Ba* highlights a shared space in which knowledge and interpretation take shape through emerging relationships (Nonaka and Konno, 1998); and Harrison and Dourish's distinction between *space* and *place* underscores how settings become meaningful through social interaction and cultural practice (Harrison and Dourish, 1996). These perspectives treat cognition and coordination as inherently relational: intelligent action is accomplished through tasks-in-interaction rather than isolated individual acts (Luff, Heath, and Greatbatch, 1992).

This locates emergence in the ordinary accomplishment of work: socio-technical effects arise through situated coordination among people, artefacts, and organisational arrangements, often in ways that are locally rational but not fully specifiable in advance.

DISTRIBUTED COGNITION AND HYBRID INTELLIGENCE In classic CSCW accounts, information technologies mediate and transform work by redistributing attention, memory, accountability, and decision-rights across people, documents, routines, and devices (Berg, 1999,

p. 376). From this perspective, intelligent action is an achievement of socio-technical assemblies whose properties cannot be reduced to any single component (Berg, 1999).

The "origin myth" of HCI frames the field as emerging, in part, from the interplay between cognitive psychology and computer science; from its very beginnings, HCI was concerned with translating selected aspects of human cognition and activity into system design (Dourish, 2001, p. 61).

The field of cognitive science, through theories of extended and distributed cognition, can formalise why intelligence is rarely "inside" a single head—or a single model. The *Extended Cognition* thesis argues that tools and environments can function as components of cognitive systems (Clark and Chalmers, 1998). In Clark and Chalmers' celebrated example of an Alzheimer patient relying on pen and paper to support his waning memory, the notebook he carried is not an external aid but part of his memory—integral to the cognitive loop that produces action. In high-stakes work, AI diagnostics, checklists, triage dashboards, and case libraries can play analogous roles as elements of the coupled system through which clinicians perceive, recall, and decide.

Distributed Cognition sharpens the point: cognitive processes are spread across people, representations, and artefacts that encode state, afford operations, and constrain moves (Hutchins, 2000). A radiology team, for instance, "thinks" through worklists, annotations, image viewers, and consultations. Likewise, in AI-supported workflows, the system's apparent "intelligence" lies not in the model alone, but in *how* human judgement and machine outputs are coordinated over time. What counts as intelligent behaviour therefore depends on the routines, tools, and locally meaningful norms through which it is enacted.

Hybrid Intelligence (HI) can be read as a contemporary consolidation of these emergence-oriented perspectives. The research agenda of the *Hybrid Intelligence Centre* characterises HI as the combination of human and machine intelligence aimed at augmenting human capabilities rather than replacing them, in order to achieve goals unreachable by either alone (Akata et al., 2020, p. 20). Another formulation stresses not only superior joint performance, but also continuous mutual improvement through learning from one another (Dellermann et al., 2019, p. 640).

Across the literature, HI systems are commonly described as mixed human–AI teams in which complementary strengths are orchestrated to achieve shared goals (Hemmer et al., 2021; Rafner et al., 2022; Tiddi et al., 2023). At the same time, recent reviews acknowledge that HI is

not a single, settled concept: it is alternately framed as an emergent property of interaction, a human-/AI-in-the-loop decision-making arrangement, a form of collective intelligence, or a broader design paradigm (Dell'Anna et al., 2024).

Despite this plurality, a consistent conceptual lineage is visible. HI explicist inherits Engelbart's augmentationist stance, with the definition by Akata et al. (2020) indivduating as the goal of AI "*augmenting human intellect* and capabilities instead of replacing them" (Akata et al., 2020, p. 20). As Rheingold observes, Engelbart's concern was with the new forms of thinking that such artefacts enable—"more effective, wider-ranging, more articulate, quicker, better-informed" cognition—hence his deliberate preference for *augmentation* over *automation* (Rheingold and Toms, 1991, p. 184). Engelbart's own formulation reinforces this systemic view: intellectual work depends on knowing the capabilities of one's tools and developing methods and heuristics for using them, with process capabilities distributed across person and artefact (Engelbart, 1963).

BRIDGING DISCIPLINES Both distributed cognition and Hybrid Intelligence converge on a shared insight: the relevant unit of analysis is neither the human nor the machine in isolation, but the relational organisation of work—how coupling, coordination, and representation allow joint performance to emerge. This continuity motivates a design stance aligned with CSCW and systems theory: rather than optimising individual components independently, one must attend to the conditions under which human and machine capacities combine productively in situated practice. In high-stakes domains, this entails recognising that effectiveness depends not only on algorithmic precision, but also on interaction quality—how information is shared, scrutinised, revised, and made accountable in context. As Suchman succinctly reminds us, "there is no such thing as a machine that acts outside of relations with humans" (Suchman and Thimm, 2024, p. 28). Defining and shaping those relations has long been a central concern of Human–Computer Interaction, articulated through a variety of terms and frameworks, each foregrounding different implications for design and responsibility (Breckner et al., 2025).

3.2 THE VOCABULARY OF AI-SUPPORTED DECISION-MAKING

The terminology used to describe sustained human–AI work shapes both expectations and design. Labels such as *teaming*, *partnership*, and *collaboration* can import assumptions about shared goals, symmetry, and oversight. The *Ethics Guidelines for Trustworthy AI* published in 2019 by the High-Level Expert Group on Artificial Intelligence (Smuha, 2019) delineate three complementary governance mechanisms to se-

cure appropriate human oversight: *human-in-the-loop* (HITL), *human-on-the-loop* (HOTL), and *human-in-command* (HIC).

Who is on the loop?

The HOTL approach refers to the capacity for human monitoring and potential intervention throughout a system's operational or design cycle. The HIC approach, by contrast, emphasises the user's authority to supervise the AI system's activities and overall impact, while retaining ultimate control over whether and how the system is deployed in a given context. Finally, the HITL model envisions a form of human oversight in which users are both permitted and encouraged to intervene during each decision cycle of the system, including, where appropriate, the adjustment of algorithmic or learning parameters themselves (Meza Martínez, Nadj, and Maedche, 2019). Yet, as noted by the Council of Europe (The Council of Europe's Ad Hoc Committee on AI, 2020), such continuous human intervention is neither feasible nor desirable in many real-world scenarios. Within the broader discourse on hybrid intelligence, "human-in-the-loop" configurations are understood as socio-technical systems capable of processing highly unstructured information (Wiethof and Bittner, 2021; Xu et al., 2021). Such configurations can achieve levels of performance unattainable by humans or machines operating independently (Holzinger, 2016). For instance, when a human operator identifies low confidence in a system's output, they can intervene to adjust or refine the input, creating a feedback loop that enhances the overall performance of the combined system. The central aim of this paradigm is to address complex decision tasks through the strategic distribution of subtasks between AI-based components and human agents in a complementary and adaptive manner (Xu et al., 2021). Nevertheless, the "human-in-the-loop" model often carries an implicit assumption that computers will perform the vast majority of work, with humans merely remaining available to intervene should an unexpected event occur. In such arrangements, the human's role risks being reduced to that of a passive or alienated auditor of the system's proper functioning. This over-reliance on automation can erode vigilance and situational awareness, ultimately leading to what Parasuraman and Manzey (2010) termed *automation complacency*.

Shneiderman (2020) describes the "Second Copernican Revolution" of Human-Centred AI (HCAI) as a new reframing of the traditional discussion on humans being "in the loop" of AI systems. In this view, HCAI instead places AI "in the loop" around humans, who remain the central locus of attention and control. Reversing the conventional approach, he calls for a *Computer-in-the-Loop* paradigm, in which "*humans work with others in teams, crews, and groups, with computers best designed as helpful tools that continuously provide information and carry out tasks, but do so under human control*" (Shneiderman, 2020, p. 113).

This human-centred inversion is echoed by Malone (2018), who proposes the complementary *Computer-in-the-Group* framing: "*Many people assume that computers will eventually do most things by themselves and that we should put 'humans in the loop' in situations where they're still needed. But I think it's more useful to realise that most things now are done by groups of people, and we should put computers into these groups in situations where that's helpful. In other words, we should move from thinking about putting humans in the loop to putting computers in the group.*" From a Human-AI Interaction perspective, "*computers should play a supportive role, amplifying people's ability to work in masterful or extraordinary ways*" (Shneiderman, 2020). In light of the importance of group dynamics in decision-making, Malone (2018) suggests that "*we should move away from thinking about putting humans in the loop to putting computers in the group.*" Shneiderman further refines this shift into what he proposes as a guiding maxim for Human-AI Interaction: "*Humans in the group; computers in the loop.*"

Against dyadic AI

Both framings challenge the dyadic Human-AI Interaction paradigm, in which humans and AI systems are conceived as symmetrical, interacting agents (Cabitza, Campagner, and Simone, 2021; Shneiderman, 2020). Winograd and Flores, 1986 "In uttering a sentence containing mental terms ('intelligent', 'perceive', 'learn'), we are adopting an orientation towards the thing referred to by the subject of the sentence as an autonomous agent. ... in using mental terms we commit ourselves to an orientation towards it as an autonomous agent" Such dyadic models fail to account for the relational, cooperative, and tacit dimensions of collective human decision-making within which AI systems are typically deployed. As Cabitza, Campagner, and Simone (2021) argue, these sociotechnical contexts are better understood as distributed networks of human expertise, coordination, and sense-making—phenomena that cannot be reduced to a single human–machine dyad.

Table 3.1: Framing terms in HAI: implied assumptions and critiques

Term	Assumption	Critique by Sources
Collaboration	Assumes common goals, reciprocal action, and often symmetric roles and contribution.	Used promiscuously (a "jingle fallacy" (Breckner et al., 2025)) across everything from decision support to co-creation. Can mask the human labour behind model building (human–human collaboration distanced and disguised) and imply reciprocity rarely present in DSS. See also the "agentistic turn" critique. (Sarkar, 2023a)
Teaming / Partnership	Confers personhood/agency to software; suggests social engagement, shared goals, and intention.	Agentic metaphors encourage users to attribute beliefs and intentions to systems (CASA effect Nass et al. (1993)), inviting expectations of benevolence/emotion and blurring accountability (e.g., accountability drift / "moral crumple zone"). Current AI lacks sensorimotor grounding and genuine intention; the metaphor can misplace credit/blame and obscure the tool nature of AI. (Nass et al., 1993; Sarkar, 2023a)

Table 3.3: Framing terms in HAI: implied assumptions and critiques

Term	Assumption	Critique by Sources
Cooperation	From CSCW: interdependent tasks, possibly with differing goals; in HAI, often implies allocating subtasks to the better-suited agent.	In classic CSCW, "cooperative work" concerns articulation work *between humans* supported by computers—conceptually distinct from human–computer "working together". Using it for AI may smuggle in undue assumptions of shared sentiments or aligned values that do not obtain in practice.
Symbiosis	Highly integrated configuration leveraging complementary strengths; often framed as *supporting* rather than replacing humans (man–computer symbiosis).	Useful for emphasising complementarity, but can overstate symmetry. A tightly coupled rhetoric may distract from the system's instrumentality and from seams/limits users should see.

Table 3.5: Conceptual framings of Human–AI relations: implicit assumptions, loci of intelligence, and critiques

Term	Assumption	Critiques
Collaboration *Assumes:* common goals, reciprocal engagement, and symmetry of roles.	Frames human–AI work as joint problem solving toward shared objectives (Seeber et al., 2020). *Locus:* distributed between human and AI as if sharing a common goal space.	Used broadly and ambiguously—from decision support to deep co-creation—leading to jingle fallacies (Sarkar, 2023a). Risks obscuring the fundamentally asymmetric relation between human agency and computational affordance, and masking human labour or accountability.
Teaming / Teammate *Assumes:* that AI systems can act as social peers with shared situational awareness and emotional reciprocity.	Assigns personhood and social cognition to machines, suggesting joint intentionality and trust dynamics. *Locus:* imagined as a socially distributed cognitive system.	Represents what Shneiderman calls the *"Teammate Fallacy"*—the belief that computers should be designed to function in teams simply because humans do (Shneiderman, 2020). Anthropomorphises AI, creating false expectations of empathy, accountability, or shared purpose that systems cannot fulfil.

Term	Assumption	Critiques
Cooperation *Assumes:* interdependent work toward possibly distinct goals.	Historically in CSCW, cooperation denotes coordination **between humans** through articulation work supported by computers (Schmitt et al., 2021). *Locus:* human; the computer mediates but does not co-act.	Using the term for AI conflates mediation with social cooperation. Misattributes intentional, normative, and communicative capacities to computational artefacts, blurring the distinction between tool and collaborator.
Partnership / Symbiosis *Assumes:* tight coupling and mutual benefit from complementary strengths.	Portrays human–machine systems as synergistic, co-evolving entities, echoing "man–machine symbiosis" (Licklider, 2008) and later notions of "symbiotic AI" (**wang2019symbiotic**; Ramchurn, Stein, and Jennings, 2021). *Locus:* hybrid organismic or sociotechnical system.	Although useful for discussing complementarity and co-adaptation, it risks romanticising interdependence and downplaying governance and human oversight. May suggest equality of roles that elides accountability or amplifies the illusion of mutual agency.

Term	Assumption	Critiques
Hybrid / Situated Intelligence *Assumes:* complementarity without symmetry; intelligence emerges from interaction embedded in context.	Combines the heterogeneous capabilities of humans and machines to achieve goals unattainable by either alone (Akata et al., 2020; Dellermann et al., 2019). Situated Interaction emphasises that meaning and performance depend on context, embodiment, and real-world practice (Dourish, 2001; Suchman, 1987). *Locus:* distributed across human, machine, and environment.	Reframes intelligence as emergent from situated action rather than as a property of either agent. Avoids the anthropomorphic drift of "team" or "partnership" metaphors, restoring the distinction between human responsibility and machine capability, while acknowledging their entanglement in practice (Cabitza, Campagner, and Simone, 2021).

In CSCW, *cooperation* refers to coordination *between humans*, supported by computers through articulation work Schmidt and Bannon, 1992. Studying *humans working together* is analytically distinct from positing *humans and computers working together as co-operators*. Preserving this distinction keeps attention on tool-mediated human practice rather than imputing social roles to artefacts.

Across Human–AI work, *collaboration* is typically defined as a setting with cooperation toward common goals and complex problem solving Ala-Luopa et al., 2024, or as sharing "the same high-level objectives, common tasks, [and] dependence" Müller, Vette, and Geenen, 2017. Human-Machine Teaming is often presented as an umbrella for human–agent/AI/automation/autonomy/computer/robot interaction (Chen and Barnes, 2014; Lyons et al., 2021; Sheridan and Parasuraman, 2005), targeting augmentation of human and machine capabilities in pursuit of shared team goals (Verhagen et al., 2025). Yet Shneiderman warns against the "Teammate Fallacy"—the assumption that computers should function in teams merely because humans do Shneiderman, 2020. This cautions against uncritically projecting human team properties onto computational systems. *Partnership* has been framed as a socio-technical system maximising human–machine synergies, with AI engaging actively with humans Ramchurn, Stein, and Jennings, 2021. Syntheses characterise Human–AI partnerships as configurations that draw on complementary strengths and sometimes even cast

humans and AI as "fellow team members who can both reactively and proactively collaborate," with discussions of roles, involvement, acceptance and reliance, and design implications Breckner et al., 2025. While informative, this rhetoric risks sliding toward anthropomorphism. Terms like *partner*, *teammate*, *collaborator*, and *co-operator* can prematurely ascribe human attributes to algorithms. Sarkar argues that even "collaboration" may be too anthropomorphising in many contexts Sarkar, 2023a. A restrained vocabulary helps maintain clarity about capabilities, accountability, and limits.

Hybrid intelligence and agency

This motivates a choice for Hybrid Intelligence (Dellermann et al., 2019) as term that emphasises how contributions are brought together without implying personhood or symmetry. For example, *hybrid collective intentionality* show how non-human components can contribute to a collective's cognitive processing without being full agents; even a simple sensor can play a role analogous to a human observer in the flow of information (Brouwer, Ferrario, and Porello, 2021). When an AI system is successfully appropriated—enhancing a user's epistemic capabilities toward their goal—a *hybrid agent* can be said to emerge, shifting some discussions (e.g., of trust) to the hybrid system as the appropriate subject (Ferrario, Facchini, and Termine, 2024). This keeps design and evaluation centred on context, protocols, complementarity, and accountability.

3.3 HUMAN–AI COLLABORATION PROTOCOLS

Emergence-oriented design

The design of emergence-oriented human-AI systems goes beyond crafting efficient interfaces for accessing medical data or implementing precise diagnostic algorithms. It involves creating environments that foster hybrid intelligence through dynamic interactions between human expertise and artificial intelligence within a shared professional space. Concretely, this means designing adaptable systems that respond to varying clinical scenarios, support different forms of human–AI collaboration, and uphold human agency while harnessing technological capabilities. These design principles suggest a shift from "component-centric design" (Carayannis and Coleman, 2005) to what can be termed "emergence-oriented design": an approach focused on creating conditions that enable the development of effective hybrid intelligence systems (Cabitza, Locoro, and Ravarini, 2020; Cabitza et al., 2025c). Rather than optimising isolated parts, designers specify how people and systems co-produce decisions in context.

A complementary HCI strand treats collaboration with intelligent systems as *joint activity*. Horvitz's mixed-initiative principles specify how initiative should shift as confidence, context, and costs change (Horvitz, 1999). Coactive design extends this to interdependence: humans and agents observe, anticipate, and adjust to one another over time (John-

son et al., 2014). What matters is less a once-and-for-all division of labour than the *protocols* that sustain collaboration: turn-taking, handovers, escalation, and deferral.

A central instrument for this is the *human–AI collaboration protocol*: the explicit rules that determine *how*, *when*, and *in what form* the AI participates in a decision task (Cabitza et al., 2023e). Such protocols stipulate "how human decision makers should interact with the machines that support them," for example, what data is made available to the AI tool; what data the AI should provide to users; at which step in an articulated process; and in what order relative to human work.

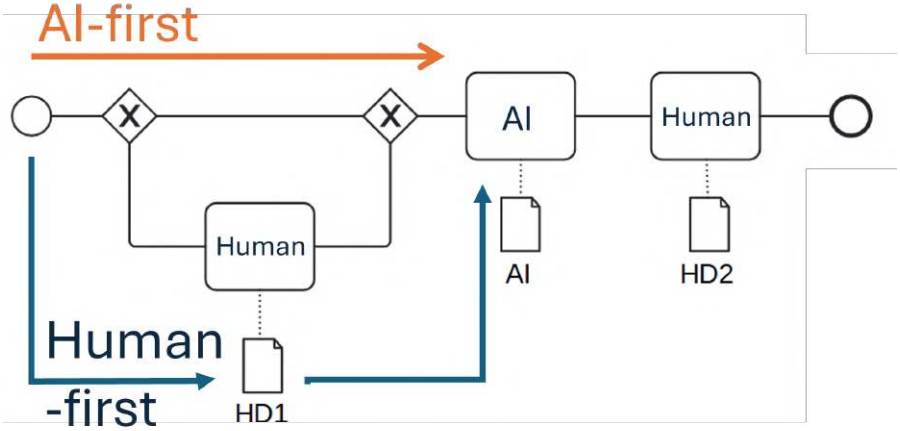


Figure 3.1: Example of a *Human-first* vs. *AI-first* Human-AI Interaction protocol, presented by Cabitza et al. (2023d). In the *AI-first* protocol, the user is first exposed to an AI response (*AI*) and after provides their own response (*HD*2). In the *Human-first* protocol, the user first provides a response without AI support (*HD*1), then they are exposed to the AI advice (*AI*), and finally provide their AI-supported final answer (*HD*2).

AI-first, Human-first protocols

In particular, Cabitza et al. (2023e) distinguishes between AI-First and Human-First interaction protocols. AI-first protocols are also defined as *concurrent* – "where AI advice is displayed concurrently with the prediction problem" (Tejeda et al., 2022). An example of AI-First interaction protocol in the radiological setting was described by Anichini, Natali, and Cabitza (2024): the X-ray image was opened on the screen at the same time as the pneumothorax detection system (AIT), whose classification of the image was immediately visible on the screen (green for "normal", red for "abnormal"), informing the radiologist of the algorithm's classification. An instance of a Human-First interaction protocol is that entailing the AI-assisted diabetic retinopathy detection task reported by (Bach et al., 2023), where ophtalmologists first had to pre-register their own unassisted decision. They termed this protocol 'hear the story first' and investigated its role as a debiasing strategy

against anchoring bias. Other terms for this protocol are update (Gajos and Mamykina, 2022; Green and Chen, 2019), sequential (Tejeda et al., 2022) or consequential (Buçinca, Malaya, and Gajos, 2021), proactive (Lai et al., 2022) or silent (Kwong et al., 2024).

Levels of Automation refer to the spectrum of how control and tasks can be divided between human and machine. Parasuraman, Sheridan, and Wickens (2000) provided a seminal model, defining four stages of functions in any human–machine system—(1) information acquisition, (2) information analysis, (3) decision and action selection, and (4) action implementation—each of which can be automated to varying degrees. A low level of automation means the human performs that function manually, whereas a high level means the AI handles it autonomously. The levels-of-automation concept highlights that automation is not all-or-nothing; designers must choose appropriate levels for each aspect of a task. Too much automation can leave users out of the loop, yet too little fails to capitalise on AI's strengths. The optimal balance often depends on context: e.g., in life-critical decisions, one might keep the human at a high level of control (low automation for decision selection) to preserve judgement, whereas for mundane repetitive tasks, high automation can reduce workload. Understanding levels of automation thus provides a vocabulary for designing adjustable autonomy—ensuring the human and AI contributions are complementary and that control can be shifted as appropriate.

Automation is not all-or-nothing

Recent empirical syntheses convert these traditions into actionable guidance. Amershi et al. (2019) emphasise making capabilities and limits legible, enabling users to summon or dismiss assistance, and maintaining an *ongoing dialogue* as conditions shift. Read together with levels of automation, these guidelines scaffold the stipulation of decision protocols: how capabilities are surfaced, how uncertainty is communicated and bounded, how people can contest or correct, and how systems evolve without destabilising practice.

Gomez et al. (2025) introduce the contrast between *advisory* and *shared control* methods of interaction. Many AI systems function in an advisory pattern: the AI makes recommendations or provides analyses, and the human user is free to accept, reject, or ignore that advice. This pattern is common in decision support because it clearly puts the human in the driver's seat (e.g., an AI system suggests a diagnosis, but the final call rests with the clinician). The advantage of a pure advisory role is clarity of responsibility and relative simplicity of implementation. However, as AI capabilities grow, more *intertwined* patterns appear in which control is dynamically shared: the system may pre-fill orders to be confirmed, adjust parameters under human-set constraints, or execute bounded actions that are reversible and logged for oversight. Such shared-control arrangements require ex-

plicit protocols for initiation, override, escalation, and audit, so that responsibility and situational awareness remain legible even as some actions are delegated.

Emergence-oriented, situated protocols articulate timing, representation, control, and escalation in ways that support complementary strengths without anthropomorphising the machine. They also give us something concrete to evaluate: not only whether an algorithm is accurate, but whether the *arrangement* by which humans and AI work together produces safer, more accountable, and more skilful decisions over time.

3.4 PROTOCOL EVALUATION AND RELIANCE PATTERNS

The efficacy of interaction protocols is mostly appraised through the lens of *appropriate reliance*: leveraging AI advice when correct and rejecting it when incorrect (Eckhardt et al., 2024; Guo et al., 2024; Lee and See, 2004; Schemmer et al., 2023; Schmitt et al., 2021).

A large metric family exists (Schaschek, Spatscheck, and Winkelmann, 2024), but many popular choices actually capture *agreement* rather than *influence*: Agreement Fraction/Percentage (Yin, Wortman Vaughan, and Wallach, 2019), Reliance Rate (Yu et al., 2019), Failure Detection (Merritt et al., 2015), and the appropriate/inappropriate reliance ratios (Brachman et al., 2022) are widely used (He, Kuiper, and Gadiraju, 2023; Wang and Yin, 2021).

Reliance patterns

It is the introduction of Human-first interaction protocols that allows to observe whether the alignment with the AI reflects or merely pre-existing concordance.

Schemmer et al. (2023) first characterise four reliance dimensions (hereafter referred to as reliance patterns), which then underpin two metrics: relative positive self-reliance and relative positive reliance. These respectively capture the degree to which decision-makers beneficially rely on their own judgement or on the AI system.

It is useful to introduce a decision table (see Table 3.6), as introduced by Cabitza et al. (2023a). This table, which builds directly on the reliance dimensions identified by Schemmer et al. (2023), complements their partial enumeration by including all possible combinations of the initial human judgment, the AI recommendation, and the final human decision. Each of these three may be correct (1) or incorrect (0) relative to a presumed reliable ground truth. Extending the work of Schemmer et al. (2023), which focused on the positive forms of reliance and non-reliance, we identify four additional reliance patterns: detrimental reliance, beneficial under-reliance, detrimental self-reliance, and beneficial over-reliance. Each pattern corresponds to distinct cognitive biases, as discussed by Cabitza et al. (2023a).

Table 3.6: Reliance patterns: decisional configurations between human decision makers and AI support. In the first three columns, 0 denotes an incorrect decision and 1 a correct decision.

Human judgement (HD1)	**AI support (AI)**	**Final decision (FHD)**	**Reliance pattern**	**Other formulations**	**Facilitating biases**
wrong (0)	wrong (0)	wrong (0)	detrimental reliance	–	automation complacency
wrong (0)	wrong (0)	right (1)	beneficial under-reliance	–	extreme algorithmic aversion
wrong (0)	right (1)	wrong (0)	detrimental self-reliance	detrimental overriding	conservatism bias
wrong (0)	right (1)	right (1)	beneficial over-reliance	correct adherence	algorithm appreciation
right (1)	wrong (0)	wrong (0)	detrimental over-reliance	detrimental adherence	automation bias
right (1)	wrong (0)	right (1)	beneficial self-reliance	corrective overriding	algorithmic aversion
right (1)	right (1)	wrong (0)	detrimental under-reliance	–	extreme algorithmic aversion
right (1)	right (1)	right (1)	beneficial reliance	–	confirmation bias (in later cases)

The introduction of decision tables helps us operationalise, measure and compare the phenomena of automation bias, conservatism bias and appropriate reliance.

AUTOMATION BIAS As anticipated in our previous overview of some human cognitive biases in human-AI interaction (Section 2.3), automation bias represents excessive reliance on AI advice even when incorrect, leading users to make errors they would not have made independently—so-called commission errors (Cabitza et al., 2023a). It occurs when individuals follow incorrect AI recommendations instead of exercising appropriate scepticism. Conceptually, automation bias reflects a failure of distrust—an inability to discard wrong AI advice when necessary. Empirically, automation bias manifests as a detrimental over-reliance pattern, where users repeatedly accept erroneous AI outputs. Its severity can be expressed through an odds ratio comparing the frequency of such over-reliant decisions to instances of beneficial self-reliance:

$$\frac{100}{N-100}\frac{N-101}{101} \tag{3.1}$$

On the left are the odds of users following incorrect AI advice; on the right, the inverse of odds of users rejecting incorrect AI advice compared to all other cases. When this ratio is close to 1, users are as likely to ignore as to follow wrong advice; when it is much larger than 1, automation bias is significant. This measure relates to the Relative Self-Reliance (RSR) index (Schemmer et al., 2023): automation bias equals zero only when RSR equals one, indicating perfect ability to reject incorrect AI suggestions.

CONSERVATISM BIAS. Conservatism bias is the mirror image of automation bias. It describes undue scepticism towards correct AI advice, leading to omission errors (Dietvorst, Simmons, and Massey, 2014). Here, users fail to adopt reliable AI recommendations, over-relying instead on their own flawed judgement. It represents a failure of trust—the inability to recognise when AI advice is right and should be accepted. In reliance-pattern terms, conservatism bias corresponds to detrimental self-reliance, i.e., situations where correct AI suggestions are dismissed. It is measured through an analogous odds ratio comparing harmful self-reliance with beneficial reliance on correct AI advice:

$$\frac{010}{N-010}\frac{N-011}{011} \tag{3.2}$$

On the left is the odds of users rejecting the correct AI advice; on the right is the inverse of odds of users following the correct AI advice.. This construct parallels the Relative AI Reliance (RAIR) index, with conservatism bias equal to zero only when RAIR equals one.

APPROPRIATE RELIANCE. Appropriate reliance integrates these two dimensions, capturing the user's ability to trust AI when it is correct and distrust it when it is wrong (Guo et al., 2024; Schemmer et al., 2023). It accounts for both the acceptance of accurate recommendations and the rejection of erroneous ones, including confidence adjustments when advice aligns with or contradicts the user's initial opinion.

Mathematically, it can be expressed as the proportion of beneficial trust and beneficial distrust patterns across all AI interactions. When the appropriate reliance score equals one, both automation bias and conservatism bias are null—indicating perfectly calibrated trust.

TECHNOLOGY IMPACT Technology Impact quantifies the practical usefulness of a decision support system in situ, capturing the extent to which AI support alters users' decision accuracy. Following Cabitza et al. (2023a), it is computed by comparing the error rate observed after receiving AI support (AIER) with the error rate in the absence of AI support (CER), expressed as an odds ratio:

$$\frac{CER}{1-CER}\frac{1-AIER}{AIER} \tag{3.3}$$

This formulation contrasts the odds of making an error when unaided with the odds of making a correct decision when supported, thereby isolating the net contribution of the AI intervention to decision outcomes.

Importantly, Technology Impact does not characterise how users rely on AI, nor whether such reliance is appropriate. Rather, it captures the aggregate behavioural effect of AI support on performance, independently of the underlying reliance patterns.

Therefore, assessing the technology impact complements measures of automation bias, conservatism bias, and appropriate reliance: taken together, these constructs allow one to distinguish between AI systems that are influential because they are genuinely helpful, and systems that appear influential due to over-reliance or persuasive effects that may undermine judgement.

In response to such risks, one prominent line of research has proposed *Explainable AI* as a means to support appropriate reliance by making AI behaviour more transparent and contestable. The following section therefore examines this promise critically, analysing the limits of explainability as a general remedy for miscalibrated trust in AI-supported decision-making.

4

PROMISES AND PERILS OF EXPLAINABLE AI

> *[. . .] yet thinking need not be reflective. For the person may not be sufficiently critical about the ideas that occur to him. He may jump at a conclusion without weighing the grounds on which it rests; he may forego or unduly shorten the act of hunting, inquiring; he may take the first 'answer' or solution, that comes to him because of mental sloth, torpor, impatience to get something settled. One can think reflectively only when one is willing to endure suspense and to undergo the trouble of searching.*
> — John Dewey, *How We Think* (Dewey, 1933)

4.1 A TYPOLOGY OF EXPLANATIONS

In high-stakes settings such as clinical diagnosis, inscrutable recommendations make integration of AI advice into human judgement fraught with ethical conundrums, complicating informed consent, audit, and redress (Langer et al., 2021).

The explainability thesis

The *explainability thesis* (Langer et al., 2021) asserts a moral duty to avoid basing the treatment of persons on opaque systems. In the words of Miller (2019, p. 23), *"Is it ethical to make important decisions about individuals without being able to explain these decisions?"*.

In their influential *Explainable Artificial Intelligence Program*, DARPA defined Explainable AI (XAI) as "models that, when combined with effective explanation techniques, enable end users to understand, appropriately trust, and effectively manage the emerging generation of AI systems" (Gunning and Aha, 2019, p. 40).

How to open the black box

The techniques to "open the black box" (Baselli, Codari, and Sardanelli, 2020) can be global (explaining the functioning of the whole AI system) or local (instance-based, case-by-case explanations of an AI output). Explanations can, at the baseline, consist of *white-box* or intrinsically interpretable models such as decision trees (Delibasic, Vukicevic, and Jovanovic, 2013) and rule-based systems (Loyola-Gonzalez, 2019). For *black box* models, their explanations can be expressed in various forms. In their review, Burkart and Huber (2021) distinguish between *graphics*, *textual description* and *multimedia* explanations. Examples of visual forms of explanations are saliency maps (Simonyan, Vedaldi, and Zisserman, 2013), activation maps (Selvaraju et al., 2017),

and feature-importance visualisations such as Interpretable Model-Agnostic Explanations (LIME) (Ribeiro, Singh, and Guestrin, 2016) and SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017). Natural language allows for great flexibility in *how* to explain (Cambria et al., 2023) and have been described as *human-friendly* (Morrison et al., 2024), allowing for interactive solutions based on dialogue (Jentzsch, Höhn, and Hochgeschwender, 2019). Other intuitive ways to present information include constrastive and counterfactual explanations (Keane and Smyth, 2020; Stepin et al., 2021; Wachter, Mittelstadt, and Russell, 2018).

Explanations are fallible meta-outputs

We can therefore see how AI explanations are often produced by post-hoc generators, which are often decoupled from the predictor. Since these models are not infallible, they introduce an unintended layer of uncertainty in the interaction (Morrison et al., 2024). This informs the pragmatic stance delineated by Cabitza et al., 2023c: explanations are treated as meta-outputs (Ghassemi, Oakden-Rayner, and Beam, 2021) designed to make advice more understandable, appropriable, and usable by intended decision-makers.

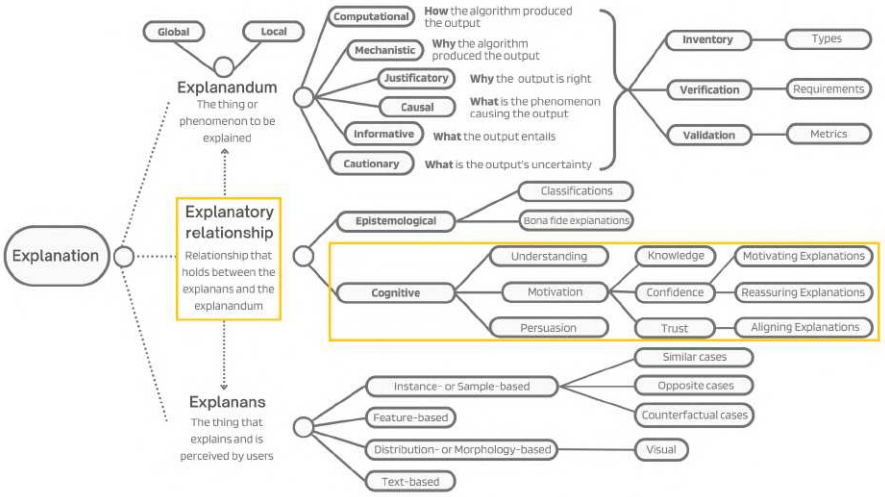


Figure 4.1: Typology of explanations introduced by Cabitza et al. (2023c), with emphasis on the *cognitive* dimension of the *explanatory relationship*.

A typology of explanations

To reason concretely about explainability, I use the typology of explanations that was introduced by Cabitza et al., 2023c and presented in Figure 4.1. This typology distinguishes three constitutive elements of explanations: the *explanandum* (what is to be explained), the *explanans* (the material offered as an explanation), and the *explanatory relationship* (how the *explanans* relates to the *explanandum*). This lens helps separate form (e.g., examples, saliency maps, rationales) from function (how the user takes it and what it changes). Crucially, the explanatory relationship may rest on epistemic implication (logic and evidence)

or on psychological implicature (how people infer meaning), and the two can diverge.

The role of the explanatory relationship

The explanatory relationship is the crucial node to classify and understand the quality and cognitive impacts of XAI. It is divided into an epistemologicaland a cognitive/psychological perspective.

The epistemological perspective evaluates the quality of the explanation on the basis of implication/logical consequence. The cognitive/psychological recognises that the impact of explanations in XAI-assisted decision support operates through several crucial cognitive pathways related to engaging the user's critical thought, affecting their emotional satisfaction and confidence, and ultimately influencing their decision alignment with the AI. Therefore this perspective evaluates explanations on the basis of its intended effects on human reasoning.

Reassuring, aligning, motivating explanations

Within the cognitive dimension of the explanatory relationship we distinguish *reassuring*, *aligning*, and *motivating* explanations.

Reassuring explanations raise confidence and satisfaction, optimising the perceived understanding and acceptance of the AI output. However, this can feed into illusions of explanatory depth (Chromik et al., 2021). Their effectiveness–which is typically evaluated using psychometric scales focusing on self-confidence, satisfaction with the system, or its perceived utility–does not necessarily translate in increased appropriate reliance.

Aligning explanations aim to secure user agreement with the AI, as their central function is persuasive rather than epistemic. While such explanations can strengthen warranted trust when the AI output is correct, they also risk misleading the decision-maker when they effectively justify an erroneous recommendation. Their impact can be assessed behaviourally, for instance by measuring the *switch fraction*—the proportion of instances in which users revise their initial judgement after seeing the AI suggestion (Yin, Wortman Vaughan, and Wallach, 2019). At an aggregated level, alignment may be quantified as the degree of agreement achieved among a group of raters interacting with the system.

By contrast, motivating explanations seek to elicit constructive distrust (Hildebrandt, 2019) and critical scrutiny, deliberately introducing a degree of cognitive friction to trigger deeper reflection. These explanations can be evaluated through psychometric instruments, for example by detecting reductions in automation bias (Goddard, Roudsari, and Wyatt, 2011; Mosier and Skitka, 1999) and increases in cognitive engagement and constructive distrust (Cabitza and Natali, 2022; Hildebrandt, 2019), or in what has been termed reasonable scepticism

towards system outputs (Ehsan and Riedl, 2020). Another indicator is observational: the frequency with which users request additional information or alternative justifications.

In the following section, I will focus on reassuring and aligning explanations with regards to their possible biasing effects.

4.2 WHEN EXPLAINABLE AI MISLEADS

As anticipated, treating explanations as meta-outputs (Ghassemi, Oakden-Rayner, and Beam, 2021) of functional modules that typically operate independently from those that generate the advice makes us aware of the added layer of uncertainty that is introduced by XAI in human-AI interaction (Morrison et al., 2024): just like predictions, they can be wrong, irrelevant, or misleading. Sarkar (2024c) talks of *exoplanations* in the context of LLMs. In particular, placebic (Eiband et al., 2019) or fluent narratives boost confidence irrespective of fidelity. In short, transparency can raise persuasive power without enhancing reliance: Bansal et al., 2021 showed that explanation can increase adherence to AI advice even in case of error.

The explainability paradox

This is an *explainability paradox*: adding explanations does not guarantee appropriate reliance and can reshape confidence in unpredictable ways (Bertrand et al., 2023). In fact, explanation can be overly persuasive, increasing over-reliance on wrong advice (the "white-box paradox", Cabitza et al. 2023d).

Cognitive biases and XAI

As shown by Bertrand et al. (2022) in their review, a range of cognitive biases can undermine the intended benefits of explanations. The mere exposure effect shows that simply providing an explanation—whether meaningful or not—can increase user confidence, as "placebic" explanations are often trusted as much as substantive ones (Eiband et al., 2019). Explanations may also reinforce confirmation bias, with users interpreting the rationale as supporting what they already believe, and recognition bias, where familiar terms or patterns make an explanation seem more credible regardless of its evidential value. The completeness bias further leads clinicians to favour longer, more detailed explanations, even when the additional information is irrelevant, potentially inflating perceived competence and promoting over-reliance. Finally, narrative or causal bias arises when users treat coherent causal stories as genuine mechanisms: a risk heightened by generative models capable of producing persuasive but unfounded explanatory narratives. In all cases, explanation does not necessarily equate to understanding—and may instead obscure the very uncertainties it is meant to clarify.

Together, these biases reveal that explanations, though intended to enhance transparency, can subtly reshape cognition in ways that reduce vigilance and increase reliance.

4.3 EXPLANATIONS AS DEBIASING STRATEGIES

In a memorable opening vignette, Miller (2023) provides an allegory of eXplainable AI in the form of two friends, *Bluster* and *Prudence*. While *Bluster* is straightforward, laconic, and unnervingly self-assured in their own advice: coming to them for support, or for an explanation, means being told what to do, point blank. On the other hand, *Prudence* embodies a reflective, cautious approach that scaffolds the user's thinking.

Reflective HCXAI

The term *Reflective HCXAI* has been introduced by Ehsan and Riedl (2020) to describe a situated, socio-technical approach to XAI grounded in *critical technical practice* (Agre, 1997) and *reflective design* (Sengers et al., 2005). They characterise this approach as both critically self-reflective about the field's implicit assumptions and practices, and value-sensitive to users and designers. In this view, Reflective HCXAI puts into question dominant views in XAI, including the prevailing notion of trust as tied to the goal of nudging users toward finding explanations plausible and simply accepting them. Instead, they argue that in certain contexts, such as misinformation detection, it may be preferable to cultivate the user's "reasonable scepticism and critical reflection," acknowledging that no model is infallible and that explanations will not always be correct. Designing for such scepticism, they contend, creates space for users to express disagreement, recognise system limitations, and engage more actively in the decision process.

Focusing on the explanatory relationship offers a means to foster desirable reliance behaviours within specific, situated contexts. In high-stakes, context-rich domains such as clinical diagnosis, motivating explanations can scaffold verification and contestation, making them preferable to reassuring glosses that merely appease. This entails the deliberate use of human cognitive debiasing techniques (Lighthall and Vazquez-Guillamet, 2016) or debiasing strategies (Bertrand et al., 2022).

Debiasing strategies

The design remedy to the shortcomings of explanations is not their abandonment, but their integration into interaction protocols that safeguard human reasoning (e.g., commit-before-reveal, paced disclosure, contrastive checks). Rather than adhering to the traditional "recommend and defend" model of the eXplaining machine, contemporary human-centred XAI research is shifting toward systems that actively provoke and engage cognition through inquiry. Instead of merely pro-

Explanations that make the user think

viding answers or conclusions, such systems prompt users with new questions, challenge assumptions, assemble evidence for and against a claim, expose uncertainty: here, I will provide an overview of interaction techniques deliberately designed to "make the user think".

Conflict-based XAI

Adversarial explanations can function as debiasing strategies. These "devil's advocate" or *provocative* explanations deliberately present information that challenges the AI's own recommendation. As Sarkar et al. (2024) describe, such provocations can take the form of "short text snippets that critique the AI-generated criteria, highlighting risks, shortcomings, and alternatives." In practice, this may involve surfacing evidence that contradicts the model's prediction or proposing an alternative outcome, prompting the user to compare and reflect. For instance, in the *Judicial AI* case proposed by Cabitza et al. (2025b) and Fregosi and Cabitza (2024), dichotomous explanations compel decision-makers to reconcile conflicting arguments, thereby actively engaging them in evaluating evidence.

This approach is closely related to, yet distinct from, the concept of *Evaluative AI* (Miller, 2023), in which the decision-support process is hypothesis-driven. Here, the AI system provides evidence both for and against hypotheses initially formulated by the human. By engaging the human first—before any framing effect can occur—this paradigm mitigates over- and under-reliance, leveraging human expertise while granting decision-makers agency to explore the relative strengths and weaknesses of multiple, self-defined options. The deliberate introduction of such "conflicting rules or knowledge" has been identified by Kliegr, Bahník, and Fürnkranz (2021) as a debiasing technique against overconfidence and underconfidence. Further examples include the works of Bhatt et al. (2021), Bussone, Stumpf, and O'Sullivan (2015), Wang et al. (2019), and Wolfe and Britt (2008).

Socratic or *reflective* AI systems lead users through questions instead of handing down answers. Rather than behaving as obedient servants, these systems act as Socratic partners – more like a tutor or a "gadfly" – that provoke deeper inquiry. Sarkar (2024a) argues that AI assistants should "challenge, not obey", reframing our "robot secretaries" as provocateurs that question the user's reasoning and offer counterpoints . For example, an AI writing assistant in the provocateur role might respond to a draft not by simply fixing errors or completing sentences, but by critiquing the argument, pointing out assumptions, and offering guiding metacognitive questions that encourage critical thinking and deeper scrutiny of AI-generated outputs. This Socratic style of interaction – inspired by the classical method of teaching through probing questions – compels the user to articulate their un-

Socratic XAI

derstanding and uncertainties.

Danry et al. (2023) present *AI-framed Questioning* as a *Human-AI co-reasoning system*, an approach in which the features or cues underpinning the AI's classification are reformulated as targeted questions, designed to prompt user reflection and actively scaffold the human reasoning process. A related line of work introduces *reflection machines*, defined by Haselager et al. as "an additional computational system to support effective and meaningful human oversight over a DSS" Haselager et al. (2024, p. 381), where provoking questions at decision-making junctures are intended to engance cognitive engagement and improve calibration of reliance (Fischer, 2025). Fischer et al. (2025) provide a taxonomy of such explanations, aimed at supporting the design and development of cognitive interventions. It features 10 questions pertaining to *Case Information*, *Relevance of Data*, *Datasets*, *Causal structure of recommendation*, *Alternatives to recommendation*, *Assumptions and expectations of the decision-maker*, *Stakeholder preferences*, *Consequences of decision*, *Change intervention*, and *Model behaviour*. For example, a question on the *assumptions and expectations of the decision-maker* prompts the user to voice their assumptions, with the aim of increasing their awareness of their own reasons and possible biases.

Fischer et al. (2025) describe this approach as a *Socratic* method. In the words of Paul and Elder (2019),

> "Socratic questioning is disciplined questioning that can be used to pursue thought in many directions and for many purposes, including: to explore complex idea, to get to the truth of things, to open up issues and problems, to uncover assumptions, to analyze concepts, to distinguish what we know from what we don't know, and to follow out logical implications of thought." (Paul and Elder, 2019, p. 4)

Socratic interventions are represented in a flourishing HCAI literature (Ang, Gollapalli, and Ng, 2023; Fischer et al., 2025; Lara, 2021; Liu et al., 2024), with empirical studies showing that Socratic method increased students' critical thinking (Dalim, 2022; Ho, Chen, and Li, 2023).

Human deliberation theories also informed the LLM-based *Deliberative AI* approach proposed by Ma et al. (2025), which has been shown to outperform feature contribution-based explanations in terms of appropriate reliance.

When AI systems directly prompting users with questions or doubts, the interaction shifts from one of machine explanation to one of human reflection.

Another category of questioning-machine techniques involves how the AI delivers its outputs – moving away from definitive, one-shot

answers (post-hoc) toward outputs that encourage comparison and analogical reasoning.

Abstention and partial explanations

Abstention-based outputs involve the AI sometimes choosing not to provide a direct answer, which paradoxically can enhance the overall decision process. Instead of always producing a potentially flawed recommendation, a well-designed AI can refuse to answer or express uncertainty when it lacks confidence. Far from being a failure of the system, this can be a powerful prompt for human engagement as a learning-to-defer strategy (Hemmer et al., 2023; Madras, Pitassi, and Zemel, 2018; Pugnana et al., 2025). Introducing an abstain option (a "selective refusal" by the model when uncertainty is high) would reduce the incidence of automation bias by preventing misleading outputs in the first place. An example is that of the application of *three–way decisions* (Campagner, Cabitza, and Ciucci, 2019; Yao, 2009), allowing for an abstain option (Campagner et al., 2024).

Similarly, the Intelligent Decision Assistant introduced by Schemmer, Kühl, and Satzger (2021) provides decision support without providing a definitive decision, bur rather by explaining its reasoning. This form of *informative guidance* is shared by *minimal rationales* or *partial explanations*. These minimal rationales are succint cues instead of exhaustive justification (Gajos and Mamykina, 2022). They only provide the core insight of an explanation—a brief pointer to what the model "noticed" or found salient—without elaborating further. A radiological example would be a system that highlights a specific region in an X-ray without giving an explicit diagnosis. Such rationales act less as conclusions and more as prompts, inviting users to interrogate the case or verify a specific aspect themselves. The guiding principle is to resist the temptation to over-explain or over-guide, as doing so risks transforming the explanation into a ready-made analysis rather than a cognitive partner.

5

THE CASE FOR FRICTION IN HUMAN-COMPUTER INTERACTION AND AI DECISION SUPPORT SYSTEMS

> *We have a huge opportunity–and a dire need–to recode our work and our technologies so they engance challenge, complexity, and connection and the valuable skills that flow from them.*
> — Don Norman, *The Design of Everyday Things* (Norman, 2013)

After introducing the idea that a certain degree of effort can be desirable to stimulate cognitive engagement in decision-making, it is useful to take a step back. Human–Computer Interaction has long been guided by the ideal of the seamless interface—a vision of interaction so smooth and transparent that it removes any possibility of friction between users and digital systems (Natale and Treré, 2024). The two goals of HCI, Helander (2014) argued, are exactly *usefulness* and *usability*. Usability has traditionally been equated with minimising effort—reducing clicks, simplifying choices, and lowering cognitive load. This includes physical friction: modern laptop keyboards are continuously refined to eliminate key sounds and reduce the effort required to press keys, reducing both effort and distraction from pressing keys and hearing their noise (Ericson, 2022). This view coupled seamlessness with lowered cognitive load, suggesting that when things "disappear," users are freed to use them without thinking and can focus beyond them on new goals (Inman and Ribes, 2019; Sarkar, 2023b). Foundational design principles and usability heuristics—most notably Nielsen's 1994 guidelines—were formulated to support the creation of user interfaces that were first and foremost *efficient*, emphasising qualities like consistency, predictability, and error prevention (Nielsen, 1994). While these principles have proven highly effective for graphical user interfaces, they rest on the assumption of relatively deterministic system behaviour and a uniform interaction flow. In contrast, AI-infused systems exhibit behaviours that are inherently dynamic, context-dependent, and often unpredictable, evolving as they learn over time. Consequently, good interfaces in the traditional sense—such as visually clean or ergonomically streamlined dashboards—do not automatically translate into effective interactions in complex human–AI settings (Amershi et al., 2019).

Usability and cognitive load

In decision-support contexts, excessive ease can be counterproductive. When an AI system makes it trivially simple to accept its rec-

ommendations, users risk becoming passive automation followers, as is the case in automation bias (Skitka, Mosier, and Burdick, 1999). Research on decision aids shows that users are more likely to accept automated suggestions with minimal scrutiny when the interface makes compliance the path of least resistance, effectively bypassing effortful reasoning (Skitka, Mosier, and Burdick, 1999). Recent work on designing with friction argues that the obsession with quickness and smoothness in interfaces has led to opaque systems that shield users from important details (Benedetti and Mauri, 2023). In such cases, a design that is overly "user-friendly" in the narrow sense can undermine precisely the cognitive engagement needed for sound judgment.

As Sarkar (2023b) notes, dominant traditions in HCI have long prioritised the minimisation of cognitive effort, treating friction as something to be removed. This orientation is reflected in the evaluation tools most commonly used to judge system quality. For instance, the *Technology Acceptance Model* (TAM) (Davis et al., 1989) explicitly includes items such as *"Interacting with the system does not require a lot of my mental effort."* Likewise, the NASA Task Load Index (NASA-TLX) (Hart and Staveland, 1988) assesses perceived mental and perceptual workload in terms of the degree of thinking, remembering, and searching required, implicitly treating high load as undesirable (Sarkar, 2023b). These instruments exemplify a broader assumption in classic usability: that good design is synonymous with simplicity, smoothness, and minimal mental demand.

5.1 BEYOND USABILITY

Traditional usability has a blind spot: a preoccupation with short-term efficiency and error avoidance, while neglecting the longer-term cognitive outcomes of interaction—understanding, skill development, and trust calibration. As Norman's framework of design experience reminds us, interaction occurs across multiple levels—visceral, behavioural, and reflective (Norman, 2007). A narrow focus on the behavioural level (efficiency and ease) may optimise immediate task performance, but it disregards the reflective processes through which users extract meaning, learn from feedback, and refine strategies. Users are often willing to tolerate friction or difficulty when it yields such higher-order benefits. In *The Design of Everyday Things*, Norman (2013, p. 109) emphasises that tasks should maintain an "appropriate level of difficulty: challenging enough to demand continued attention, yet not so difficult as to induce frustration." Likewise, minor usability shortcomings may be offset by a "strongly positive visceral response" that sustains engagement and motivation (Norman, 2013, p. 106). Frictional designs, such as speculative and critical designs, are defined

by Pierce (2021) as *transproductional*, meaning they offer value—such as critique, reflection, debate, and discourse—that goes beyond the aim of achieving a commercially viable product (Pierce, 2021). The value of this work is not in solving problems quickly, but in creating meaningful experiences tailored to unique needs. Even in game design, where enjoyment and user-friendliness are paramount, *deliberate inefficiencies* or *intentional friction* in the user interface can promote user reflection and create meaningful player experiences (Silva, Cardoso, and Giesteira, 2022).

Transposed to AI decision support, these "higher rewards" manifest as calibrated trust, deeper insight, and confidence in one's reasoning process. An interface paradigm that seeks to eliminate all friction risks also eliminating the very opportunities for reflection, learning, and cognitive growth that make human–AI collaboration meaningful.

Designers and researchers are thus reconsidering the long-standing ethos of maximising automation, efficiency, and cognitive ease in system design.

Critical HCI

In HCI, this critique takes the form of reflective design (Sengers et al., 2005) and slow technology (Hallnäs and Redström, 2001) as well as seamful design. In seamful design (Chalmers, 2003), uncertainty, provenance, and breakdowns are revealed so that users can interrogate and repair systems rather than defer to them. Inman and Ribes's concept of "beautiful seams" (Inman and Ribes, 2019) similarly advocates for interfaces that reveal seams or gaps instead of hiding them, on the theory that exposing some of the system's inner workings or uncertainties prompts users to engage more critically. While a seamless design would conceal these complexities for the sake of simplicity, a seamful (frictional) design brings them to the forefront, trusting the user to grapple with them.

Critical Design (Bardzell et al., 2012) is a broad movement that leverages technology to raise critical questions and provoke reflection on cultural and societal norms. It promotes reflection and critique by subverting expectations and assumptions, sometimes achieved by making technology intentionally "unfriendly" (Cox et al., 2016; Dunne, 2008). It is related to speculative design (Dunne, 2008). Ohm and Frankle (2018) point to *desirable inefficiencies* in digital systems design – *"when the efficient alternative fails to provide or protect some essential human value, such as fairness or trust"*(Ohm and Frankle, 2018, p. 777). The guiding ethos of these critical approaches is "in tension with progression" (Pierce, 2021), intended as the movement toward final production and solution-finding. In the words of Frischmann and Selinger (2019), tolerating some inefficiency, is instrumental to resist the dominance of efficiency and productivity logics and to promote human flourishing

through the exercise and development of human capabilities.

Pierce's early work introduced the idea of "digital limitations" (Pierce, 2012) through the design of counterfunctional things: artifacts that intentionally complicate aspects of their own use. Rather than optimising efficiency or convenience, these designs strategically withhold or disrupt functionality to encourage more intentional engagement. This orientation was shaped by insights from nudging theory and research on choice overload, both of which suggest that well-placed constraints can support more deliberate decision-making. He later developed this line of inquiry into a broader framework of frictional design (Pierce, 2021), rooted in his work on "undesigned" and alternative design practices (Pierce, 2012).

The framework outlines several characteristic tendencies by which friction can be intentionally introduced. *Divergent* frictions uproot familiar conventions, creating experiences that feel unusual enough to provoke curiosity or unease. Others adopt a more directly *oppositional* stance, using friction as a means to criticise prevailing norms and expectations. A further mode, *accelerational* friction, works by exaggerating current trajectories to an extreme point—pushing the familiar into discomfort or absurdity to expose its underlying dynamics. *Counterfactual* friction, by contrast, invites users to consider alternative histories or possible worlds, drawing attention to contingency and the non-inevitability of the present. Finally, *analogical* friction resists literal or direct representation, instead encouraging users to interpret through association, metaphor, or inference (Pierce, 2021).

Deliberately introducing these small frictions can surface the consequences of user actions and invite reflection (Cox et al., 2016). These approaches do not reject usability point blank; they calibrate it to the *stakes* and *learning goals* of the task. In expert domains, a degree of cognitive friction can be a feature rather than a bug: it helps sustain vigilance, preserve skills, and expose mismatch between model competence and case specifics. In place of the *doctrine of simplicity* (Sarkar, 2023b), a countervailing paradigm is gaining traction—one that purposefully incorporates cognitive friction or cognitive forcing functions (Buçinca, Malaya, and Gajos, 2021) into human–AI interactions to stimulate reflective thinking and sustain users' active engagement with the task. The motivation for this shift stems from growing evidence that highly "seamless" AI support can encourage users to default to fast, intuitive, and often uncritical System 1 reasoning, rather than engaging the slower, more analytical System 2 processes required for careful judgement (Kahneman, 2013; Rastogi et al., 2022). Introducing small, well-calibrated "speed bumps" into the interaction (Frischmann and Benesch, 2023; Mejtoft et al., 2023)—such as prompts that require

Cognitive forcing functions

justification, reflection, or explicit comparison—can function as micro-boundaries (Cox et al., 2016) that interrupt automatic acceptance of AI recommendations and sustain active user engagement.

Within this framing, disfluency is valued for its potential to act as a cue signalling the need for deeper processing (Alter et al., 2007). As noted in cognitive neuroscience, experiences of conflict or difficulty trigger activity in the anterior cingulate cortex, which in turn alerts prefrontal regions involved in deliberative control (Botvinick et al., 2001).

As Miller (2023) argues, a well-designed decision-support system should do more than present outcomes—it should assist users in mapping the landscape of possible choices, clarifying consequences and underlying values, and supporting the weighing of trade-offs in a manner that remains intelligible to human reasoning. In contexts where over-reliance and premature closure are common, the design challenge is not to make interactions ever smoother, but to introduce desirable difficulties that sustain analytical engagement—small, well-calibrated interruptions that "are designed with intention, and introduced with care" (Cox et al., 2016, p. 1391) to prompt reflection, mitigate bias, and preserve agency. This calls for decision support systems that leverage cognitive friction and complexity instead of obfuscating it, inspiring the name *Frictional AI* (Cabitza et al., 2024b; Natali, 2023).

5.2 DESIRABLE DIFFICULTY AND DISFLUENCY

The concept of "desirable difficulty" emerged in the field of cognitive psychology and education, based on the paradoxical finding that making information or a task harder to learn can benefit long-term retention and performance.

The concept is supported by a wealth of literature stemming from work by researchers like Robert Bjork (Bjork, Bjork, et al., 2011; Bjork, 1994). The core mechanism of desirable difficulty is the creation of conditions that make information harder to learn in the moment, which in turn encourages people to process the information more deeply, slowly, and effortfully than they otherwise would (Taylor et al., 2020). This deeper processing leads to better integration of the information with existing knowledge, resulting in better long-term memory (Craik and Tulving, 1975; Taylor et al., 2020).

Some of the reported strategies for desirable difficulty in the literature are spacing out repetitions of studying information, reading aloud (Taylor et al., 2020), or discarding typing in favour of the more effortful writing with pen on paper, which has been shown through neuroimaging to engage the brain more deeply, showing heightened activation in areas connected to cognitive processing (Marano et al., 2025).

Figure 5.1: Example of the *Sans Forgetica* typeface, reading "Desirable difficulty or disfluency?"

However, some difficulties are just disfluent (Taylor et al., 2020). A well-known example is *Sans Forgetica* (Fig 5.1), a deliberately disfluent typeface designed to slow reading and thereby encourage deeper processing. Although widely promoted—garnering media attention and design awards—the font has since become a case study in the limitations of disfluency-based interventions.

The theoretical premise behind Sans Forgetica was straightforward: perceptual difficulty may act as a metacognitive cue, signalling that more careful, systematic processing (System 2) is required. By making text slightly harder to read, the font was expected to trigger such deeper encoding. This simplicity led to a wave of interest in whether disfluent fonts could boost memory without changing content.

However, empirical results have consistently shown that Sans Forgetica increases felt difficulty without improving memory performance (Taylor et al., 2020). Participants reliably judged the font as harder to read, yet recall and comprehension remained equivalent to standard fonts (e.g., Arial), and in some tasks—such as word-pair recall—performance was actually worse. In other words, the intervention succeeded in producing difficulty but failed to produce desirable difficulty.

Desirable difficulty vs. Disfluency

This outcome aligns with meta-analyses showing that perceptual disfluency generally has negligible effects on learning outcomes in text-based contexts (Xie, Zhou, and Liu, 2018), pointing to the need for careful evaluation of friction-based design practices, rather than assuming that difficulty alone will enhance learning.

5.3 FRICTIONAL STRATEGIES IN THE LITERATURE

The literature on applying friction in real-world user studies—often framed under concepts like design friction, cognitive forcing, or Frictional AI—demonstrates that intentionally introduced difficulties can yield benefits, particularly in promoting reflection, mitigating biases, and increasing user agency, though often at the cost of subjective preference or efficiency of Human–AI interaction.

In high-stakes domains and complex decision-making scenarios, friction is primarily studied as a mechanism to compel users to engage analytically rather than over-rely on automated suggestions.

Cognitive forcing interventions, such as implementing a thirty-second timeout before an AI recommendation is shown or requiring users to actively request the recommendation, significantly reduced overreliance on AI suggestions compared to simpler explainable XAI approaches in the study by Buçinca, Malaya, and Gajos (2021). However, systems that reduced overreliance the most typically received the least favorable subjective ratings from participants, highlighting a trade-off between subjective trust, preference, and performance (Buçinca, Malaya, and Gajos, 2021). When bias mitigation techniques such as requiring decision justification and pre-committing a decision before accessing AI support were explored with ophthalmologists using AI-assisted decision support tools, researchers encountered substantial concern over a decrease in efficiency (Bach et al., 2023), as the time pressure faced by clinicians limits the appeal of mitigation strategies that require additional user input.

Buçinca, Malaya, and Gajos (2021) and Jong et al. (2025) observed that cognitive forcing functions disproportionately benefit individuals with high *Need for Cognition* (a trait reflecting a motivation to engage in effortful mental activities, Cacioppo and Petty 1982) which suggests that these friction-based interventions may produce intervention-generated inequalities.

Need for Cognition and cognitive forcing

In a study of GenAI assistants, introducing moderate friction (a chatbot plus a summary) evoked significantly higher levels of agency and perceived usefulness than either no-friction (regular chatbot) or high-friction (chatbot plus a detailed form) interfaces (Malaguti et al., 2025). Critically, the introduction of friction in this context did not have a significant negative effect on perceived usefulness, cognitive load, or task performance compared to the frictionless AI mode. Users reported higher control, confidence, and trust in the AI systems incorporating friction elements like summaries and forms. (Malaguti et al., 2025)

Reicherts, Park, and Rogers (2022) devised a *ProberBot*–a conversational agent designed to probe and prompt users. Testing it in a stock investment task found that participants appreciated how the bot made them slow down to think through decisions in a structured way, viewing it as a valuable cognitive tool that helped mitigate impulsive behavior. However, integrating these probes presents challenges, as users sometimes perceived them as intrusive or irrelevant.

User studies exploring *provocations* (AI-generated critiques of AI suggestions) in knowledge work found they had strong qualitative effects, successfully inducing metacognition and critical thinking and coun-

teracting the tendency for mechanized convergence (Drosos, Sarkar, Toronto, et al., 2025). However, quantitative results were inconclusive, and surprisingly, there was some evidence of potential overreliance on the provocations themselves, as participants in the control condition sometimes reported considering the possibility of AI error more than the provocation condition.(Drosos, Sarkar, Toronto, et al., 2025)

Programmed inefficiencies

Cabitza et al. use the term *programmed inefficiencies* to describe features intentionally built into decision support systems that make it impossible or undesirable for the human operator to behave as a passive automation consumer (Cabitza et al., 2019). In their 2019 work, they illustrate this with the design of a medical AI system that avoids giving a single definitive answer. Rather, it might output multiple partially conflicting indicators or an ambiguous visualization that the human must interpret. For example, instead of a binary diagnosis, the system could present a "jigsaw puzzle" of clues—redundant or even contradictory pieces of information about the case that the clinician has to assemble and make sense of (Cabitza et al., 2019). This deliberate ambiguity forces the user to synthesize information, exercising their expertise.

The role of ambiguity

Ambiguity can be an asset when used to highlight uncertainty rather than conceal it. Gaver, Beaver, and Benford, 2003 identified ambiguity as a means to provoke interpretation and speculation, arguing that designers can purposefully use imprecision to foreground the multiplicity of possible truths. In decision-support, this can take the form of vague visualisations (Assale, Bordogna, and Cabitza, 2020)—for instance, confidence intervals, probabilistic sets, or non-exclusive categories—that resist the illusion of determinism and encourage users to reason about uncertainty instead of anchoring on point estimates. In this sense, the AI system acts less as an oracle (Miller and Masarie Jr, 1990) and more as a muse (Sarkar, 2024b), which provokes meaning-making without claiming authority.

Conflict-based strategies exploit the epistemic value of disagreement. Rather than optimising for harmony between user and system, they introduce argumentative friction to expose blind spots and counter biases. Cai, Arawjo, and Glassman, 2024 identify several forms of productive antagonism, ranging from adversarial (zero-sum) to argumentative (value-challenging) and personal (behaviour-challenging) forms, all of which can be designed to stimulate self-reflection. In decision support, this translates into agonistic configurations, where multiple models present opposing arguments for a decision (Hildebrandt, 2018). These models may differ in architecture, training data, or optimisation criteria (e.g., sensitivity versus specificity), thereby

surfacing the assumptions underpinning each conclusion.

A related direction is the *Human–AI Deliberation* paradigm proposed by Ma et al. (2025). In this approach, the AI system—implemented as a large language model–based assistant—does not simply provide answers or explanations; rather, it actively engages the user in structured discussion around conflicting viewpoints. Through conversational exchange and targeted presentation of information, the system aims to support users in weighing opposing perspectives and forming more considered judgements. In their mixed-methods evaluation, the authors report that such *Deliberative AI* encourages more appropriate reliance on the system and leads to improved task performance when compared with conventional XAI interfaces. However, users also reported notably lower satisfaction when interacting with the deliberative system than when performing the task unaided. The authors suggest that this decrease in perceived usability stems from the system drawing attention to unresolved tensions and forcing users to do more cognitive work. This aligns with earlier findings that highlight a tension between promoting reflective, effortful reasoning and preserving a positive user experience (Buçinca, Malaya, and Gajos, 2021). In fact conflict must be proportionate: as Aicher et al. (2024) caution, indiscriminate opposition can reinforce defensive reasoning rather than reflection. Following Huang, Hsu, and Ku (2012), an intelligent system should adjust the frequency, timing, and selection of counter-arguments dynamically, balancing provocation with receptivity.

Conflict-based strategies can also have an explicit pedagogical aim, *de facto* amounting to a scaffolding-based strategy. They guide users to examine their own reasoning, but do so by offering a helping hand. Reicherts, Park, and Rogers (2022) propose designing interfaces that "encourage users to step back and think about what they are doing," framing AI interaction as a learning opportunity. Dialectic decision support systems (Jarupathirun et al., 2007) do not prescribe what users should think, but instead scaffold how they should think by encouraging the critical comparison of alternatives. Rather than Socrates (Ang, Gollapalli, and Ng, 2023; Fischer et al., 2025; Lara, 2021; Liu et al., 2024), the philosophical inspiration for this approach is Hegel in the a multi-perspective reconciliation framing of thesis—antithesis–synthesis, which promotes the critical comparison of alternatives. Chae, Courtney, and Haynes (2005) argue that a dialectic approach is apt to deal with wicked problems, — for which "There are no criteria for correctness" and "no well defined solutions" (Skaburskis, 2008, p. 278). Schwenk and Valacich (1994) similarly point to the relevance of the dialectic decision support for unstructured tasks.

Other scaffolding mechanisms include AI-framed questioning (Danry et al., 2023)—prompts that ask users to justify, defend, or challenge their assumptions—and explicit decision justification requirements. For instance, when clinicians were asked to write short rationales before reviewing AI advice, this practice helped them organise their thoughts and reflect on their reasoning, even though it raised concerns about time efficiency (Bach et al., 2023).

Designers have long recognised that not all discomfort is detrimental. Benford et al. (2012) describe uncomfortable interactions as those that deliberately unsettle the user to create memorable, meaningful experiences. Discomfort is *the dark side of fun* (Benford et al., 2018), and can be elicited as in the previously-mentioned case of deliberate inefficiencies and intentional friction is game design (Silva, Cardoso, and Giesteira, 2022). Those uncomfortable interactions are a form of the micro-boundaries proposed by (Cox et al., 2016), or *nudging-through-friction* – alerts that disrupt the user's auto-pilot mode to point them towards system explanations.

Time itself can act as a debiasing resource. A growing body of research demonstrates that ordering and temporal pacing of AI advice critically shape reliance patterns (Buçinca, Malaya, and Gajos, 2021; Lai and Tan, 2019; Wang et al., 2019). Human-first protocols require the user to form an independent judgement before the AI provides its recommendation. This simple reordering forces the user to engage their own analytic reasoning before being anchored by the machine's output. Empirical studies show that such sequencing markedly reduces over-trust and increases error detection.

"Slow" interaction designs extend this logic by introducing deliberate latency between system stages, encouraging users to deliberate or cross-check before proceeding (Springer and Whittaker, 2019). Kliegr, Bahník, and Fürnkranz (2021) describe how time mitigates anchoring and recognition heuristics: under time pressure, people tend to favour familiar or easily recognised options, even when inappropriate. Allowing more time, or deliberately designing for "slowing down," helps users override this bias. Abstention mechanisms apply the same logic of restraint. Systems may choose not to respond when confidence is low (selective prediction), or provide only partial information. Recent studies show that partial explanations—which highlight errors without offering corrections—reduce over-reliance on wrong advice, albeit at some cost to reliance on correct ones (Jong et al., 2025).

Across these modalities, frictional AI replaces the doctrine of seamlessness with a pedagogy of reflection, or negotiated complexity (Sarkar, 2023b). Ambiguity draws attention to uncertainty; conflict

exposes assumptions; discomfort interrupts complacency; timing and abstention reorder attention; and scaffolding consolidates learning.

6

SCIENTIFIC GAP AND RESEARCH PROGRAMME

> *We have a huge opportunity–and a dire need–to recode our work and our technologies so they engance challenge, complexity, and connection and the valuable skills that flow from them.*
> — Matthew Beane, *The Skill Code* (Beane, 2024)

The literatures reviewed converge toward a shared realisation that mirrors Shneiderman 2020's "second Copernican revolution" in Human–AI Interaction and Hybrid Intelligence: the shift from better models to better interactions. What determines whether AI yields benefit or harm is not accuracy alone, but the protocol that governs who attends to what, when, and why.

However, existing work on clinical AI and decision support has predominantly assessed systems in terms of predictive performance, with comparatively less attention to how these systems reshape professional judgement, learning trajectories, and responsibility in everyday practice. Research in medical AI has begun to highlight mismatches between reported performance and clinical impact, but systematic evaluative frameworks that account for human–AI interaction patterns remain rare. At the same time, CSCW and studies of clinical work have shown how decision-support tools reconfigure documentation, coordination, and accountability. Theories of extended and distributed cognition similarly provide rich vocabularies for understanding sociotechnical reasoning, but they have seldom been operationalised into concrete design and assessment tools for AI decision support. Finally, while HCI and XAI research has explored issues of trust, transparency, and user experience, much of this work assumes that usability and explainability correspond to an unquestionable benefit, with limited engagement with the possibility that ease and persuasive explanations may degrade professional skill and vigilance over time.

Together, these bodies of work leave at least three gaps. First, we lack a consolidated, interaction-centred account of hybrid intelligence in clinical settings that integrates CSCW, cognitive, and technical perspectives. Second, there is no widely adopted evaluative grammar for assessing AI's influence on human decision-making beyond aggregate accuracy. Third, design strategies that deliberately use friction and disfluency to sustain expertise remain under-theorised and under-evaluated in medical AI.

6.1 FROM MODEL TO PROTOCOL

Building on the notion of Human–AI Collaboration Protocols (Cabitza et al., 2023e; Wilson et al., 2025), we reconceptualised decision support as a designed protocol rather than a static artefact. The concept of *emergence-oriented design* (Cabitza et al., 2025c) provides the philosophical grounding for this shift, portraying intelligence as distributed and de-individuated across socio-technical practice. Within this view, intelligence "lies not in the system but in its situated use": design must therefore orchestrate relations—sequencing, timing, and commitment—through which hybrid intelligence unfolds.

6.2 FROM ACCURACY TO COGNITIVE EFFECTS

Through the *Human–AI Interaction Assessment* tool (Natali, Campagner, and Cabitza, 2024), I extend evaluation from model performance to decision influence, reliance patterns, and distributive effects across expertise levels. These contributions collectively advance an evaluative grammar that accounts for behavioural, cognitive, and organisational consequences of AI use.

6.3 FROM USABILITY TO FRICTION

Drawing on interaction design and cognitive psychology, my research reframes friction—traditionally seen as inefficiency—as a resource for reflection and skill sustainment. The studies on frictional AI (Cabitza et al., 2023b, 2024b; Rubegni et al., 2025) collectively advance Frictional AI as a design paradigm that counters automation bias and epistemic sclerosis by embedding "desirable inefficiencies" into human–AI protocols.

These strands constitute a coherent research programme that positions itself within interaction-centred evaluations for Hybrid Intelligence—integrating design frameworks (Natali et al., 2024), empirical evaluation (Natali, Campagner, and Cabitza, 2024; Natali et al., 2023), and design experimentation (Cabitza et al., 2024b; Rubegni et al., 2025). In doing so, it contributes a unifying theoretical and methodological lens through which the benefits and harms of AI can be understood as outcomes specific Human–AI Collaboration Protocols.

Part II

FINDINGS

7

STUDIES OVERVIEW

This part presents the original research contributions of the thesis, synthesising a set of peer-reviewed studies and preliminary findings into a coherent narrative, organised around the type of contribution each body of work makes to the overarching research problem: conceptual, methodological, and design-oriented.

Conceptual contributions (Chapter 8) reconceptualise human–AI collaboration through an emergence-oriented lens, and articulate core risks such as deskilling and epistemic sclerosis.

Methodological contributions (Chapter 9) advance how the impact of AI on human decision-making is evaluated, moving beyond accuracy-centred metrics.

Design-oriented empirical contributions (Chapter 10) investigate frictional interaction protocols in AI-assisted clinical decision-making and assess their effects on reasoning, reliance, and professional agency.

Each chapter brings together multiple empirical or conceptual works and is preceded by a short framing section that clarifies the motivations for the included studies and how they advance the thesis argument.

7.1 AUTHOR CONTRIBUTIONS AND ORGANISATION OF THE STUDIES

The main findings reported are anchored in publications and completed studies to which I made the primary conceptual, methodological, or analytical contribution. These works constitute the core empirical and conceptual backbone of the dissertation and they are presented in detail, including their motivation, study design, methodological approach, and key findings.

Published works include the full bibliographic reference. Contributions labelled as *Preliminary findings* report completed studies that have not yet undergone peer review.Where authorship order does not correspond to first authorship, my specific contributions are made explicit.

In addition, each part includes a subsection titled *Further work*, which narratively integrates publications to which I contributed substantially but for which I was not the primary author, as well as first-author works whose relevance to the thesis concerns a specific

analytical dimension rather than the full research question (e.g. the work on epistemic sclerosis in Section 8.3.2, and the Human–AI interaction dimension of the assessment tool presented in Section 9.2.2). These contributions extend and contextualise the core findings presented in the main sections, supporting their interpretation through complementary perspectives.

Importantly, the thesis makes an original and independent contribution by synthesising these studies into a unified framework for understanding, evaluating, and designing *frictional* Human–AI interaction protocols—an integration that goes beyond the scope of any individual publication.

Together, the studies presented in this part substantiate the central claim of the thesis: that the consequences of AI in clinical decision-making emerge not from model performance alone, but from how human and machine contributions are structured, evaluated, and experienced in practice.

Table 7.1: Conceptual contributions underpinning Chapter 8

Study	Author contribution	Thesis contribution
Chiara Natali et al. (2024). "Humans in the Group, Computers in the Coop. Comparison of Individual and Collective Improvement in Cognitive Tasks in Adjunct AI Settings." In: *IFIP Working Conference on Human Work Interaction Design*. Springer, pp. 174–191	Organised and managed experimental sessions; conducted the analyses and led the writing of the manuscript.	Provides empirical grounding for an emergence-oriented view of human–AI collaboration, showing that the effects of AI on performance and reliance depend on interaction protocols and collective configurations rather than model performance alone. Establishes adjunct and group-based AI as meaningful alternatives to individual, dyadic decision support.
Ben Wilson et al. (2025). "Dimensions of human-machine combination: prompting the development of deployable intelligent decision systems for situated clinical contexts." In: *Computer Supported Cooperative Work (CSCW)*, pp. 1–57	Originated and authored the disambiguation section on *combination*, *cooperation*, and *hybridity*; refined terminology concerning control relations; developed and wrote the material connecting the framework to the EU AI Act, model cards, and intended use; co-authored subsequent revisions.	Supplies the conceptual and analytical vocabulary used throughout the thesis to characterise human–AI interaction as a situated, relational space. Enables later methodological and design contributions by formalising how control, timing, overlap, and influence shape hybrid decision-making.
Federico Cabitza et al. (2025c). "Beyond cyborgs: the cybork idea for the de-individuation of (artificial) intelligence and an emergence-oriented design." In: *AI & SOCIETY* 40.5, pp. 3333–3348	Structured the overall contribution by integrating three distinct essays by the co-authors; originated and authored the sections on *Extended Cognition* and *Implications for Socio-Technical Design*.	Provides the philosophical foundation for the thesis by reframing human–AI systems as emergent cognitive assemblages rather than optimisation pipelines. Justifies the shift away from representational and accuracy-centric accounts of AI evaluation.

Table 7.3: Conceptual contributions underpinning Chapter 8

Study	Author contribution	Thesis contribution
Chiara Natali et al. (2025). "AI-induced Deskilling in Medicine: A Mixed-Method Review and Research Agenda for Healthcare and Beyond." In: *Artificial Intelligence Review* 58.11, pp. 1–40	Oversaw and coordinated the writing of the contribution; collaboratively authored the full manuscript, with particular responsibility for the research agenda and design implications.	Establishes deskilling as a systemic and design-mediated risk of AI-supported work, rather than an individual deficit. Motivates the thesis's focus on interaction design as a means of preserving professional agency and clinical competence.
Chiara Natali and Federico Cabitza (2025). "Make Some Noise for Ground Truthing! Frictional design against epistemic sclerosis in Decision Support Systems." In	Following conceptual input from the co-author on *epistemic sclerosis* and *perspectivism*, independently conducted the literature review; authored the full manuscript; developed the theoretical argument and articulated the design implications.	Introduces epistemic sclerosis as a distinct failure mode of AI-mediated decision-making. Strengthens the thesis's claim that deliberate friction is required to sustain epistemic diversity, contestability, and reflective judgement in clinical practice.

Table 7.5: Methodological contributions underpinning Chapter 9

Study	Author contribution	Thesis contribution
Chiara Natali et al. (2023). "Color shadows 2: Assessing the impact of xai on diagnostic decision-making." In: *World Conference on Explainable Artificial Intelligence*. Springer, pp. 618–629	Led the conceptual framing and writing of the contribution; deployed the human–AI interaction assessment framework; analysed reliance patterns, automation bias, algorithmic aversion, and technology impact.	Provides the first full empirical instantiation in the thesis of a *beyond-accuracy* evaluation of XAI. Demonstrates that explanations systematically reshape diagnostic reasoning, reliance, and performance distributions across users, thereby motivating interaction-sensitive assessment over purely predictive benchmarks.
Chiara Natali, Andrea Campagner, and Federico Cabitza (2024). "Answering the Call to Go Beyond Accuracy: An Online Tool for the Multidimensional Assessment of Decision Support Systems." In: *BIOSTEC (2)*, pp. 219–229	Presented and framed the human–AI interaction assessment tool co-developed within the research laboratory; articulated its novelty within the beyond-accuracy discourse; foregrounded interactional dimensions of evaluation.	Establishes the methodological backbone of the thesis by operationalising AI impact assessment as a multi-dimensional, interaction-aware process. Reframes evaluation from model-centred performance to situated consequences of AI use, enabling systematic comparison of benefits, risks, and reliance patterns in decision support.
Federico Cabitza et al. (2024a). "Explanations considered harmful: the impact of misleading explanations on accuracy in hybrid human-ai decision making." In: *World conference on explainable artificial intelligence*. Springer, pp. 255–269	Led the experimental work, including organising and managing experimental sessions; co-authored the full manuscript and analysed the data pertaining to human-AI interaction assessment; framed the necessity of empirical evaluation of explanations as an application of the beyond-accuracy framework rather than an a priori validation of explainability.	Challenges the assumption that explanations are inherently beneficial by empirically showing their capacity to induce over-reliance and distorted judgement. Strengthens the thesis's methodological claim that XAI must be evaluated as an interactional intervention with measurable cognitive and behavioural effects.

Table 7.7: Design-oriented empirical contributions underpinning Chapter 10

Study	Author contribution	Thesis contribution
Federico Cabitza et al. (2024b). "Never tell me the odds: Investigating pro-hoc explanations in medical decision making." In: *Artificial intelligence in medicine* 150, p. 102819	Guided the interpretation of results and discussion; led the writing of the sections on frictional AI and the design implications.	Introduces post-hoc friction as a deliberate interaction strategy to counteract uncritical acceptance of AI advice. Demonstrates how reflective delays and evaluative prompts can preserve user judgement without suppressing AI support, substantiating friction as a designable property of human–AI interaction.
Federico Cabitza et al. (2023b). "Let me think! investigating the effect of explanations feeding doubts about the AI advice." In: *International cross-domain conference for machine learning and knowledge extraction*. Springer, pp. 155–169	Co-authored the paper's Introduction, Discussion, and Conclusion to frame the study within the Reflective XAI and cognitive forcing discourses.	Operationalises reflective XAI as an interaction paradigm that promotes active reasoning rather than passive consumption of explanations. Provides empirical support for cognitive forcing functions as a means of sustaining epistemic agency in AI-assisted decision-making.
Elisa Rubegni et al. (2025). "Oracles slip on frictionless marble: The case for productive friction in AI-Supported Radiological Work." In: *Unpublished*	Designed the profiling and evaluation questionnaires; co-led experimental sessions (prototype interaction and interviews); led questionnaire analysis and thematic interview analysis; authored the manuscript and the full set of design implications.	Empirically contrasts oracular and evaluative AI interaction styles, showing that reducing decisional closure and increasing epistemic openness mitigates over-reliance. Establishes evaluative AI as a concrete instantiation of frictional interaction protocols in practice.
Giulia Anichini, Chiara Natali, and Federico Cabitza (2024). "Invisible to machines: designing AI that supports vision work in radiology." In: *Computer Supported Cooperative Work (CSCW)* 33.4, pp. 993–1036	Positioned the contribution within the CSCW literature (Section *Foundations in CSCW*); organised findings into thematic clusters; authored the entire *Design Implications* section.	Grounds frictional AI principles within socio-technical and CSCW-informed design traditions, showing through practitioner interviews how their professional trajectories can be better supported via AI systems that are *open*, *multiple* and *auxiliary*.

8

CONCEPTUALISING EMERGENCE-ORIENTED HUMAN–AI INTERACTION AND ITS IMPLICATIONS

> *The human mind isn't a computer; it cannot progress in an orderly fashion down a list of candidate moves and rank them by a score down to the hundredth of a pawn the way a chess machine does. Even the most disciplined human mind wanders in the heat of competition. This is both a weakness and a strength of human cognition. Sometimes these undisciplined wanderings only weaken your analysis. Other times they lead to inspiration, to beautiful or paradoxical moves that were not on your initial list of candidates.*
>
> — Garry Kasparov, *Deep Thinking*
>
> (Kasparov and Greengard, 2017)

8.1 INTRODUCTION

This part presents the conceptual contributions of the thesis. Drawing on a set of peer-reviewed publications, it develops an emergence-oriented account of human–AI interaction in clinical and other knowledge-intensive settings. Rather than treating the effects of AI as intrinsic properties of algorithms or explanation techniques, the works presented here collectively argue that such effects arise from how human and artificial agents are configured to interact over time.

The first contribution establishes deskilling as a structural and emergent risk of AI-supported decision-making. Synthesising evidence across a broad and heterogeneous literature, this work reframes deskilling not as an episodic side-effect of automation, but as a cumulative and often invisible consequence of repeated patterns of AI-mediated delegation. This analysis is complemented by the notion of epistemic sclerosis, which captures the infrastructural mechanisms—such as ground-truth fixation, disagreement suppression, and feedback loops between data, models, and practice—through which skill erosion can become stabilised and resistant to correction.

Building on this diagnosis, the second contribution shifts attention from individual users and systems to human–AI interaction protocols as the locus where emergent effects are configured. Through the analysis of centaur and group-based configurations, including adjunct forms of AI integration, this work shows that the same AI system can give rise to qualitatively different outcomes depending on its position, timing, and authority within the decision-making process. Interaction

protocols are thus treated as designable configurations that shape reliance, deliberation, and long-term competence.

Finally, this part situates these insights within a broader, distributed view of human–AI systems. Drawing on work on cybork configurations and dimensions of human–machine combination, it extends the emergence-oriented perspective beyond dyadic interactions, providing conceptual vocabulary to describe how agency, responsibility, and cognition are distributed across socio-technical assemblages. Together, the publications presented in this part establish the conceptual foundations of the thesis and motivate the need for interaction-sensitive methods of evaluation.

The investigation of hybrid intelligence has been the unifying thread of my research. Because hybrid intelligence is inherently multifaceted—spanning cognition, interaction design, governance, and collective practice—each of my conceptual and empirical studies has, either explicitly or implicitly, addressed this topic. Most notably, I have examined its philosophical grounding (Cabitza et al., 2025c), the conditions required for its emergence (Cabitza and Natali, 2022), the interaction protocols that enable it (Cabitza et al., 2023d; Wilson et al., 2025), and the impact of different interaction paradigms on the decision-making process of individuals and groups through empirical investigation (Natali et al., 2024).

For clarity and focus in this chapter, I consolidate these efforts into two core contributions. First, I present a conceptual and terminological framing of hybrid intelligence, drawing on insights developed through high-level theoretical work and co-authored analyses (Cabitza et al., 2025c; Wilson et al., 2025). Second, I report findings on the design of interaction protocols that can sustain hybrid intelligence, building on the frameworks proposed in Cabitza et al. (2023d) and Wilson et al. (2025), and detail the empirical evidence I collected in a user study on the impact of different interaction paradigms across individuals and groups (Natali et al., 2024), assessed through the metrics of human-AI interaction quality outlined in Cabitza et al. (2023a) and Natali, Campagner, and Cabitza (2024).

These contributions lay the conceptual and practical foundations for understanding hybrid human–AI decision-making and for designing interactions that enable hybrid intelligence to emerge in practice. Crucially, they also motivate the need to assess human–AI interaction beyond combined accuracy, recognising that hybrid intelligence—as both a goal and a design orientation—is greater than the sum of its parts.

8.2 HUMAN–AI COLLABORATION PROTOCOLS AS CONFIGURATIONS OF EMERGENCE

8.2.1 *Moving AI to the periphery: the* Human-first *approach and collective intelligence*

Bibliographic reference

Chiara Natali et al. (2024). "Humans in the Group, Computers in the Coop. Comparison of Individual and Collective Improvement in Cognitive Tasks in Adjunct AI Settings." In: *IFIP Working Conference on Human Work Interaction Design*. Springer, pp. 174–191[1]

This study provides one of the first controlled empirical examinations of technology dominance in group deliberation settings, extending a body of work that has so far largely focused on individual human–AI dyads. Technology dominance is operationalised through two complementary phenomena: automation bias, i.e. detrimental over-reliance on incorrect AI advice, and algorithmic aversion, i.e. detrimental rejection of correct AI recommendations.

Building on prior research on Human–AI Collaboration Protocols (HAI-CPs), the study examines whether the well-established benefits of the centaur model in individual settings generalise to collective decision-making, and whether group deliberation alters the role played by AI in hybrid intelligence systems. In particular, it interrogates whether AI acts as a dominant epistemic authority or as an adjunct contributor within groups, and whether collective intelligence modulates reliance patterns when compared to individual human–AI interaction.

To this end, the study contrasts individual human–AI dyads (Centaurs) with group-based configurations (Bigae), further distinguishing group protocols according to the timing of AI advice. This allows an explicit test of how interaction structure, rather than AI accuracy alone, shapes performance, bias, and trust.

8.2.1.1 *Research questions*

RQ1 How do different HAI-CPs—individual human–AI dyads (Centaurs) versus group-based human–AI configurations (Bigae)—affect decision-making performance in cognitively demanding, non-domain-specific tasks?

RQ2 Within group-based HAI-CPs, how does the timing of AI advice (AI-first vs Group-first) influence accuracy, automation bias, and algorithmic aversion?

1 Open access available at https://boa.unimib.it/handle/10281/572325

8.2.1.2 *Methods*

PARTICIPANTS Forty-six Master's students participated, forming 22 Centaur dyads and six Bigae groups. Group assignment was random. Group composition ensured gender diversity (mean per group: 2.71 women, 1.28 men), consistent with prior findings on collective intelligence, though no causal claims are made on this basis.

COLLABORATION PROTOCOLS Three HAI-CPs were experimentally compared. Protocol naming draws on Greek mythology and Roman chariot racing metaphors: AI-supported individuals are referred to as *Centaurs*, AI-supported groups as *Bigae* (two-horse chariots), and individual group members as *Aurigae* (charioteers).

Across all protocols, participants were required to first provide an independent, unaided decision before any group deliberation or exposure to AI advice. Consequently, all HAI-CPs followed a *human-first* decision structure.

In the *Centaur* protocol (1 human + AI), individual participants interacted directly with the AI system after their initial decision and produced a final AI-supported answer without any group interaction.

In the *AI-First Bigae* protocol (4 humans + AI), group members received AI advice prior to engaging in collective deliberation, after which they jointly produced a single group-AI decision.

In the *Group-First Bigae* protocol (4 humans + AI), group members first deliberated and reached an initial collective decision without AI support; only subsequently did they consult the AI and produce a final group-AI decision.

This experimental design disentangles individual decision-making, group deliberation, and AI advice, enabling both inter-protocol and intra-protocol comparisons of decision accuracy and reliance patterns across successive decision stages.

TASK AND MEASURES Participants solved 19 challenging logic puzzles (numerical, alphanumeric, deductive, graphical, and anagrams) from the website *YouMath*, each with four response options. AI advice was simulated via a Wizard-of-Oz GPT-like interface (Figure 8.1.

AI accuracy was deliberately set at 68.42% (13 correct, 6 incorrect), exceeding chance but remaining clearly fallible. Correct answers were front-loaded to establish baseline trust.

Each puzzle followed a fixed 180-second structure, detailed in Figure 8.2. All participants started with an individual phase lasting one minute, during which they were exposed to the logic puzzle and they reported their initial response. After protocol-specific deliberation and

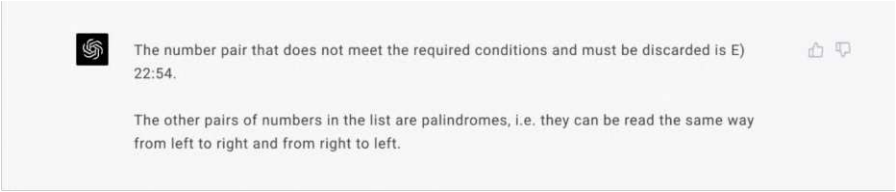


Figure 8.1: The simulated GPT-like interface reporting one of the puzzles (translated from Italian)

AI exposure, they reported their final decision (as *Centaurs* or as *Bigae*).

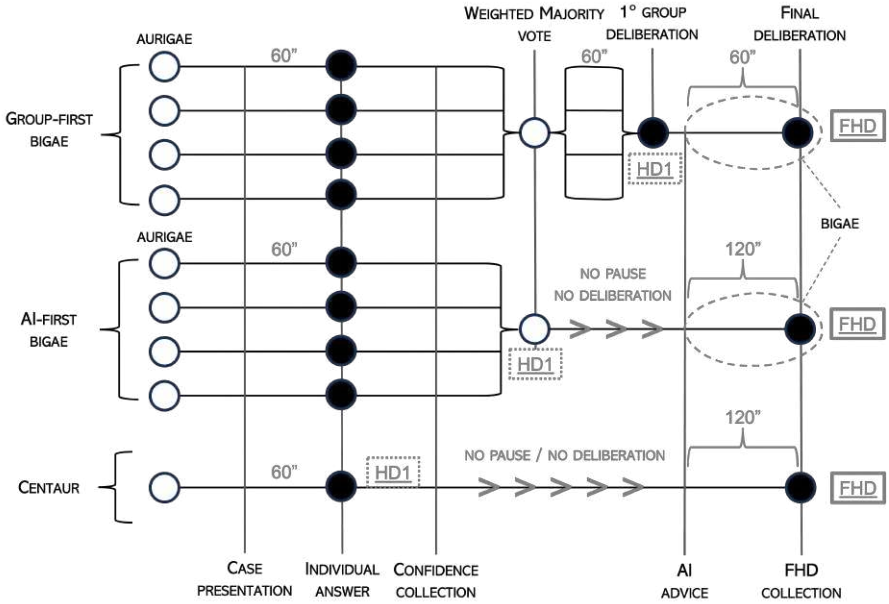


Figure 8.2: Deliberation scheme for the three experimental configurations.

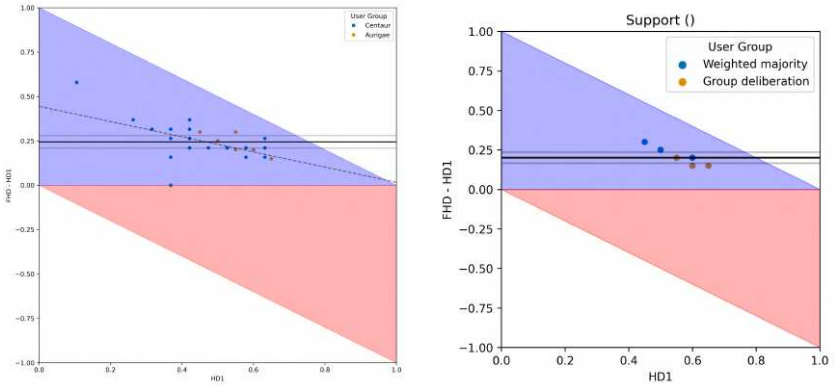
Pre- and post-experiment questionnaires assessed trust and perceived utility of AI, drawing on validated scales (AI Anxiety Scale; Technology Acceptance Model). TAM items were aggregated into a Utility score (Cronbach's = .85).

Automation and conservatism biases were quantified using reliance patterns across decision stages (initial human decision, AI advice, final decision), and data was elaborated through the *Human-AI Interaction Assessment* tool.[2]

8.2.1.3 *Results*

COLLECTIVE VS INDIVIDUAL HYBRID PERFORMANCE. Baseline group performance was below the AI: 57.89% for AI-first groups and 50% for Group-first, compared with the AI's 68.42%. After collaboration, however, both group conditions overtook the AI: AI-first

2 Human-AI Interaction tool available at https://mudilab.github.io/metimeter-frame-2025/tools.html



(a) Benefit Diagram: Aurigae vs. Centaurs

(b) Benefit Diagram: Weighted Majority vs. Group Deliberation Bigae

Figure 8.3: Diagrams generated by the *Human Interaction Metimeter Tool*

rose to 78.95% and Group-first to 75%, corresponding to gains of +21.06% and +25% over initial group performance (Figure 8.3). These gains were sizeable but, given the sample, not statistically significant over the AI baseline (AI-first: +10.53%, t(56)=1.6, p=.109; Group-first: +6.58%, t(75)=1, p=.300). The final performance of Centaurs, 69.14%, manginally and non-significantly exceeded AI accuracy (t(417)=0.3, p=.756).

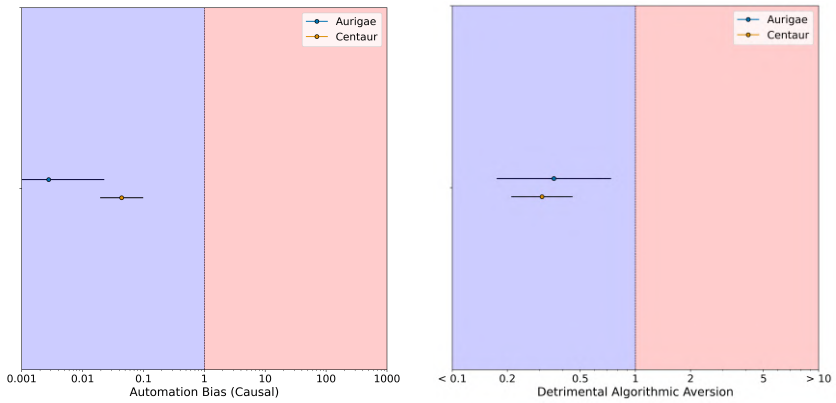
After AI exposure and deliberation, AI-First Bigae reduced errors by 58% compared to their baseline individual answers. Group-First Bigae achieved a 16% reduction in errors through human deliberation alone, followed by a further 41% reduction after AI consultation.

AI AS PERFORMANCE LEVELER. Across both individuals and groups, AI acted as a performance leveler. Lower initial performers benefitted disproportionately from AI support, reducing overall variance in performance. This effect was confirmed by a significant negative correlation between initial performance and improvement (r = 0.55; p = .002; slope -0.43).

AUTOMATION BIAS, CONSERVATISM BIAS, TRUST As shown in Figure 8.4, Automation bias was significantly lower in groups (Bigae) than in individuals (Centaurs). Instead, Conservatism bias showed no significant mean differences but greater variance in groups, indicating heterogeneous strategies of AI acceptance.

In particular, Group-First Bigae exhibited lower automation bias and higher conservatism bias than AI-First Bigae, indicating more critical engagement with AI advice when deliberation preceded exposure.

Trust in AI increased post-experiment in all conditions, but the increase was significantly larger in group settings (p = .028, d = .31).



(a) Detrimental Algorithmic Aversion (Conservatism Bias): Bigae vs. Centaurs

(b) Automation bias: Bigae vs. Centaurs

Figure 8.4: Cognitive impact diagram

Although Group-First Bigae ended with slightly lower final accuracy (3.95%), they relied less on AI, supporting the interpretation of AI as an adjunct rather than dominant contributor in this configuration.

8.2.1.4 *Limitations*

The study's strength—its highly controlled design—is also its primary limitation. Tasks were abstract logic puzzles, the AI support was simulated through a *Wizard-of-Oz*, and collaboration protocols were tightly constrained. While these choices were necessary to isolate reliance dynamics, they limit direct generalisation to real-world professional settings.

The front-loaded sequence of correct AI answers may have influenced trust formation, though this strategy is well-documented in prior XAI research and was applied consistently across conditions to enable comparison.

Finally, the AI was intentionally treated as a fallible advisor rather than a high-performing expert system. This choice prioritised the study of interaction dynamics over raw AI capability.

8.2.2 Further work: *The socio-technical dimensions of human-AI combination and emergence-oriented design*

Bibliographic references

Wilson, B., Natali, C., Roach, M., Scott, D., Rahat, A., Rawlinson, D., & Cabitza, F. (2025). "Dimensions of human-machine combination: prompting the development of deployable intelligent decision systems for situated clinical contexts." In: *Computer Supported Cooperative Work (CSCW)*, pp. 1–57[3]

Cabitza, F., Natali, C., Varanini, F., & Gunkel, D. (2025). "Beyond cyborgs: the cybork idea for the de-individuation of (artificial) intelligence and an emergence-oriented design." In: *AI & SOCIETY* 40.5, pp. 3333–3348[4]

In the paper *Dimensions of Human–Machine Combinations* (Wilson et al., 2025), we advanced a fundamentally interactional account of human–AI collaboration in clinical decision-making. The central claim of the paper is that the persistent "AI chasm" between promising algorithmic performance and real-world clinical benefit cannot be explained by model limitations alone, but by a systematic failure to specify, design, and evaluate how humans and algorithmic systems are combined in practice.

Rather than treating AI as a stand-alone decision aid whose effects can be inferred from bench performance, the paper reframes clinical AI as a sociotechnical intervention, whose success depends on eight contextual dimensions that determine how AI actually combines with clinicians in practice: (1) who participates in the decision episode (participating agents); (2) who retains authority or must merely receive AI output (control relations); (3) how much of the clinician's task the AI really touches (task overlap); (4) when, and in what order, human and AI contributions occur (temporal patterning); (5) how close each agent is to the information needed (informational proximity); (6) how much of that information is shared across agents (informational overlap); (7) whether humans can shape inputs, intentionally or otherwise (input influence); and (8) what forms of output the AI is allowed to emit, beyond "alert/flag" (output representation coverage). An overview of the dimensions is presented in Table 8.1

A review of the literature makes under the lenses of these dimensions makes visible that most clinical AI studies still assume a 1-human/1-AI dyad, human-as-controller, single-touch, low-overlap setting, and thereby under-specify the very interactional space in which automation bias, over-reliance, or distributive effects actually

3 Open access version available at https://link.springer.com/article/10.1007/s10606-025-09514-4 or https://boa.unimib.it/handle/10281/559542

4 Open access version available at https://link.springer.com/article/10.1007/s00146-025-02191-3 and https://boa.unimib.it/handle/10281/559543

arise.

Having observed the emergent effects from protocols in the previous study, this framework turns the sequence-based (Human-first, AI-first) human-AI interaction protocol into a fully granular, multi-dimensional specification framework.

In (Cabitza et al., 2025c), we reframed the question of where intelligence operates in human–AI systems, conceptualising intelligence as an emergent property of situated work practices, terming *cybork* the coordinated activity of humans and technologies. Building on cybernetic notions of feedback and regulation, and on socio-technical traditions that view work as irreducibly human–technical, the cybork departs from system metaphors that presuppose separable social and technical components. Instead, it emphasises their co-constitutive entanglement in practice.

From the perspective of extended and distributed cognition, the cybork treats technological artefacts not as external aids but as integral elements of expanded cognitive systems, shaping how perception, reasoning, and decision-making unfold. Intelligence, in this view, is neither embedded in algorithms nor merely augmented in users, but emerges within shared spaces of action—akin to Nonaka's *Ba* (Nonaka and Konno, 1998)—where human judgement, computational mediation, and collective knowledge continuously interact. This de-individuated understanding of intelligence motivates a corresponding shift in socio-technical design: rather than specifying components, roles, or interaction sequences, designers are called to adopt emergence-oriented approaches that cultivate the conditions under which effective human–AI collaboration can form, stabilise, and evolve through use.

Conceptually, the cybork extends and problematises the cyborg metaphor. While the cyborg foregrounds the augmentation of an individual organism through technological prostheses, it retains a residual commitment to the individual as the primary locus of agency and cognition. The cybork, by contrast, shifts the unit of analysis from the individual to collective, situated activity. What matters is not who or what acts, but how action is organised, sustained, and made meaningful within a socio-material practice.

The AI-assisted medical équipe exemplifies this perspective. In such settings, clinicians and AI systems form a cybork insofar as diagnostic work is accomplished through their entangled contributions. The AI does not "support" the clinician in isolation, nor does the clinician simply "oversee" the AI. Instead, diagnostic judgement emerges from the coupled performance of data-driven pattern recognition and hu-

Table 8.1: Dimensions of human–machine combination in clinical AI

Dimension	What it asks (design / evaluation prompt)	Typical positions in current clinical AI
Participating agents	Who actually takes part in the decision episode (how many humans, how many AI/algorithmic agents)?	Almost all studies are 1H–1AI dyads; ∼60–70 situated studies use this setting; no genuine multi-human–multi-AI arrangements.
Control relations	Who is meant to be in charge of resolution — does the human authorise/override, act as peer, or merely receive AI instructions?	Dominant pattern: human-as-controller (human in command, AI advisory). Some human-as-recipient in automated screening (ophthalmology). Virtually no human–AI peer relations are modelled.
Task overlap	How much of the human task-set the AI actually helps with in that episode — is it a narrow adjunct or a broad co-worker?	Most systems provide narrow, local help (lesion marking, single-disease flag, one-step triage), i.e. low–moderate overlap between human and AI task-space.
Temporal patterning	When and how often human(s) and AI meet; in what order contributions occur (continuous, episodic, event-triggered).	Mostly single-touch / event-triggered interactions; AI often speaks first (AI-first CDS alert, AI-first image labelling); the ordering of contributions is rarely discussed explicitly.
Informational proximity	How easy it is, in that setting, for an agent to retrieve additional relevant information (labs, history, images, notes).	Often low–medium: the AI works off a curated stream (image, waveform) while clinicians have EHR/use-context; the two information spaces are not aligned.
Informational overlap	How much of the same information human(s) and AI see and can reference explicitly.	Frequently low overlap: AI sees latent signals the human cannot; human sees social/contextual/EHR the AI cannot; this asymmetry is seldom made visible to users.
Input influence	Whether the human(s) can shape what the AI takes as input — intentionally (add/remove features, correct metadata) or unintentionally (phrasing, order of entry).	Most clinical AI is no-input-influence: fixed input fields, fixed image, fixed pre-processing; human cannot steer. Some settings show only unintentional influence (e.g. dispatch calls).
Output representation coverage	What kinds of outputs the AI is allowed to produce — only positive hits, or also negatives, abstentions, and uncertainty.	Predominantly positive-only escalation ("alert", "refer", "flag"). Very few systems emit safe-to-ignore, inconclusive, or explanation-only messages.

man interpretive, ethical, and experiential reasoning. What counts as a correct, acceptable, or trustworthy decision is negotiated within the practice itself, through the ongoing enactment of collaboration protocols.

This conceptual move has direct implications for how human–AI collaboration protocols are understood. Rather than being predefined rules that govern interaction between separate agents, collaboration protocols are reinterpreted as emergent patterns of coordination within cybork practices. In this sense, the cybork functions less as a model to be implemented and more as an orienting concept. It directs analytical and design attention away from components, roles, and functions, and towards relationships, practices, and trajectories of use. By doing so, it provides the conceptual grounding needed to study human–AI collaboration protocols not as static interaction patterns, but as evolving achievements of collective work.

.

8.3 DESKILLING AS AN EMERGENT STRUCTURAL RISK OF AI-SUPPORTED DECISION-MAKING

8.3.1 *AI-induced deskilling and upskilling inhibition*

Bibliographic reference
Natali, C., Marconi, L., Dias Duran, L. D., & Cabitza, F. (2025). AI-induced deskilling in medicine: a mixed-method review and research agenda for healthcare and beyond. *Artificial Intelligence Review*, 58(11), 356.[5]

This literature review examines two interrelated but conceptually distinct threats arising from the integration of AI into clinical practice: deskilling and upskilling inhibition. Deskilling refers to the degradation of previously acquired clinical competencies due to reduced practice or technological substitution, whereas upskilling inhibition denotes the suppression of opportunities to acquire new or advanced skills—particularly among trainees—due to systematic over-reliance on AI systems.

In clinical contexts, deskilling manifests as a gradual shift from hands-on, experience-driven diagnosis and treatment towards supervisory validation of algorithmic outputs. This transition risks eroding not only technical and procedural skills, but also the tacit judgment, interpretive sensitivity, and professional confidence underpinning safe medical practice. Upskilling inhibition, by contrast, disrupts the developmental trajectory of medical training: when AI systems consistently resolve diagnostic or interpretive tasks, novices are deprived of the progressive challenge and responsibility necessary to cultivate diagnostic reasoning, autonomy, and confidence.

Despite increasing concern, empirical understanding of how, where, and for whom AI-induced deskilling and upskilling inhibition occur remains fragmented. These phenomena are gradual, often imperceptible, and disproportionately affect forms of expertise—such as tacit knowledge, heuristics, and interpersonal competence—that resist straightforward quantification. This study addresses this gap by systematically mapping the manifestations, mechanisms, and implications of AI-induced deskilling and upskilling inhibition in medicine.

8.3.1.1 *Research questions*

RQ1 Which clinical competencies are most frequently identified as vulnerable to AI-induced deskilling in the medical literature?

5 Open access available at https://link.springer.com/article/10.1007/s10462-025-11352-1 or https://boa.unimib.it/handle/10281/565822

RQ2 How does AI adoption risk inhibiting skill acquisition and professional development, particularly among trainees and early-career clinicians?

RQ3 What socio-technical, organisational, and design mechanisms are implicated in these processes, and how might they be mitigated?

8.3.1.2 *Methods*

A mixed-method review design was adopted, combining a systematic literature review with a narrative synthesis.

The systematic review followed PRISMA guidelines (Page et al., 2021) and was anchored in the PACES-MRCP(UK) framework (Ghafur et al., 2017), a widely accepted clinical competency model used to assess core medical skills, including physical examination, clinical reasoning, communication, and judgment. PACES was employed as an empirically grounded proxy to classify deskilling concerns according to established clinical standards. Searches across *PubMed*, *Scopus*, and *Web of Science* yielded a final corpus of 22 peer-reviewed studies, prioritised for depth and conceptual relevance, following the inclusion and exclusion criteria detailed in Table 8.3.

Criteria	**Description**
Inclusion (IC1)	Studies completely written in English.
Inclusion (IC2)	Studies with a central focus on healthcare or medicine and artificial intelligence.
Inclusion (IC3)	Studies directly and significantly dealing with the theme of deskilling.
Exclusion (EC1)	Papers merely mentioning the notion of deskilling without elaboration.
Exclusion (EC2)	Papers representing commentaries without specific relevance.
Exclusion (EC3)	Papers outside the specific domain of AI in healthcare or medicine.

Table 8.3: Inclusion and exclusion criteria for the systematic review

Complementing this, a narrative review of 62 additional sources explored broader implications for Human–AI Interaction, organisational skill dynamics, training, and ethics.

These sources were analysed using qualitative thematic synthesis and organised within a two-dimensional categorisation framework comprising:

- Human–AI Interaction
  - Impact of AI on human skills in organisations
  - Deskilling types and AI influence on clinical decision-making
  - Physician expectations and deskilling concerns
  - Patient-centred perspectives and safety
- AI Adoption and Integration
  - Incorporation of domain knowledge in AI models
  - Training and education
  - Ethical and legal implications

This hybrid approach enabled fine-grained mapping of clinical skill erosion alongside systemic, organisational, and ethical dynamics.

8.3.1.3 *Results*

SYSTEMATIC REVIEW FINDINGS In addition to PACES-aligned competencies, two further categories emerged. The first is *organisational vulnerabilities*, including reduced team situational awareness, fragile workflows, and loss of collective expertise. *AI-specific risks* described the emerging phenomena of as automation bias, inability to contest algorithmic outputs, and heightened risk of error during system failures.

The findings of the systematic review are detailed in Tables **??**.

Across the surveyed literature, concerns cluster around a progressive weakening of core clinical skills, relational practices, and collective resilience in AI-supported care. At the level of physical examination and procedural competence, multiple studies report deterioration of bedside assessment and manual skills as clinicians increasingly rely on algorithmic outputs rather than direct patient engagement (Hallowell et al., 2023; Levy, Jotkowitz, and Chowers, 2019; Lu, 2016; Monteith et al., 2022; Rafner et al., 2022; Ruskin et al., 2020). This effect is particularly pronounced for trainees (Aquino et al., 2023; Chen et al., 2021; Monteith et al., 2022), for whom automation of routine tasks reduces exposure to baseline cases and constrains opportunities for skill acquisition.

In the domain of clinical communication, the literature consistently points to a degradation of doctor–patient interaction (Aquino et al., 2023; Levy, Jotkowitz, and Chowers, 2019; Lu, 2016; Parchmann et al., 2024). Attention diverted towards interfaces and system outputs is associated with reduced empathic engagement, weakened therapeutic relationships, and diminished capacity to manage patient concerns, raising broader questions about the preservation of relational and interpretive aspects of care (Akudjedu et al., 2023; Dias Duran, 2021; Lu, 2016; Rafner et al., 2022).

Concerns intensify around diagnostic reasoning and clinical judgement (Cabitza, Rasoini, and Gensini, 2017; Kashou et al., 2024; Rafner et al., 2022; Smith and Baumann, 2020). Habitual deference to AI recommendations is linked to reduced diagnostic accuracy, erosion of clinical knowledge, narrowing of independent assessment, and declining confidence in unaided judgement, suggesting a shift from active sense-making to confirmatory oversight (Cabitza, Rasoini, and Gensini, 2017; Dias Duran, 2021; Monteith et al., 2022; Rafner et al., 2022).

Beyond individual competencies, the review identifies risks to patient welfare, encompassing both safety and moral dimensions (Dias Duran, 2021; Hallowell et al., 2023; Parchmann et al., 2024; Stogiannos et al., 2025). These include inattentive acceptance of AI outputs, diminished ethical sensitivity, and threats to beneficence, dignity, and trust—particularly when clinicians' moral agency is attenuated by automated decision pathways.

At the organisational level, systemic vulnerabilities emerge. Increased dependence on AI is associated with loss of collective expertise, reduced team situational awareness, and fragile workflows that are difficult to sustain or recover when systems fail. These dynamics extend to team cognition, where coordination and shared understanding deteriorate as responsibility and attention are redistributed across human–AI assemblages (Monteith et al., 2022; Rafner et al., 2022; Smith and Baumann, 2020; Sparrow and Hatherley, 2019).

Finally, the review highlights a set of AI-specific epistemic and interactional risks. These include automation bias, difficulties in understanding or contesting algorithmic outputs, reduced motivation for in-depth learning, and errors arising from poorly aligned human–AI interaction (Akudjedu et al., 2023; Choudhury and Chaudhry, 2024; Kashou et al., 2024; Lu, 2016; Panesar et al., 2020; Ruskin et al., 2020).

NARRATIVE REVIEW FINDINGS The narrative synthesis contextualised these findings within broader socio-technical dynamics, reported in table 8.10.

Widespread over-reliance on AI is consistently associated with automation bias and erosion of critical thinking, as habitual deferral to algorithmic outputs narrows diagnostic reasoning and weakens independent judgement (Banerjee et al., 2021; Cabitza, Rasoini, and Gensini, 2017; Cabitza, 2021b; Campbell et al., 2020; Green, 2019; Kapoor, Walters, and Al-Aswad, 2019; Lu, 2016; Monteith et al., 2022; Wessel, 2023). Closely related is a broader transformation of professional roles, in which clinicians are repositioned from primary decision-makers to supervisors of automated systems, with attendant losses of autonomy and devaluation of tacit, experience-based knowledge (Aslam and Hoyle, 2022; Kundu, 2021; Mosch et al., 2022; Rafner et al., 2022; Sambasivan and Veeraraghavan, 2022).

Table 8.4: Concerns related to *physical examination* and *clinical communication* in AI-supported care.

Clinical skill	Concern	Design implications	References
Physical examination / identifying physical signs	Deterioration of physical exam skills when clinicians rely on AI outputs instead of bedside assessment.	Protect dedicated exam time; "AI-off" drills; require documented physical findings before consulting AI; periodic competency checks tied to supervision.	Hallowell et al. 2023; Levy, Jotkowitz, and Chowers 2019; Lu 2016; Monteith et al. 2022; Rafner et al. 2022; Ruskin et al. 2020
Physical examination / procedural skills	Deterioration or loss of manual skills with increased automation.	Rotation quotas for manual procedures; simulation refreshers; failure-mode training; fallback protocols for system outages.	Duran et al. 2023; Hallowell et al. 2023; Rafner et al. 2022; Ruskin et al. 2020; Smith and Baumann 2020
Skill development (trainees)	Reduced opportunities to develop baseline competencies as routine tasks are automated.	Case-allocation rules that guarantee trainee exposure; staged autonomy; commit-before-reveal interactions for trainees.	Aquino et al. 2023; Chen et al. 2021; Monteith et al. 2022
Clinical communication / managing concerns	Worsening of doctor–patient communication as attention shifts to system interfaces.	Interface minimalism during encounters; "patient-first" interaction checklists; scribing support that preserves eye contact and listening.	Aquino et al. 2023; Levy, Jotkowitz, and Chowers 2019; Lu 2016; Parchmann et al. 2024
Clinical communication / relationship	Deterioration of the physician–patient relationship and empathic engagement.	Training on relational work with AI; shared decision-making prompts; transparency scripts about AI's role and limits.	**akudjedu2023**; **duran2021**; Levy, Jotkowitz, and Chowers 2019; Lu 2016; Rafner et al. 2022

Table 8.6: Concerns related to *differential diagnosis, clinical judgement, patient welfare, organisational* and *AI-specific* risks, identified in the Systematic Review.

Clinical skill / area	**Concern**	**Design implications**	**References**
Differential diagnosis	Decrease in diagnostic accuracy with habitual deference to algorithmic suggestions.	Commit-before-reveal; second-reader or adversarial checks; uncertainty exposure; require rationale alignment before accepting AI output.	Cabitza, Rasoini, and Gensini 2017; Kashou et al. 2024; Rafner et al. 2022; Smith and Baumann 2020
Differential diagnosis (knowledge base)	Erosion of clinical knowledge and interpretive skill when AI pre-processes signs and tests.	Protected reasoning time; forced consideration of alternatives; periodic "AI-absent" case conferences.	**hallowell2023**; Choudhury and Chaudhry 2024; Koplin et al. 2025; Levy, Jotkowitz, and Chowers 2019; Nakagawa et al. 2023; Rafner et al. 2022; Sparrow and Hatherley 2019
Clinical judgement	Poorer clinical reasoning and judgement; narrowing of independent assessment.	Reflection prompts; justification capture; graded autonomy; audit of override/accept rates.	Aquino et al. 2023; Cabitza, Rasoini, and Gensini 2017; Koplin et al. 2025; Levy, Jotkowitz, and Chowers 2019; Lu 2016; Parchmann et al. 2024; Rafner et al. 2022
Clinical judgement (confidence)	Unwillingness to provide a definitive clinical assessment unaided.	Confidence-building via supervised independent reads; "speak-first" rounds; mentorship on warranted dissent.	Cabitza, Rasoini, and Gensini 2017; Dias Duran 2021; Monteith et al. 2022; Rafner et al. 2022
Maintaining patient welfare	Deterioration of moral/ethical skills; risks to beneficence and dignity.	Ethics case reviews; patient-centred KPI dashboards; explicit harm–benefit checks when adopting AI.	Dias Duran 2021; Hallowell et al. 2023; Parchmann et al. 2024; Stogiannos et al. 2025

Table 8.8: Concerns related to *patient welfare*, *organisational resilience* and *cognition* risks, identified in the Systematic Review.

Clinical skill / area	**Concern**	**Design implications**	**References**
Maintaining patient welfare (safety)	Undermining patient safety through inattentive deference to AI.	Safety huddles focused on AI failure modes; red-team testing; escalation triggers for atypical cases.	Levy, Jotkowitz, and Chowers 2019; Rafner et al. 2022
Organisational resilience	System fragility / loss of collective expertise when AI becomes irreplaceable.	Continuity plans for AI outage; cross-skilling; preserve manual pathways; regular resilience exercises.	Monteith et al. 2022; Rafner et al. 2022; Smith and Baumann 2020; Sparrow and Hatherley 2019
Team cognition	Reduced situational awareness and teamwork quality.	Shared displays of model limits; team-based verification; role clarity for human/AI responsibilities.	Panesar et al. 2020; Smith and Baumann 2020
AI–HCI risks	Errors arising from human–AI interaction and interface design.	Human-factors evaluation; alert load governance; usability testing with failure injection.	Panesar et al. 2020; Ruskin et al. 2020
Dependence / motivation to learn	Increased dependence on AI; reduced motivation for in-depth learning.	Learning goals tied to independent cases; spaced retrieval; periodic AI-free weeks.	Akudjedu et al. 2023; Chen et al. 2021; Choudhury and Chaudhry 2024; Nakagawa et al. 2023; Ruskin et al. 2020
Epistemic critique	Inability to understand and challenge AI outputs.	Model factsheets; uncertainty displays; calibration training; mandatory challenge–response steps.	Akudjedu et al. 2023; Choudhury and Chaudhry 2024; Kashou et al. 2024; Lu 2016; Panesar et al. 2020; Ruskin et al. 2020

A third theme concerns education and training, where the literature recognises the necessity of AI literacy but repeatedly warns that poorly balanced curricular reforms risk displacing foundational clinical competencies rather than complementing them (Banerjee et al., 2021; Lu, 2016; Rao, 2023; Vallor, 2015; Zhang et al., 2023; Zulkipli, Alam, and Lim, 2023). Ethical analyses add a further layer, highlighting forms of moral and ethical deskilling: diminished ethical sensitivity, blurred accountability, and emerging responsibility gaps when decision authority is partially delegated to opaque systems (Da Silva et al., 2022; Dias Duran, 2021; Drabiak et al., 2023; Gerke, Minssen, and Cohen, 2020; Hallowell et al., 2023; Iqbal et al., 2022; Vallor, 2015).

At the system level, the review identifies growing socio-technical vulnerabilities, including technical dependency, reduced capacity to verify or override AI failures, and increased organisational fragility with potential legal and societal consequences (Aquino et al., 2023; Cabitza, Rasoini, and Gensini, 2017; Drabiak et al., 2023; Hoff, 2011; Morley et al., 2020; Panesar et al., 2020; Sparrow and Hatherley, 2019; Tsai, Fridsma, and Gatti, 2003). Against this predominantly cautionary backdrop, the literature also articulates a countervailing theme centred on human–AI synergy (Aslam and Hoyle, 2022; Kundu, 2021; Mofatteh, 2021; Nelson et al., 2020; Rafner et al., 2022). Under the banner of hybrid intelligence, several contributions point to the possibility of skill preservation or enhancement when human–AI collaboration is deliberately structured to maintain human primacy in judgement and to support reflective, resilient practice rather than passive reliance (Aslam and Hoyle, 2022; Kundu, 2021; Mofatteh, 2021; Nelson et al., 2020; Rafner et al., 2022).

8.3.1.4 *Limitations*

Several limitations must be acknowledged. First, the PACES framework, while clinically grounded, was developed for structured assessment environments and may not fully capture the complexity of real-world clinical expertise. It was adopted as a pragmatic proxy rather than a definitive representation of professional competence.

Second, literature-based synthesis is inherently constrained by publication bias and the rapid pace of AI deployment, which may outstrip formal evaluation. The reviewed literature disproportionately reflects speculative concerns and early-stage implementations, underscoring the need for longitudinal, in-situ, and mixed-methods research.

Finally, the scarcity of empirical measures distinguishing skill erosion from inhibited skill acquisition remains a major methodological gap. Addressing this gap is essential for moving beyond descriptive concern towards actionable governance and design strategies.

8.3.1.5 *Design implications and practical interventions*

We proposed a research agenda (Tables 8.12 and 8.14) that includes proposed countermeasures against these risks.

Drawing on the mixed-method review, the research agenda frames deskilling not as an inevitable by-product of AI, but as a contingent risk that can be systematically studied and actively mitigated through design, evaluation, education, and governance articulates. This requires a shift away from narrow, performance-centric evaluations of AI towards a more skill-sensitive and longitudinal perspective. Conceptually, it first calls for sharper distinctions between deskilling, upskilling inhibition, and legitimate forms of skill transformation, in order to avoid conflating the loss of existing expertise with changes in professional roles or the emergence of new competencies. Empirically, it emphasises the need for real-time and long-term studies that trace how sustained human–AI interaction reshapes judgement, competence, and outcomes over time, rather than relying on short-term accuracy gains.

Methodologically, the agenda foregrounds qualitative and mixed-methods approaches to capture dimensions that quantitative metrics routinely miss, such as professional identity, trust, autonomy, and safety culture. In parallel, it explicitly addresses publication bias by encouraging the reporting of negative findings and by broadening evaluation criteria beyond accuracy to include indicators of cognitive engagement, skill retention, and professional agency. From a design perspective, it advocates for hybrid-intelligence approaches in which AI systems are deliberately engineered to keep practitioners cognitively involved—through context-aware, frictional, and protocol-driven interactions—rather than positioning humans as passive validators of automated outputs.

Beyond system design, the agenda stresses the importance of interdisciplinary collaboration, with clinicians involved early and continuously in AI development to ensure alignment with practice and training needs. Educational initiatives are positioned as a core mitigation strategy, calling for AI literacy to be embedded in professional curricula alongside the cultivation of non-automatable skills such as ethical judgement, empathy, and complex sense-making. Finally, it argues for cross-domain learning and stronger policy and governance frameworks, drawing lessons from other high-stakes sectors to develop evidence-based safeguards that protect professional autonomy, accountability, and ethical deployment.

Table 8.10: Salient cross-cutting themes emerging from the narrative review, and associated concerns or opportunities with representative references.

THEME	CONCERNS & OPPORTUNITIES	REFERENCES
Over-reliance and Critical Thinking Erosion	Automation bias, diminished clinical judgement, and reduced diagnostic reasoning due to habitual deferral to AI outputs.	Banerjee et al., 2021; Cabitza, Rasoini, and Gensini, 2017; Cabitza, 2021b; Campbell et al., 2020; Green, 2019; Kapoor, Walters, and Al-Aswad, 2019; Lu, 2016; Monteith et al., 2022; Wessel, 2023
Professional Role Transformation	Shift from clinician to supervisor; redefinition of professional identity; loss of autonomy; devaluation of tacit knowledge.	Aslam and Hoyle, 2022; Kundu, 2021; Mosch et al., 2022; Rafner et al., 2022; Sambasivan and Veeraraghavan, 2022
Training and Education Challenges	Need to revise curricula; ensure skill preservation; promote AI literacy while maintaining human competencies.	Banerjee et al., 2021; Lu, 2016; Rao, 2023; Vallor, 2015; Zhang et al., 2023; Zulkipli, Alam, and Lim, 2023
Ethical and Moral Deskilling	Decline in ethical sensitivity and moral judgement; challenges to human accountability and responsibility gaps.	Da Silva et al., 2022; Dias Duran, 2021; Drabiak et al., 2023; Gerke, Minssen, and Cohen, 2020; Hallowell et al., 2023; Iqbal et al., 2022; Vallor, 2015
Socio-Technical Vulnerabilities	Increased system fragility; technical dependency; inability to verify or override AI failures; broader societal and legal consequences.	Aquino et al., 2023; Cabitza, Rasoini, and Gensini, 2017; Drabiak et al., 2023; Hoff, 2011; Morley et al., 2020; Panesar et al., 2020; Sparrow and Hatherley, 2019; Tsai, Fridsma, and Gatti, 2003
Human–AI Synergy and Mitigation Potential	Recognition of hybrid intelligence; potential for skill enhancement and resilience through thoughtful integration and role delineation.	Aslam and Hoyle, 2022; Kundu, 2021; Mofatteh, 2021; Nelson et al., 2020; Rafner et al., 2022

Table 8.12: Research agenda on *AI deskilling* evaluation and mitigation.

Research domain	**Key focus areas and strategies**	**References**
Conceptual clarification	Distinguishing AI-induced deskilling, upskilling inhibition, and re-skilling phenomena; clarifying skill transformation versus outright loss or inhibited skill acquisition.	Cabitza 2021a; Dias Duran 2021; Green 2019; Rafner et al. 2022; Vallor 2015; Wessel 2023; Winter and Carusi 2022
Empirical investigation	Real-time monitoring and longitudinal studies to detect practitioner–AI interaction shifts; evaluation of long-term effects on competencies, judgement quality, and outcomes.	Campbell et al. 2020; Tsai, Fridsma, and Gatti 2003
Qualitative and mixed-methods approaches	Interviews, focus groups, and case studies combined with quantitative measures to capture professional identity, job satisfaction, trust, and safety implications.	Aquino et al. 2023; Banerjee et al. 2021; Bunch, Jones, and Psirides 2023; Kundu 2021; Lennartz et al. 2021; Maassen et al. 2021; Mosch et al. 2022; Nelson et al. 2020; Winter and Carusi 2022
Mitigating publication bias	Explicit reporting of failures and negative results; evaluation metrics beyond accuracy, including skill retention, cognitive engagement, and professional autonomy.	Dickersin et al. 1987; Natali, Campagner, and Cabitza 2024
Hybrid intelligence and AI design	Design of AI systems that actively engage practitioners through context-aware and frictional interactions; explicit collaboration protocols guiding effective human–AI decision-making.	Cabitza, Campagner, and Sconfienza 2021; Cabitza et al. 2025b, 2024b; Dellermann et al. 2019; Miller 2023; Schemmer, Kühl, and Satzger 2021

Table 8.14: Research agenda on *AI deskilling* evaluation and mitigation.

Research domain	**Key focus areas and strategies**	**References**
Interdisciplinary collaboration	Early and sustained clinician involvement in AI development; interdisciplinary feedback loops to iteratively refine system behaviour and integration practices.	Iqbal et al. 2022; Mosch et al. 2022; Sambasivan and Veeraraghavan 2022
Educational initiatives and complementary human-centric skills	Integration of AI literacy into professional education and continuing development; cultivation of non-automatable skills such as ethical judgement, empathy, and complex sensemaking.	Aslam and Hoyle 2022; Kundu 2021; Rao 2023; Vallor 2015; Zhang et al. 2023
Cross-domain insights	Adaptation of evaluation strategies and governance principles from high-stakes domains such as healthcare, law, finance, aviation, and military decision-making.	Amer, Hilmi, and El Kezazy 2024; Browning 2024; Carrel 2018; Golfetti, Napoletano, and Cichomska 2021; Odonkor et al. 2024; Ruan 2020; Talib et al. 2025; Vallor 2013
Policy and governance	Evidence-based policies addressing AI-induced skill dynamics; safeguards for professional autonomy, accountability, and ethical deployment in practice.	Drabiak et al. 2023; Lu 2016

8.3.2 *Further work: Epistemic sclerosis and organisational brittleness*

Bibliographic reference

Natali, C & Cabitza, F. (2025). Make Some Noise for Ground Truthing! Frictional design against epistemic sclerosis in Decision Support Systems. *SJIS Preprints* (Forthcoming). 15. https://aisel.aisnet.org/sjis_preprints/15 [6]

This contribution is a speculative position paper on the limits of current ground truthing practices (ie., the work of classifying and labelling data for the creation of a ground truth for ML model training). Medical illnesses are not singular, stable entities. Drawing on Mol 2002's notion of the body multiple, diseases unfold differently across clinical practices, diagnostic tools, and interpretive contexts, meaning that classification is always contingent and negotiated rather than fixed. Through a narrative survey of the Information System and CSCW literature, we show how medical labelling is an inherently uncertain and interpretative rather than a direct reflection of reality. This multiplicity becomes suppressed in current machine learning pipelines, where ground truth is treated as a singular, authoritative fact, rather than an artefact of human interpretative labour. Moreover, the validity of ground truth suffers from lack of situatedness, as annotation is frequently performed in isolation, requiring experts to classify static images without the critical contextual information (e.g., prior patient images) that would be essential in real-world clinical decision-making–an action that would amount to "a genuine act of malpractice" in clinical reality.

While this paper explores a broader design space in the topic of ground truthing, I focus on its discussion of epistemic sclerosis (inspired by a previous work by Cabitza 2021a) insofar as it explains the durability and institutionalisation of deskilling.

This study advances the concept of epistemic sclerosis to describe a distinctive failure mode of AI-supported decision making in medicine, whereby expert knowledge becomes progressively rigid, self-referential, and resistant to revision. Rather than emerging from a single design flaw, epistemic sclerosis is shown to arise from a cyclical socio-technical process that begins well before deployment, at the level of ground truth construction, and is subsequently reinforced through routine clinical use of AI systems.

The analysis starts from a critical observation: contemporary medical AI systems are trained on labels produced by experts through

6 Open access available at https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1014&context=sjis_preprints

ground truthing practices that are widely treated as technical preliminaries rather than epistemic acts. Once encoded into datasets, these labels acquire infrastructural authority, functioning as if they were objective facts rather than historically situated, negotiated, and often contested interpretations. In standard machine-learning pipelines, any inter-rater disagreement encountered during this process is typically framed as annotation noise to be eliminated through consensus mechanisms such as majority voting or adjudication. While operationally convenient, this approach systematically suppresses epistemic ambiguity, collapsing multiple legitimate clinical perspectives into a single, supposedly definitive "gold standard".

Epistemic sclerosis emerges precisely at this point of collapse. In high-stakes medical contexts, variation in expert judgement is rarely reducible to error; instead, it often reflects classificatory inadequacies, ontological uncertainty, or the intrinsic ambiguity of complex clinical phenomena. By treating disagreement as *noise* (Kahneman, Sibony, and Sunstein, 2021) rather than an epistemic signal, dominant ground truthing practices convert what should be a site of ongoing knowledge production into a mechanism of knowledge fixation. Labels crystallise prematurely, embedding historically contingent interpretations into datasets that subsequently anchor model behaviour.

Once deployed, AI systems trained on these frozen representations enter clinical practice as decision support tools. Here, a reinforcing feedback loop takes hold. Machine-generated outputs—derived from past labels—are increasingly accepted as authoritative, subtly reshaping clinicians' expectations and interpretative habits. Over time, practitioners are exposed less frequently to epistemic friction, fewer borderline cases are actively debated, and opportunities for reflective reinterpretation diminish. Knowledge thus circulates in a closed loop: expert judgements inform labels; labels train models; models stabilise and reproduce those same judgements in practice. This loop constitutes the core dynamic of epistemic sclerosis.

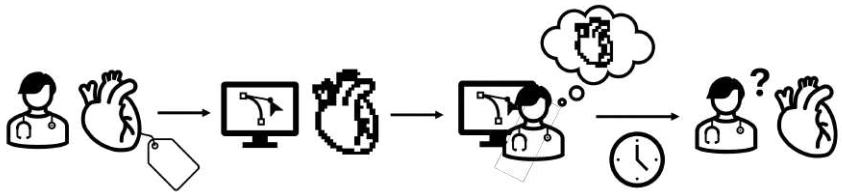


Figure 8.5: Epistemic sclerosis.

The consequences are not merely theoretical. The study reviews evidence in the literature of several downstream effects associated with this process, including deskilling, decision atrophy, and a gradual

detachment from professional responsibility. As AI outputs come to be perceived as epistemically superior or "more objective", clinicians may exert less constructive distrust toward system recommendations, reducing their sensitivity to concept drift and evolving medical realities. In extreme scenarios, AI-derived classifications risk becoming the dominant—if not sole—source of medical knowledge, narrowing the epistemic space in which learning, adaptation, and innovation can occur.

Crucially, the paper argues that epistemic sclerosis is not an inevitable consequence of automation, but the result of specific design and organisational choices—particularly those that privilege efficiency, standardisation, and apparent certainty over deliberation and epistemic plurality. To counter this tendency, the study introduces frictional ground truthing as a design-oriented response. Drawing on perspectives from CSCW, HCI, and classification theory, frictional ground truthing reframes labelling as an iterative, revisable, and explicitly contestable practice.

Within this framework, epistemic sclerosis is understood to take hold at three critical moments: when documentation fails to support organisational learning; when labelling workflows exclude alternative expert perspectives; and when AI outputs present classifications as unappealable conclusions rather than provisional interpretations. Correspondingly, the principles of *Openness*, *Multiplicity*, and *Auxiliarity* are proposed as epistemic countermeasures. *Openness* preserves the traceability and revisability of ground truth decisions (Table 8.16); *Multiplicity* sustains disagreement and alternative interpretations rather than erasing them (Table 8.18); and *Auxiliarity* ensures that AI systems support reflective judgement instead of replacing it (Table 8.20).

Table 8.16: Strategies supporting *Openness* in the categorisation phase.

Strategy	Description	Design Implications	References
Reflexive documentation	Record labelling decisions, actor rationales, and epistemic assumptions so work remains inspectable and accountable.	Dataset Nutrition Labels; Datasheets; embedded rationales; confidence capture; versioned docs; model/annotator positionality; metadata capture.	Cabitza et al. 2020; Cambo and Gergle 2022; Gebru et al. 2021; Hollan, Hutchins, and Kirsh 2000; Lebovitz, Levina, and Lifshitz-Assaf 2021; Miceli et al. 2021
Participatory labelling	Co-create and evolve label schemas with expert annotators and developers through structured deliberation.	Deliberation sessions; shared labelling guides; co-authored schema artefacts.	Miceli and Posada 2022.
Adaptive review cycles	Iteratively refine labels as knowledge and practice evolve; reopen cases on signal.	Feedback loops between schema and annotation; user-flagged cases for re-evaluation; structured stress tests; MAMA-style iteration.	Pustejovsky and Moszkowicz 2012; Zajac 2022

Table 8.18: Strategies supporting *Multiplicity* in the annotation phase.

Strategy	Description	Design Implications	References
Multi-labelling & fuzzy categorisation	Allow multiple valid labels, ranked alternatives, spectrum/probabilistic assignments.	Multi-label standards; TWD models; range-based systems; probabilistic/spectrum labelling; disagreement-aware aggregation.	Aroyo and Welty 2014; Barrett, Chen, and Zhang 2023; Campagner et al. 2021; Chen, Weld, and Zhang 2021; Davani, Díaz, and Prabhakaran 2022; Yao 2009
Jury learning & multiverse analysis	Aggregate interpretations via deliberative, perspective-sensitive modelling beyond majority vote.	Jury-based model composition; multiverse pipelines; schema/perspective switching in UIs.	Cambo and Gergle 2022; Gordon et al. 2022; Lebovitz, Levina, and Lifshitz-Assaf 2021; Steegen et al. 2016
Social labelling interfaces	Surface annotator disagreement; support collaborative interpretation.	Disagreement-awareness UIs; collaborative tools; gamified engagement.	Muller et al. 2021
Tailored labelling interfaces	Adapt UI to annotator expertise, task complexity, cognitive state.	Annotator-adaptive UIs; cognitive load–aware features.	Muller et al. 2021
Uncertainty-aware annotation	Capture and preserve uncertainty explicitly.	Confidence-weighted labels; uncertainty flags.	Cabitza et al. 2020

Table 8.20: Strategies supporting *Auxiliarity* in the decision support phase.

Strategy	**Description**	**Design Implications**	**References**
Multi-label, agonistic & evaluative outputs	Preserve multiple expert viewpoints in AI outputs; avoid monolithic classifications.	Multi-perspective output displays; conformal prediction; present alternative plausible decisions.	Davani, Díaz, and Prabhakaran 2022; Shafer and Vovk 2008
Disagreement awareness & abstention	Expose uncertainty and allow abstention on low-confidence or high-disagreement cases.	Disagreement flagging; abstention; visualisation of annotation variance.	Cabitza et al. 2020; Davani, Díaz, and Prabhakaran 2022
Jury & multiverse decision support	Let users toggle among competing ground-truth interpretations during diagnosis.	Schema-switching UIs; perspective-aware diagnostic pathways.	Cambo and Gergle 2022; Dragicevic et al. 2019; Gordon et al. 2022; Steegen et al. 2016
Cognitive forcing	Introduce targeted friction to prevent uncritical acceptance of AI outputs.	Micro-boundaries; time gates; decision-confirmation prompts.	Buçinca, Malaya, and Gajos 2021; Cox et al. 2016
Constructive distrust	Normalise reflective scepticism as an interaction norm; make seams/limits visible.	Seamful UIs; trust-calibration through friction; uncertainty-forward displays.	**nais**; Chalmers 2003; Hildebrandt 2019; Inman and Ribes 2019; Naiseh et al. 2021

9

EVALUATING HUMAN–AI INTERACTION BEYOND ACCURACY

> *The prioritization of performance values is so entrenched in the field [of Machine Learning Research] that generic success terms, such as "success", "progress", or "improvement" are used as synonyms for performance and accuracy. However, one might alternatively invoke generic success to mean increasingly safe, consensual, or participatory ML that reckons with impacted communities and the environment.*
> — Abeba Birhane *et al., The Values Encoded in Machine Learning Research* (Birhane et al., 2022, p. 179)

9.1 INTRODUCTION

This part presents the methodological contributions of the thesis. Starting from the conceptual premise that the effects of AI emerge from interactional configurations rather than from model performance alone, it addresses the limitations of accuracy-centred evaluation paradigms in capturing the real-world impact of AI-supported decision-making.

The first set of publications introduce a beyond-accuracy framework for assessing human–AI interaction (Natali, Campagner, and Cabitza, 2024; Natali et al., 2023). By operationalising constructs such as reliance patterns, automation bias, conservatism bias, and technology impact, this work provides methodological tools to evaluate how AI influences human judgment, rather than merely whether it produces correct outputs. These measures are explicitly designed to capture interaction-level phenomena identified in the conceptual part of the thesis, including early manifestations of over-reliance and competence erosion.

The contribution presented in Section 9.2.1 demonstrates the application of this methodological perspective in a clinical setting, through the assessment of explainable AI in diagnostic decision-making, analysing the impact of eXplainable AI on accuracy, reliance, and decision revision across different levels of expertise and case complexity. This work shows that the impact of XAI is conditional and cannot be inferred from performance metrics alone, as *beyond-accuracy* evaluation reveals effects—both beneficial and detrimental—that would remain invisible under conventional assessment strategies (Natali et al., 2023).

The final contribution in this part (Cabitza et al., 2024a) addresses the limits of explainability known in the literature as *eXplainability para-*

doxes (Bertrand et al., 2022; Morrison et al., 2024). Through empirical investigation of misleading explanations, it shows that explanations can distort decision-making by increasing confidence without improving understanding. Explainability is thus treated not as an intrinsic virtue of AI systems, but as a *meta-output* whose effects must be empirically assessed. Collectively, the methodological contributions presented here establish how human–AI interaction protocols can be evaluated in ways that are commensurate with their emergent and socio-technical nature.

9.2 ASSESSING AI'S IMPACT ON HUMAN DECISION-MAKING: INDIVIDUAL AND EXPERIENCE-GROUP LEVEL

9.2.1 *Assessing the impact of XAI on radiological diagnostic tasks*

Bibliographic reference

Natali, C., Famiglini, L., Campagner, A., La Maida, G. A., Gallazzi, E., & Cabitza, F. (2023). *Color Shadows 2: Assessing the Impact of XAI on Diagnostic Decision-Making*. In: Longo, L. (ed), *Explainable Artificial Intelligence*. xAI 2023. Communications in Computer and Information Science, vol. 1901. Springer, Cham. https://doi.org/10.1007/978-3-031-44064-9_33[1]

Thoraco-lumbar fracture detection from X-ray imaging remains a diagnostically demanding task, with persistently high error rates despite clinical expertise. While explainable AI has been proposed as a means to support human decision-making in such complex settings, its actual impact on diagnostic performance and reliance behaviour remains contested. In particular, explanations may either enhance human judgement or exacerbate dysfunctional effects such as automation bias or algorithmic aversion. This study therefore adopts a socio-technical evaluation of XAI, examining not only accuracy gains but also patterns of reliance within a deliberately structured Human–AI Collaboration Protocol, tested longitudinally in a realistic diagnostic task.

9.2.1.1 *Research questions*

The study addressed the following questions:

RQ1 What is the impact of XAI support on diagnostic accuracy?

- RQ1a Does this impact differ by radiological expertise?
- RQ1b Does this impact differ by case complexity?

RQ2 Does XAI support induce automation bias?

RQ3 Does XAI support induce conservatism bias?

1 Open access available at https://boa.unimib.it/retrieve/99640799-4a64-4331-bb2f-53b77aba112e/Natali-2023-XAI-AAM.pdf

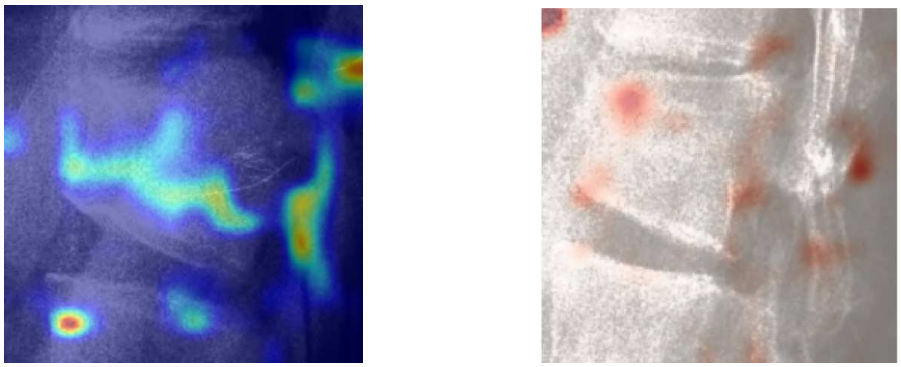


Figure 9.1: Two examples of AM, based on the moderate level of detail (left-hand side: traditional coloring, right-hand side: semantic coloring)

Figure 9.2: Two examples of x-rays and their corresponding activation map. On the left-hand side the low-level AMs is used, while on the right-hand side is shown the high-level AM.

9.2.1.2 *Methods*

PARTICIPANTS Sixteen orthopaedic clinicians participated in the study. Half were residents or specialists with 5 years of professional experience (lower expertise), and half were specialists or sub-specialists with >5 years of experience (higher expertise).

XAI SYSTEM AND STIMULI Participants were supported by a decision support system providing both binary AI advice (fracture vs no fracture) and visual explanations in the form of Activation Maps (AMs) generated via Class Activation Mapping.

AMs highlighted image regions most influential for the model's prediction. To reflect a heterogeneity of XAI solutions, explanations varied along two dimensions: feature detail level (low vs high), and colouring scheme (traditional red–blue saliency vs semantic colouring aligned with the AI's classification outcome).

Participants could interactively compare the original X-ray and the AM using an interactive comparison slider.

Importantly, the study did not aim to compare AM variants against each other, but to assess the overall impact of XAI support.

From the test dataset, two physicians selected 18 X-rays (balanced for fracture presence and case complexity), ensuring interpretability despite reduced resolution (800800 pixels). The sample included both low- and high-complexity cases to enable stratified analyses.

TASK AND PROCEDURE The study followed a longitudinal within-subject design comprising two reading sessions separated by a three-week wash-out period.

During the first session, participants independently reviewed all 18 X-rays in randomised order and provided a diagnosis, confidence rating, and perceived case complexity, without AI support.

Three weeks later, participants reviewed the same cases (again randomised), this time receiving both the AI classification and the corresponding AM. They then provided a final diagnosis, confidence rating, and perceived utility of the XAI support.

Diagnostic performance was compared pre- and post-XAI support. Analyses were stratified by clinician expertise and case complexity. Automation bias and algorithmic aversion were operationalised via reliance patterns derived from changes between initial human judgement (HD1), AI advice, and final human decision (HD2).

9.2.1.3 *Results*

OVERALL DIAGNOSTIC PERFORMANCE Human–AI hybrid performance (Humans+XAI) achieved a mean accuracy of 0.89 (95% CI [0.70, 1.00]), outperforming both unaided humans (0.79 [0.58, 0.97]) and the AI alone (0.77 [0.66, 0.88]). The improvement from pre-XAI to post-XAI was statistically significant (Wilcoxon test p = .007; BH-adjusted p = .035) with a large effect size (d = 1.23), exceeding one standard deviation.

Although two participants showed a reduction in accuracy and two showed no change, the collective effect was robustly positive, as shown by the Benefit Diagram (Figure 9.3).

EXPERTISE-STRATIFIED EFFECTS For lower-expertise clinicians, the XAI effect did not reach statistical significance (p = .115; BH-adjusted p = .144), despite a moderate-to-large effect size (d = 0.79). In contrast, higher-expertise clinicians exhibited a significant improvement (p = .017; BH-adjusted p = .042) with a very large effect size (d = 1.76). This asymmetry suggests that effective use of AM-based explanations presupposes substantial domain-specific visual expertise.

CASE COMPLEXITY EFFECTS XAI support significantly improved diagnostic accuracy in high-complexity cases (p = .027; BH-adjusted p = .045; d = 1.07), whereas no meaningful effect was observed for low-complexity cases (p = .689; d = 0.14). This indicates that XAI

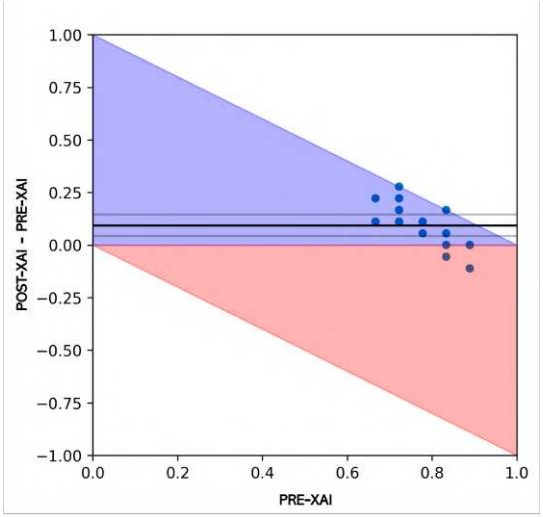


Figure 9.3: Benefit diagram comparing the accuracy of the unaided human (pre-XAI) with the effect of the XAI intervention, showing a clear benefit.

yields benefits primarily where unaided human performance is most challenged.

AUTOMATION BIAS AND CONSERVATISM BIAS Analysis of reliance patterns revealed no evidence of automation bias (Figure 9.4: incorrect AI advice was not systematically over-adopted.

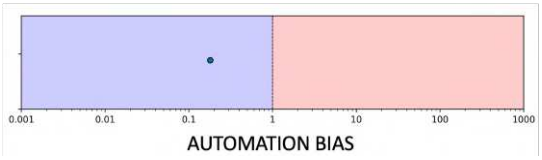


Figure 9.4: Benefit diagram related to the phenomenon of automation bias (AB). Red indicates a presence of AB effect, while blue indicates its absence.

Likewise, conservatism bias was minimal (Figure 9.5, with clinicians generally incorporating correct AI advice rather than dismissing it.

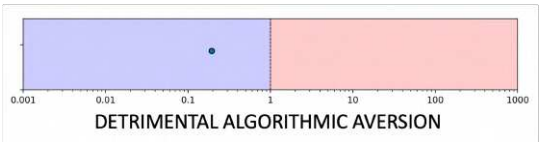


Figure 9.5: Benefit diagram related to the phenomenon of conservatism bias, or *detrimental algorithmic aversion*. Red indicates a presence of AB effect, while blue indicates its absence.

9.2.1.4 *Limitations*

The study is constrained by a limited number of cases and participants, which may reduce statistical power in subgroup analyses. Additionally, the design does not disentangle the effect of explanations from that of AI advice alone, as both were presented simultaneously. Future work should isolate these components and examine more granular interaction protocols, including intermediate decision checkpoints.

9.2.1.5 *Summary*

The study shows that XAI support based on activation maps yields a statistically significant and practically relevant improvement in diagnostic accuracy for thoraco-lumbar fracture detection, with Human+XAI performance (accuracy = 0.89) exceeding both unaided clinicians and the AI alone. This benefit is strongly moderated by expertise and case complexity: accuracy gains are significant and large for more experienced clinicians and for high-complexity cases, while effects for less experienced users and low-complexity cases are non-significant. Crucially, the introduction of XAI does not induce automation bias nor detrimental algorithmic aversion, as evidenced by stable and appropriate reliance patterns.

9.2.2 *Further work: The* Human-AI Interaction assessment tool

Bibliographic reference
Natali, C., Campagner, A., & Cabitza, F. (2024). Answering the Call to Go Beyond Accuracy: An Online Tool for the Multidimensional Assessment of Decision Support Systems. In *BIOSTEC* (2) (pp. 219-229).[2]

The starting point of the contribution is the recognition that accuracy, while indispensable, is methodologically insufficient as a standalone criterion for DSS evaluation. Empirical evidence from the machine learning literature shows a persistent over-reliance on performance metrics, with limited attention paid to negative downstream effects or interactional consequences. The methodological problem, therefore, is not simply that "accuracy is overvalued", but that existing evaluation practices lack the tools needed to capture how AI systems behave once embedded in human decision-making processes. The assessment tool presented in this study is explicitly designed to address this gap.

At the time the paper was written (October 2023), a cursory Scopus search for the term "Beyond Accuracy" indicated an emerging body of work in Computer Science and Engineering, with 42 conference papers and journal articles featuring the expression in their title. By January 2026, this figure had increased to 88 publications.[3].

Methodologically, the tool operationalises a multi-dimensional framework for DSS quality assessment, in which accuracy is retained but embedded within a broader set of complementary dimensions. These dimensions—robustness, data similarity, calibration, utility, data reliability, and human interaction—are treated as independent yet composable modules, allowing evaluators to adapt the assessment to available data and contextual constraints (Figure 9.6). This modularity is a deliberate design choice, acknowledging that comprehensive evaluation is often infeasible in practice, while still promoting methodological rigour where possible.

Crucially, the contribution moves beyond model-centric validation by providing formal methods to assess human–AI interaction effects. The human interaction module of the tool is explicitly methodological in nature: it defines experimental protocols, metrics, and visualisations that enable evaluators to measure how AI advice influences human judgement. Central to this module is the adoption of a human-first decision-making protocol, in which an unaided human judgement is

2 Open access available at https://boa.unimib.it/handle/10281/572326

3 `query: TITLE ( "beyond accuracy" ) AND ( LIMIT-TO ( SUBJAREA , "COMP" ) OR LIMIT-TO ( SUBJAREA , "ENGI" ) ) AND ( LIMIT-TO ( DOCTYPE , "cp" ) OR LIMIT-TO ( DOCTYPE , "ar" ) )`

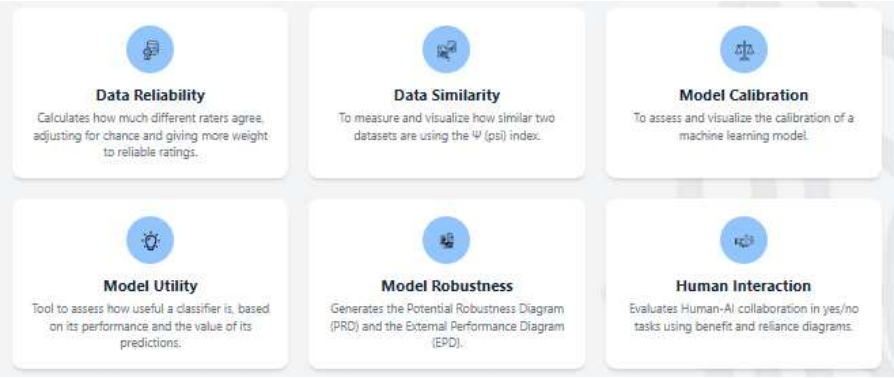


Figure 9.6: The dimensions of DSS quality assessment available as of January 2026: robustness, data similarity, calibration, utility, data reliability, and human interaction. Screenshot of the webpage https://mudilab.github.io/metimeter-frame-2025/tools.html.

elicited before exposure to AI output, followed by a final, post-AI decision. This protocol is not proposed as a normative interaction design, but as a measurement strategy that makes AI influence empirically observable.

The framework of reliance patterns, described in Section 3.4, is leveraged as the methodological device for classifying decision shifts. By encoding all possible combinations of correct and incorrect human decisions, AI recommendations, and final outcomes, the framework enables a systematic mapping between observable behaviour and well-established cognitive biases (Figure 9.7). Importantly, this mapping is not left at a descriptive level: the tool defines computable metrics—Automation Bias, Conservatism Bias (Figure 9.8), and Technology Impact that quantify these effects using odds ratios. In doing so, the methodology translates abstract concerns about over-reliance, under-reliance, and technological dominance into reproducible, comparative measures.

The notion of decision benefit further reinforces the methodological orientation of the contribution. Rather than treating benefit as an implicit consequence of improved accuracy, the tool defines it explicitly as a difference measure between human performance with and without AI support, under controlled conditions. The associated benefit diagram (Figure 9.9) is introduced as a visual analytic method, designed to reveal not only average effects but also heterogeneity across users and baseline skill levels, by allowing stratified analyses according to user type, AI support, and study. Methodologically, this allows evaluators to detect cases in which AI support produces uneven, negligible, or even negative effects—outcomes that would remain invisible under aggregate accuracy reporting.

Reliance Patterns

Study A

HD1	AI	FHD	Count (AI)	Count (XAI)
0	0	0	79 (18.8%)	89 (21.2%)
0	0	1	0 (0.00%)	1 (0.24%)
0	1	0	57 (13.6%)	45 (10.7%)
0	1	1	59 (14.0%)	12 (2.86%)
1	0	0	11 (2.62%)	1 (0.24%)
1	0	1	36 (8.57%)	35 (8.33%)
1	1	0	0 (0.00%)	0 (0.00%)
1	1	1	178 (42.4%)	237 (56.4%)
			tot: 420	tot: 420
53.57%	70.00%	65.00%	AI	
65.00%	70.00%	67.86%		XAI

Figure 9.7: The reliance pattern table defines all potential decision-making and AI reliance patterns between human decision-makers and their AI-based Decision Support systems. In the first three columns, '0' indicates an incorrect decision and '1' stands for a correct decision. For each decision pattern, we characterize the decision-maker's kind of reliance toward the AI system, according to whether they accept or discard the AI's advice and whether this is right or wrong. Additionally, we identify the main cognitive biases associated with each pattern. Generated with the tool available at https://mudilab.github.io/metimeter-frame-2025/tools.html.

The methodological contribution culminates in an explicit alignment with the concept of technovigilance. Here, technovigilance is not framed as a general ethical stance, but as a methodological commitment: the continuous, structured re-assessment of DSS performance and effects after deployment. The tool is positioned as enabling this form of ongoing oversight by supporting repeated measurements, comparative analyses over time, and the detection of emergent biases or degradations in human–AI collaboration. Evaluation, in this view, becomes an iterative practice rather than a one-off validation step.

The tool is under continuous refinement. In this chapter, I refer to the functionalities available at the time of publication of this contribution

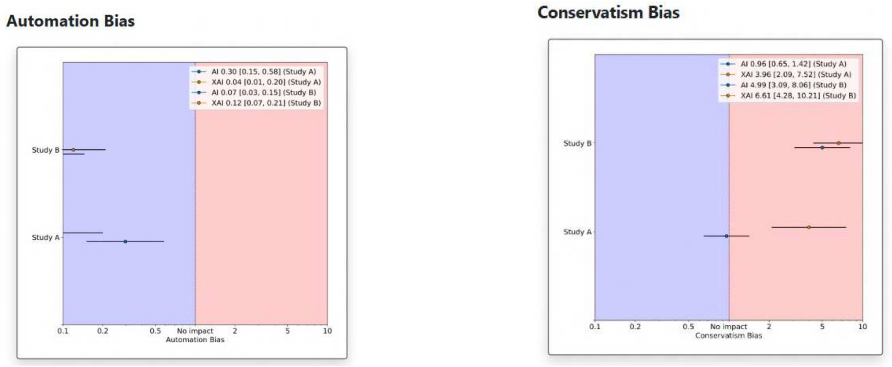


Figure 9.8: Visualisations of Automation Bias (left) and Conservatism Bias (right)

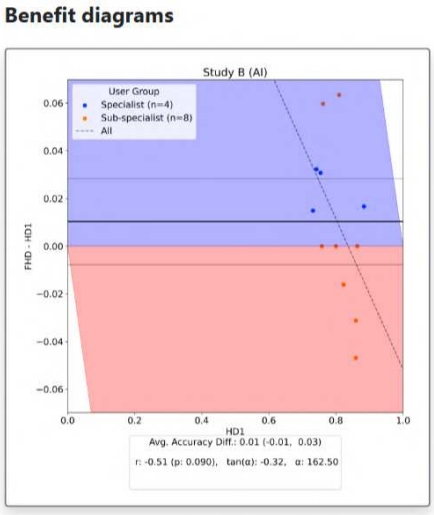


Figure 9.9: The benefit diagrams: the dots represent the accuracies of the humans, and the black lines the average difference in accuracy between the pre-AI and the post-AI decisions, along with the corresponding 95% confidence interval. The blue region denotes an improvement in error rates, while the red region denotes a worsening.

(Natali, Campagner, and Cabitza, 2024), which were employed in the user studies reported in this thesis (Cabitza et al., 2024a,b; Natali et al., 2023, 2024). More recent additions include paired plots enabling within-subject comparisons of mean accuracy before and after AI support (i.e., first versus final decision), thereby allowing stratified analyses across higher- and lower-performing participants. The tool also incorporates protocol diagrams, as investigated in (Cabitza et al., 2025a), which identify the decision-making protocol that maximises final accuracy given observed patterns of appropriate reliance. Five alternative human–AI decision protocols are considered, differing in the timing, role, and degree of human and AI involvement in the decision process.

In line with a technovigilance perspective, the tool continues to evolve as new evaluation needs and methodological opportunities emerge.

9.3 STUDYING THE EXPLAINABILITY PARADOXES

A recurring assumption in explainable AI (XAI) is that augmenting advice with explanations will improve decision quality and reliance calibration. Yet, explanations may be *misleading* even when they are linguistically coherent: a "correct" explanation can align with an AI's output while remaining irrelevant to the task's true rationale (Morrison et al., 2024). This decoupling motivates the broader *explainability paradox*: not only persuasive explanations can make users accept wrong advice (the "white-box paradox"), but, symmetrically, a poor explanation can have users casting doubt on the validity of an otherwise correct advice.

We studied the latter, neglected case: *coherent–irrelevant* explanations that "look right" because they follow the answer, yet are not pertinent to solving the task. Through what we termed the *XAI halo effect* (Cabitza et al., 2024a), we show how perceived explanation quality (or lack thereof) can "bleed" onto the perceived quality of advice, harming appropriate reliance even when advice is correct. We also introduce a lightweight taxonomy of misleading explanations, classified according to pertinence and relevance (Figure 9.10).

9.3.1 *The impact of misleading explanations on Human–AI decision making*

Bibliographic reference

Cabitza, F., Fregosi, C., Campagner, A., & Natali, C. (2024). *Explanations Considered Harmful: The Impact of Misleading Explanations on Accuracy in Hybrid Human–AI Decision Making*. In: Longo et al. (eds) *xAI 2024*, CCIS 2156, Springer.[4]

Explainable AI (XAI) is widely promoted as a mechanism to support appropriate reliance in human–AI decision-making. However, a growing body of evidence has shown that explanations may paradoxically increase inappropriate reliance, particularly when persuasive explanations accompany incorrect advice—a phenomenon commonly referred to as the white-box paradox. This study extends the scope of this paradox by shifting analytical attention to a less explored but equally consequential configuration: accurate AI advice paired with misleading explanations.

Building on a minimal but operational framework that distinguishes explanation quality along two dimensions—coherence with the advice and relevance to the task—this work investigates explanations that are coherent yet irrelevant (Figure 9.10). Such explanations do not contradict the AI's output, but fail to provide information that is genuinely

4 Open access available at: https://boa.unimib.it/handle/10281/499419

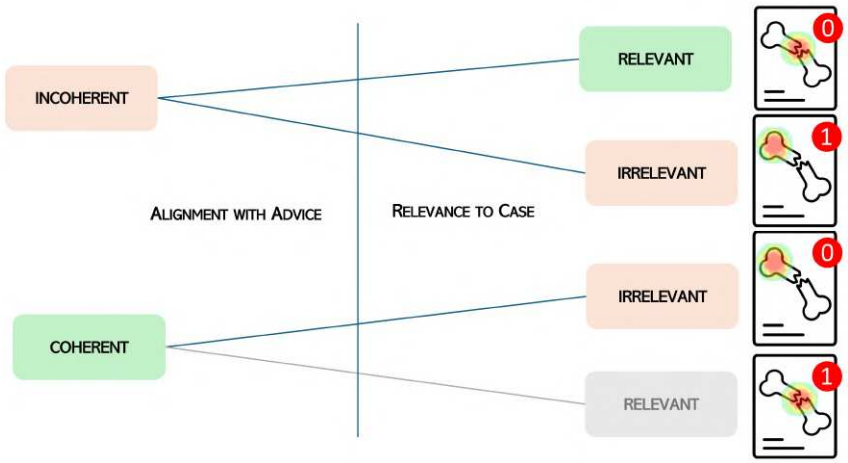


Figure 9.10: The lightweight taxonomy of *misleading explanations* according to *coherence* to the AI advice and *relevance* to the case.

useful for validating it. We hypothesise that these explanations can exert a halo effect on users' judgement, biasing their assessment of the advice itself. This phenomenon, termed the *XAI halo effect*, complements the white-box paradox and together they constitute what we describe as the broader explainability paradox.

AI (advice)	XAI (explanation)	FHD (final decision)
0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1

WHITE BOX PARADOX

XAI HALO EFFECT

Figure 9.11: Reliance pattern-based description of the White Box Paradox and Halo Effect. The WBP corresponds to a incorrect AI advice (AI=0), persuasive explanation (XAI=0) resulting in a wrong final decision due to over-reliance (FH=0). The HE corresponds to a correct AI advice (AI=1) accompanied by an incoherent or irrelevant explanation (XAI=0), causing under-reliance on the correct AI output (FH=0).

9.3.1.1 *Research questions*

RQ1 Does correct AI advice coupled with a misleading explanation negatively affect user accuracy? If so, are some user strata more susceptible to this effect?

RQ2 Does correct AI advice coupled with a misleading explanation affect users' confidence in their decisions?

9.3.1.2 *Methods*

PARTICIPANTS AND TASK The study involved 22 Master's students enrolled in an Artificial Intelligence programme, who individually completed 19 moderate-to-hard logic puzzles akin to those used in psychometric assessments of general intelligence. Tasks covered numerical and alphanumeric reasoning, deductive logic, graphical interpretation, and anagrams.

Participants interacted with a simulated ChatGPT-like interface, implemented via a Wizard-of-Oz design, which returned both an answer and an explanation for each puzzle. Of the 19 AI responses, 13 were correct and 6 incorrect. Among the correct responses, 6 were paired with misleading explanations (coherent with the advice but irrelevant to the task), while 7 were paired with correct explanations. All incorrect AI responses were accompanied by misleading explanations. To foster initial trust calibration and avoid early algorithmic aversion, 9 correct AI responses were presented consecutively at the beginning of the task sequence. Figure 9.12 reports the pattern of correctness/incorrectness for each question and their related explanations.

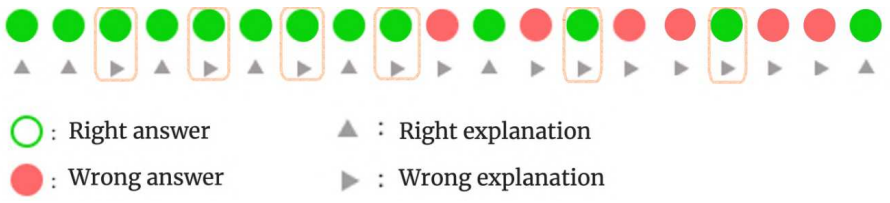


Figure 9.12: The distribution of correct and incorrect AI responses within the experimental question set. 13 AI advices were correct, and 6 incorrect. Among the correct cases, 6 displayed *misleading* XAI explanations of the *coherent/irrelevant* type.

After each AI-supported interaction, participants selected a final answer from four options and reported their confidence on a 4-point ordinal scale (from "not sure at all" to "almost certain"), designed to mitigate central tendency bias. The study was conducted in person, and responses were collected via LimeSurvey.

MEASURES Given the non-normality of accuracy scores and the ordinal nature of confidence ratings, Mann–Whitney U tests (normal approximation, = .05) were used to compare participant-level aggregated outcomes across the two focal configurations: *Correct AI + Correct XAI* vs. *Correct AI + Misleading XAI*. Effect sizes were reported using both standardized measures and Common Language Effect Size (CLES), in line with the exploratory nature of the study.

To examine differential susceptibility, participants were stratified by performance quartiles (Q1 vs Q4) and contrasted via Mann–Whitney U. Interaction-centred analyses further mapped item-level outcomes to Technology Impact and Conservatism Bias diagrams (Figures 9.15 and 9.14).

9.3.1.3 *Results*

IMPACT OF MISLEADING EXPLANATIONS ON ACCURACY AND RELIANCE As shown in Figure 9.13, accuracy was significantly lower in *Correct AI + Misleading XAI* than in *Correct AI + Correct XAI* ($p < .001$; standardised effect $= 0.51$; CLES $= 0.79$). Thus, in $\sim$4 out of 5 pairwise comparisons, participants performed better with correct than misleading explanations despite identical (correct) advice.

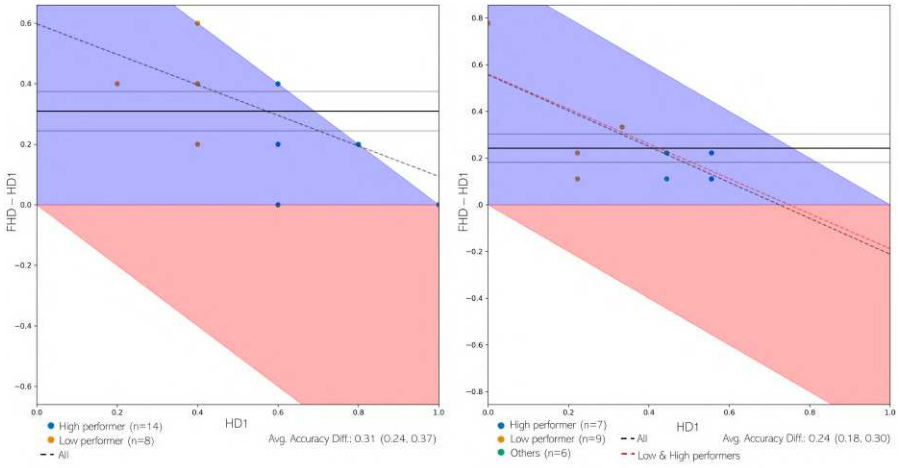


Figure 9.13: Benefit diagrams for the "correct AI and correct XAI" (left) and "correct AI and misleading XAI" (right) cases. Generated with the tool available at https://mudilab.github.io/dss-quality-assessment/.

An analysis of reliance patterns, using Technology Impact (Figure 9.15) and Conservatism Bias (Figure 9.14) metrics, further contextualised these findings. While AI assistance had an overall positive effect on decision accuracy irrespective of explanation quality, misleading explanations were associated with significantly higher conservatism bias.

Participants exposed to misleading explanations were more likely to remain anchored to their initial judgements, even in the presence

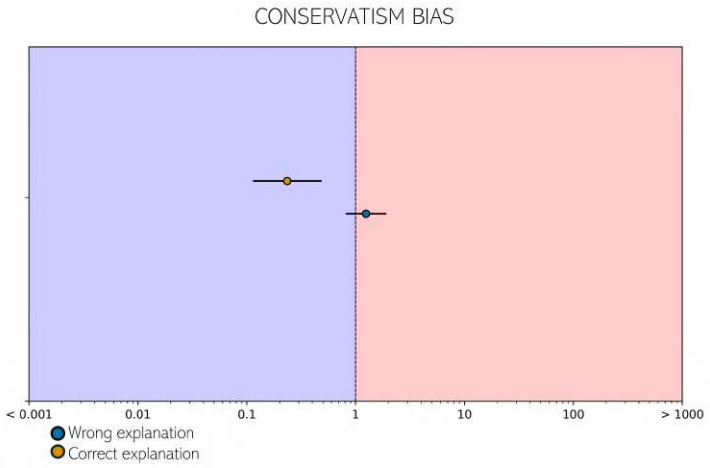


Figure 9.14: Conservatism Bias diagram, the red region denotes an overall negative effect of the AI intervention, while the blue region denotes an overall positive ef- fect. Generated with the tool available at https://mudilab.github.io/dss-quality-assessment/.

of correct AI advice, signalling a subtle but systematic distortion of reliance dynamics.

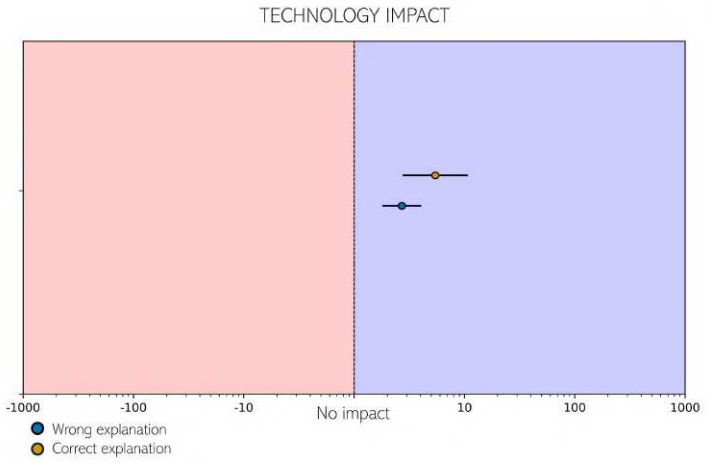


Figure 9.15: Technology Impact diagram, the red region denotes an overall negative effect of the AI intervention, while the blue region denotes an overall positive effect. Generated with the tool available at https://mudilab.github.io/dss-quality-assessment/.

To examine heterogeneity effects, participants were stratified into low- and high-performers based on the first and fourth quartiles of overall accuracy. The detrimental impact of misleading explanations was significantly stronger for low-performers ($p = .04$, medium effect size $= 0.4$). This suggests that misleading explanations disproportionately harm those users who would benefit most from AI support, amplifying rather than mitigating performance disparities.

IMPACT OF MISLEADING EXPLANATIONS ON CONFIDENCE In contrast to accuracy, confidence ratings did not differ significantly between conditions ($p = .086$), and the observed effect size was small (0.26). Participants expressed comparable confidence whether the explanation was correct or misleading, despite the substantial performance gap. This dissociation between confidence and accuracy indicates that users did not recognise the inadequacy of the misleading explanations, lending further support to the notion of an XAI halo effect: explanations can silently impair judgement without triggering epistemic doubt. .

LIMITATIONS AND VALIDITY This study is exploratory and subject to several limitations. The sample size ($N = 22$) and the limited number of tasks (19) constrain statistical power and generalisability. Accordingly, greater interpretive weight is placed on effect sizes rather than significance testing.

Moreover, the interpretation rests on two simplifying assumptions. First, that participants perceived the tasks as sufficiently difficult to warrant engagement with AI advice; Second, that exposure to explanations did not allow participants to form overly stable independent judgements. Both assumptions are supported by concurrent baseline data showing that unaided human performance was substantially lower than AI accuracy. If anything, these conditions render the reported effects conservative estimates.

9.3.1.4 *Findings*

The findings presented in this study point towards an important — and often underappreciated — limitation of explainable AI: explanations actively shape how users interpret and respond to those outputs, and not always for the better. In particular, the study demonstrates the existence of what we term the XAI halo effect: when an explanation is coherent with the advice it accompanies but irrelevant to the case at hand, users tend to discard the advice uncritically, even when the correct answer would have been within reach - a form of negativity bias (Bertrand et al., 2022).

Explanations can mislead, and critically, they can do so when the underlying advice is accurate. Our previous account of the white-box paradox (Cabitza et al., 2023d) concerns persuasive explanations propping up incorrect decisions; these results show the mirror image, as misleading explanations destabilise trust in correct decisions. The net effect is not simply "unreliable explanation", but a distortion of the decision-making process itself.

Moreover, the harm is not evenly distributed across users. Lower-performing participants were substantially more vulnerable to misleading explanations than higher performers, widening the performance gap. This raises ethical questions about the distributive effects of XAI:

systems intended to "support" decision-makers may in fact disproportionately impair those who are most in need of reliable support. In practical terms, this calls for further study over whether explainability can unintentionally act as a skill amplifier for those who are already strong, and a skill suppressant for those who are not.

10

FINDINGS ON FRICTIONAL DESIGN IN AI-ASSISTED RADIOLOGICAL TASKS

> *Human interaction is often perceived, from an engineer's mindset, as complicated, inefficient, noisy, and slow. Part of making something "frictionless" is getting the human part out of the way.*
> — David Byrne, *Eliminating the Human* (Byrne, 2017)

INTRODUCTION

This part presents the empirical and design-oriented contribution of the thesis: Frictional AI.

The first set of studies investigates pro-hoc (Cabitza et al., 2024b) and reflective forms of explanation (Cabitza et al., 2023b) as design strategies that interrupt automatic reliance and promote deliberation. This work responds directly to the explainability paradoxes identified in the methodological part of the thesis, presenting instead explanations that are framed as moments of reflection rather than as authoritative justifications.

The second contribution extends this design perspective to the question how decision-support systems can be designed to align with clinical workflows and professional practices. Through qualitative and mixed-method studies in radiological contexts–a proof-of-concept user study on hypothesis-driven support (Rubegni et al., 2025) and an ethnomethodological account on the adoption of AI systems by radiologists (Anichini, Natali, and Cabitza, 2024)–this work evaluates how frictional design choices are perceived, negotiated, and appropriated by clinicians in practice.

The studies presented in this part show that friction is neither a universal remedy nor a usability defect. Instead, it emerges as a situated and bounded interactional strategy, whose effectiveness depends on task characteristics, user expertise, and organisational context. This final part grounds the thesis' conceptual and methodological contributions in empirical practice, demonstrating how emergence-oriented insights can inform the design of human–AI systems that support professional judgment.

10.1 PROMOTING APPROPRIATE RELIANCE THROUGH FRICTIONAL AI IN CLINICAL DECISION-SUPPORT SYSTEMS

10.1.1 *Investigation of* pro-hoc explanations *in radiological decision-making*

Bibliographic reference
Cabitza, F., Natali, C., Famiglini, L., Campagner, A., Caccavella, V., & Gallazzi, E. (2024). Never tell me the odds: Investigating pro-hoc explanations in medical decision making. Artificial intelligence in medicine, 150, 102819.[1]

This study investigates a specific instantiation of frictional AI: pro-hoc explanations, a form of decision support that deliberately withholds categorical AI advice and instead offers evidence aligned with a clinician's own tentative judgement. Unlike post-hoc explanations—which rationalise a machine's output after the fact—pro-hoc explanations replace machine advice altogether. The system first records the clinician's unaided judgement (a human-first, second-opinion protocol) and then presents explanatory evidence corresponding both to that judgement and to its counterfactual alternative.

The underlying motivation is to counteract well-documented cognitive risks of AI-supported decision making—most notably automation bias, complacency, and longer-term deskilling—without abandoning the benefits of decision support. By introducing desirable inefficiency through evidentiary rather than prescriptive support, pro-hoc explanations are hypothesised to promote reflective reasoning while preserving clinician agency.

Within the broad space of explainable AI, this work focuses on example-based explanations, operationalised as the presentation of similar past cases with known ground truth. While similar-case retrieval has been explored extensively in medical AI, prior work has typically framed it as a complement to, or justification for, algorithmic predictions. The contribution here is conceptually distinct: similar cases are provided as the sole form of AI support in a binary diagnostic task, explicitly abstaining from prediction, confidence scores, or probabilistic outputs.

10.1.1.1 *Research questions*

The study addresses five interrelated research questions:

RQ1 What is the impact of replacing AI classifications with example-based pro-hoc explanations on diagnostic performance?

1 Open access available at https://www.sciencedirect.com/science/article/pii/S0933365724000617 or https://www.boa.unimib.it/handle/10281/466918

RQ2 Do the effects of pro-hoc explanations differ between clinicians with different levels of expertise (residents vs specialists)?

RQ3 How do clinicians perceive the usefulness of pro-hoc explanations?

RQ4 How do pro-hoc explanations affect clinicians' confidence in their final decisions?

RQ5 How do perceived task complexity and confidence relate to actual diagnostic accuracy, and how do they interact with pro-hoc support?

10.1.1.2 *Methods*

An empirical user study was conducted in the context of radiological detection of vertebral fractures in spine X-rays. Sixteen orthopaedic physicians participated: ten board-certified spine specialists and six residents. Each clinician evaluated 18 cases (total decisions: $N = 288$), selected for representativeness and varying diagnostic difficulty. All images were presented at 800 × 800 px resolution within a web questionnaire.

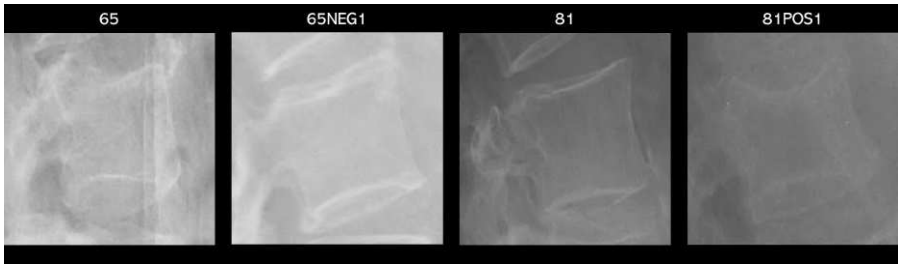


Figure 10.1: Spine x-ray cases presented to the participants.

For each case, clinicians first provided an unaided diagnosis (HD1), alongside self-reported confidence and perceived case complexity (6-point ordinal scales).

The system then retrieved three similar cases using cosine similarity: two supporting the clinician's initial judgement and one counterfactual case from the opposite class. After reviewing these pro-hoc explanations, clinicians provided a final diagnosis (FHD) and updated confidence rating.

No categorical AI advice or probability estimates were provided at any stage. Analyses were conducted using non-parametric tests (= .05). In addition to conventional performance metrics, the study examined reliance on AI through reliance-pattern analysis (stable vs. changed diagnoses); shifts in reported confidence; sensitivity to case complexity, and the Technology Impact (TI) measure and Number of Decisions Needed (NDN) indices. NDN is an adaptation of the

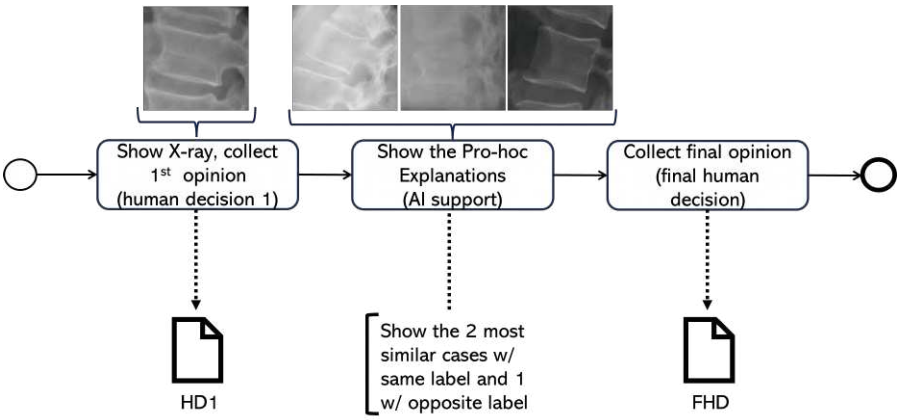


Figure 10.2: The human-first interaction protcol examined.

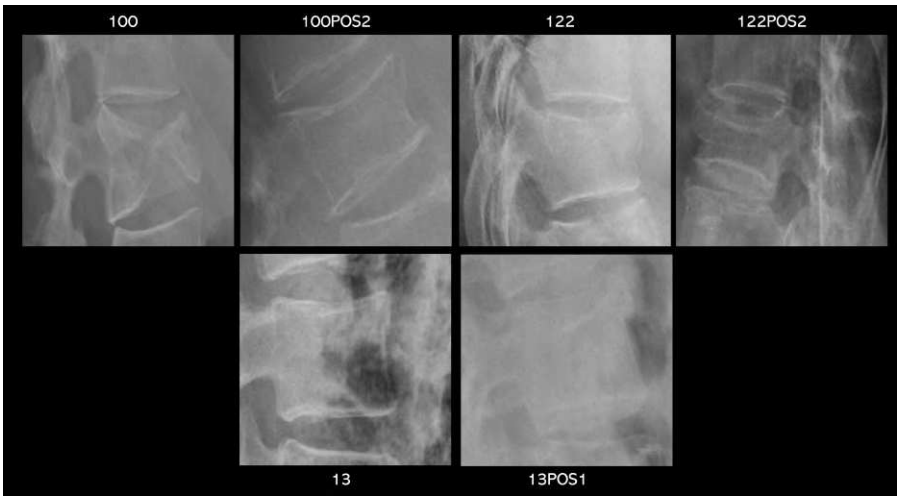


Figure 10.3: Spine x-ray cases presented to the participants.

epidemiological Number Needed to Treat, and quantifies how many supported decisions are required to prevent one error relative to the unsupported condition.

10.1.1.3 *Results*

DECISION PERFORMANCE Overall diagnostic accuracy increased marginally from 78.8% (pre-support) to 80.9% (post-support). This improvement was not statistically significant (p = .53; effect size = .05), yet its practical relevance becomes clearer when examined through decision changes and NDN.

Out of 288 decisions, only 16 changed after exposure to pro-hoc explanations. Crucially, 11 of these changes corrected an initially wrong diagnosis, while only 5 introduced new errors. This corresponds to a 10% reduction in total diagnostic errors (from 61 to 55), yielding an NDN of 50—that is, one error prevented every 50 supported decisions. Small effect sizes of this magnitude are consistent with prior XAI

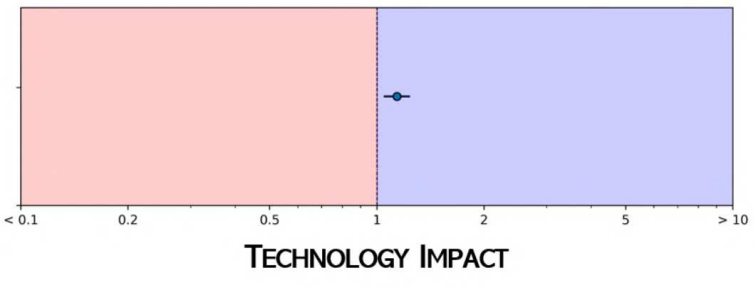


Figure 10.4: Technology impact diagram generated via *Metimeter*.

studies where explanatory support is decoupled from algorithmic prediction.

EXPERTISE-RELATED DIFFERENCES Unaided residents outperformed specialists at baseline (mean accuracy .83 vs .76), albeit non-significantly, plausibly due to greater task engagement. However, this gap disappeared with pro-hoc support.

Specialists benefitted more from the intervention: 60% improved their accuracy post-support, compared to none of the residents. The effect size for specialists was small-to-moderate (0.30), with an NDN of 6, indicating one prevented error every six aided diagnoses. Specialists were also more likely to revise their decisions, particularly on cases perceived as complex, and did so more often in the correct direction. Residents, by contrast, exhibited higher rates of fixation and a greater relative susceptibility to automation bias, despite reporting higher perceived usefulness.

PERCEIVED USEFULNESS Pro-hoc explanations were perceived as highly useful overall (mean $=$ 2.7 on a 4-point scale, 95% CI $[2.62, 2.87]$), with a significant majority of ratings in the upper half of the scale ($p < .001$). Residents rated usefulness significantly higher than specialists ($p = .010$, $Z = 2.57$, standardised es $= .16$, common-language es $= .59$), despite deriving less objective performance benefit—highlighting a dissociation between subjective appreciation and measurable augmentation.

CONFIDENCE Exposure to similar cases led to a modest, non-significant increase in confidence overall ($p = .176, Z = 1.353, es = 0.056, NDN = 32$), with stronger effects among specialists ($p = .07724$, es $= .093$, $NDN = 19$). Importantly, when confidence did change, it increased in two-thirds of cases, a pattern that was statistically significant ($p = .0014$; $NDN = 5$). Thus, even when decisions were unchanged, pro-hoc explanations often reinforced clinicians' confidence in their judgement.

RELATION BETWEEN COMPLEXITY, CONFIDENCE, AND ACCURACY Self-reported confidence and perceived case complexity proved to be strong and reliable proxies for actual diagnostic accuracy. Confidence correlated positively with accuracy ($r = +.48$), while perceived complexity correlated negatively ($r = \breve{}.39$). Higher perceived complexity was also associated with lower confidence ($r = \breve{}.78$) and higher perceived usefulness of pro-hoc explanations ($r = +.32$), supporting the ecological validity of the study and the relevance of subjective measures in evaluating AI support.

TIMING Residents completed the task 15% more slowly than specialists (+4 minutes on average), consistent with deeper task engagement. In unaided reading, residents showed higher baseline accuracy than specialists (mean .83, SD .06 vs .76, SD .08; effect size 1.03, $T = 2.1$, $p = .50$), a non-significant but notable pattern.

10.1.1.4 *Limitations*

The study is exploratory and limited by a small sample of clinicians and cases, constraining statistical power and generalisability. Nevertheless, the observed effect sizes and NDN values provide actionable estimates for powering future, larger-scale studies and for situating pro-hoc explanations within the broader design space of frictional AI.

10.1.1.5 *Summary and Design Implications*

The results indicate that the benefit of pro-hoc explanations by examples manifests itself in a nuanced and context-dependent manner. While the overall improvement in diagnostic accuracy was small, the Number of Decisions Needed shows that relatively few supported decisions are required to prevent an error, particularly for expert clinicians, for whom one mistake was avoided every six aided diagnoses. This, combined with the absence of direct AI predictions, suggests that pro-hoc explanations belong to the class of frictional AI solutions associated with a low risk of over-reliance and deskilling, as they preserve clinicians' interpretative responsibility rather than substituting it.

The findings further suggest several concrete design implications: (i) subjective measures such as perceived case complexity and diagnostic confidence should be systematically collected, as they reliably correlate with actual accuracy and may be used to dynamically modulate the level of friction introduced by the system; (ii) presenting both positive and negative similar cases can support diagnostic reasoning, with benefits that are perceived more strongly by less experienced clinicians, even when objective gains are greater for experts; and (iii) pro-hoc explanations tend to increase clinicians' confidence in their final decisions, sometimes independently of decision changes, which

may be interpreted as a marker of positive user experience and usability.

Overall, the study positions pro-hoc explanations as a promising frictional design strategy for clinical decision support—one that prioritises reflective reasoning, cautious engagement, and human agency over marginal efficiency gains, and whose empirical effect sizes provide a concrete basis for powering and designing future large-scale evaluations.

10.1.2 *Further work:* Reflective XAI *as friction in orthopedic radiology*

Bibliographic reference
Cabitza, F., Campagner, A., Famiglini, L., Natali, C., Caccavella, V., & Gallazzi, E. (2023). Let me think! investigating the effect of explanations feeding doubts about the AI advice. In *International cross-domain conference for machine learning and knowledge extraction* (pp. 155-169). Cham: Springer Nature Switzerland.[2]

This study investigated a form of explainable AI deliberately designed not to reinforce trust or optimise immediate performance, but to induce reflection and epistemic caution in AI-assisted decision-making. The proposed support, termed *Reflective XAI*, operationalised explanation as a prompt for deliberation rather than reassurance. Instead of justifying the AI's recommendation, it exposed users to contrasting evidence about the system's own fallibility, shifting attention from what the AI suggests to how dependable that suggestion may be in this specific instance.

Concretely, the reflective support followed an AI-first diagnostic protocol in orthopaedic radiology and took the form of post-hoc explanations-by-example. After receiving the AI's categorical advice and a standard pixel attribution map, clinicians were shown two highly similar cases retrieved from the model's prior predictions: one that the AI had classified correctly and one that it had misclassified, both associated with the same predicted label as the case at hand. Each example was accompanied by its own attribution map. The intended function of this design was evaluative rather than confirmatory: by juxtaposing successful and failed past decisions on similar cases, the system implicitly asked users to assess the local reliability of the AI, effectively posing an *"Are you sure?"* challenge rather than an endorsement.

Sixteen orthopaedists (residents and specialists) participated in the study, each diagnosing 18 spine X-rays under the HAI-CP delineated in Figure 10.6. The evaluation focused on three dimensions: diagnostic performance, decision confidence, and perceived utility of the reflective support.

The results were counter-intuitive in several respects. First, the introduction of reflective XAI did not improve diagnostic accuracy. On the contrary, although the overall effect was small and statistically non-significant, accuracy slightly decreased after exposure to the reflective support. Changes in decisions were rare (around 4% of cases), but when they occurred they were more often detrimental than beneficial. Notably, these negative shifts disproportionately involved more experienced clinicians (Figure 10.7), including cases in which experts who

2 Open access available at https://boa.unimib.it/handle/10281/456600

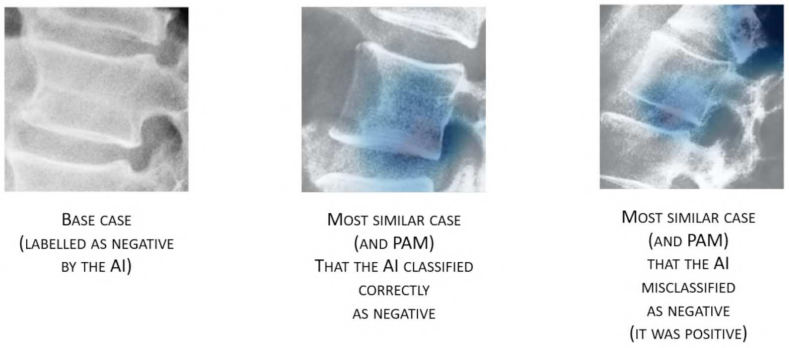


Figure 10.5: Two cases shown to the participants in the user study: on the left, the base cases associated with an advice of no fractures (negative); in the top case the label was right; in the bottom case, the label was wrong; on the right, the two most similar cases with the corresponding pixel attribution maps (PAMs) associated with each base case; they indicate, in the middle, the case correctly identified by the AI and, on the right, the case incorrectly indicated by the AI as positive (to the presence of fractures). This means that in the first base case, the middle XAI case should have reinforced the idea that the AI was right; in the second base case, conversely, the misclassified case on the right should have prompted users to be cautious of the AI's advice.

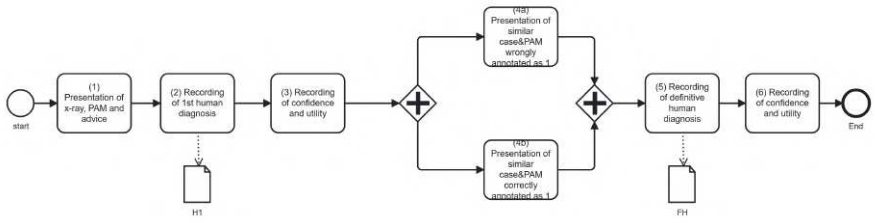


Figure 10.6: Human-AI Interaction Protocol, represented in BPMN notation, adopted in the user study, and the main data collected. Steps 1-3 are performed on the first page of the questionnaire; Steps 4-6 are in the second page of the questionnaire. The protocol is repeated for each base case. The similar cases are displayed in the same page, at the same moment (step 4).

had been correct with AI-only support revised their decisions incorrectly after seeing the reflective examples. From a narrow performance perspective, the reflective XAI thus appeared, at best, inconsequential—and at worst mildly misleading.

Second, the support did not globally increase or decrease decision confidence. However, expertise again played a differentiating role: less experienced clinicians tended to report lower confidence after seeing the reflective examples, whereas more experienced clinicians showed a modest increase in confidence. This divergence suggests that the

same epistemically ambiguous information—two similar cases pulling in opposite directions—was processed very differently depending on prior domain expertise.

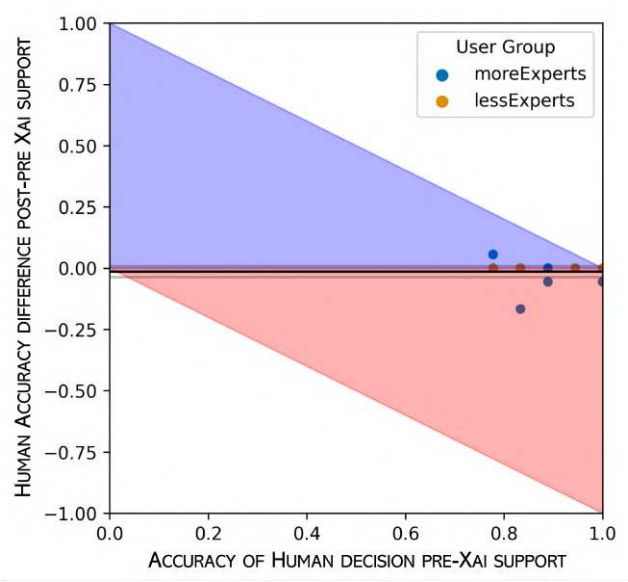


Figure 10.7: Benefit diagram of the introduction of the XAI support (after the AI support). Each point corresponds to a single participant in the study, the solid black line represents the average post-pre XAI support accuracy difference, while the shaded grey lines represent the corresponding 95% confidence interval. The red region of the diagram denotes a worsening, in terms of accuracy, between post- and pre-XAI support; while the blue region denotes an improvement in terms of accuracy. The confidence interval of the aggregate effect contains the zero line, although the average line is slightly below it, so no significant difference in accuracy could be detected in the user study. Generated with `https://haiiassessment.pythonanywhere.com/`

Third, and most strikingly, the reflective XAI was valued by users despite (or perhaps because of) its inconclusiveness. While perceived usefulness was moderate overall, it was significantly higher among more experienced clinicians, who judged the support helpful even though it did not improve—and sometimes worsened—their objective performance.

These findings challenge instrumental views of explanation as a means to optimise here-and-now decisions. Reflective XAI did not function as a performance enhancer; rather, it acted as a deliberation-inducing mechanism that made the AI's competence appear contingent, situated, and open to scrutiny. For experts in particular, the value of the system seemed to lie not in being guided towards the correct answer, but in being prompted to reason about the AI as a fallible

collaborator, drawing on analogical judgement, expectations about case difficulty, and professional experience.

From this perspective, the apparent paradox—lower or unchanged accuracy coupled with higher perceived utility—becomes intelligible. Reflective XAI is not aimed at maximising short-term correctness, but at countering automation bias, preserving epistemic agency, and sustaining clinicians' sense of responsibility over time. Its contribution should therefore be assessed less in terms of immediate decision outcomes, and more in terms of its potential long-term effects on reliance patterns, skill preservation, and the maintenance of reflective judgement in AI-mediated practice.

10.2 FRICTIONAL AI, RADIOLOGISTS' PREFERENCES AND WORKFLOW LOGIC

In this section, I examine whether frictional interaction protocols align with radiologists' established diagnostic practices and workflow constraints.

10.2.1 *Preliminary findings on radiologists' perspectives on frictional AI decision support*

Bibliographic reference
Rubegni, E., Natali, C., Ayoub, O., Rizzo, S. M. R., Valsecchi, C. & Facchini, A. (2025). "Oracles slip on frictionless marble: The case for productive friction in AI-Supported Radiological Work." *Preliminary findings.*

Early clinical AI systems in radiology have largely adopted an *oracular* interaction paradigm (Miller and Masarie Jr, 1990), in which the system outputs a single predicted finding or diagnosis for the clinician to accept or reject. While such systems can reduce omission errors by flagging subtle abnormalities, they risk misaligning with radiologists' actual diagnostic practice. Radiological interpretation is not a passive classification task, but a hypothesis-driven sensemaking process: clinicians typically begin with a clinical question or provisional hypothesis, iteratively inspect images for confirmatory or disconfirmatory evidence, generate differential diagnoses under uncertainty, and, where needed, consult reference cases or external resources.

By contrast, a naïve "AI second reader" that produces a single top-ranked answer may either undermine clinician confidence when it conflicts with an existing hypothesis or induce premature anchoring when it aligns with it. In both cases, control over the diagnostic process risks shifting from the clinician to the system. This tension motivates the exploration of alternative AI interaction strategies that preserve clinician agency while still providing computational support.

Th recent conceptualisation of *Evaluative* or *Hypothesis-driven* AI (Miller, 2023) proposes a shift away from recommendation-centric designs toward interfaces that support human reasoning by surfacing evidence for and against multiple decision options. Rather than persuading users toward a particular conclusion, evaluative systems aim to scaffold abductive reasoning by enabling users to actively generate, inspect, and test hypotheses. In the education setting, López-Pernas et al., 2025, p. 34 noted that "Evaluative AI can not only enhance transparency but also support the critical engagement required *[...]* to maintain autonomy over AI-augmented decisions". However, prior literature also cautions that simply presenting counter-arguments may

backfire, reinforcing fixation or defensive reasoning. Compared to typical XAI outputs accompanied by recommendations, hypothesis-driven explanation pose an increased cognitive load on decision-makers (Miller, 2023). This suggests that not only what information is presented, but how users engage with it, is critical. Against this background, the study presented in this Section empirically contrasts this study contrasts two AI interaction paradigms: an oracular interface that delivers a single suggested diagnosis, and an hypothesis-driven interface that surfaces multiple diagnostic hypotheses with supporting evidence, allowing the clinician to remain the principal agent in sensemaking.

10.2.1.1 *Research questions*

RQ1 How willing are radiologists to adopt oracular versus evaluative AI decision-support interfaces?

RQ2 How willing are radiologists to adopt oracular versus evaluative AI decision-support interfaces?

RQ3 Which interaction paradigm better supports radiologists' diagnostic reasoning process?

METHODS The study followed a mixed-method, within-participants experimental design, in which each radiologist interacted with both prototypes in counterbalanced order. This structure enabled direct comparison of interaction effects while controlling for individual diagnostic style.

PARTICIPANTS Eleven fully qualified radiologists participated in the study. All were practising specialists working in regional hospitals in the canton of Ticino, Switzerland. Participants were recruited via university networks using snowball sampling.

THE TWO PROTOTYPES Both prototypes were web-based and designed to isolate interaction style rather than algorithmic capability. Visual layout, patient metadata, and MRI imagery were held constant.

A curated dataset of 23 MRI cases spanning varying diagnostic complexity was used. Each prototype presented 10 cases per participant, drawn from this pool.

The *Oracular* interface (Figures 10.8, 10.9) presented a single top-ranked diagnostic recommendation. Participants could accept or reject the suggestion; disagreement required them to articulate an alternative hypothesis.

The *Hypothesis-driven* interface (Figures 10.10, 10.11) presented three to four plausible diagnostic hypotheses per case. When selected, each

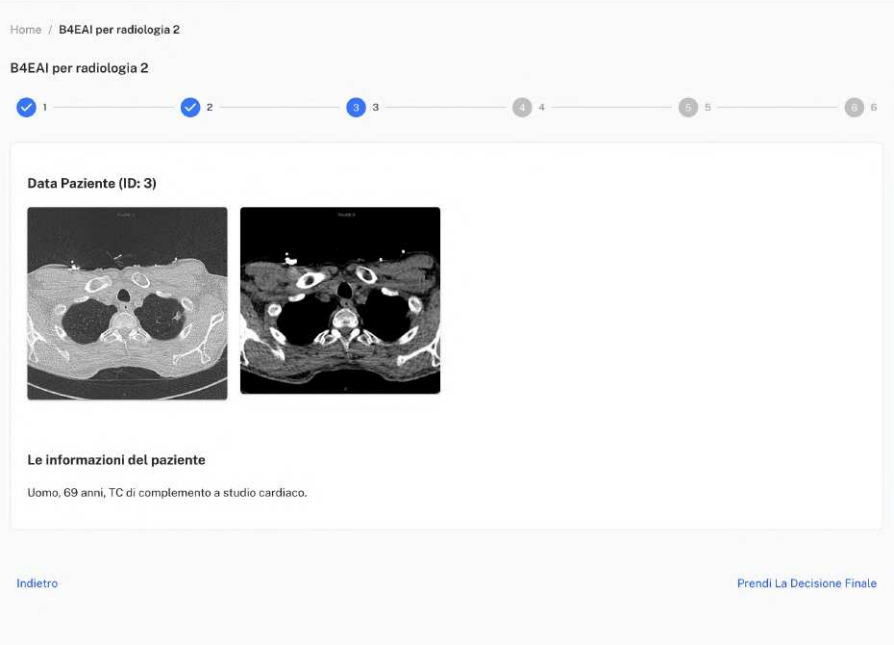


Figure 10.8: A screenshot from the Oracular prototype

hypothesis was accompanied by concise textual evidence and exemplar images. Participants could freely inspect and compare options before committing to a diagnosis.

TASK AND PROCEDURE The study was conducted remotely and proceeded in four phases:

- Baseline Case Review (Phase 0): Participants independently completed an online baseline assessment. classified 20 MRI studies without AI support at least 2 weeks prior to the prototype interaction session. This included demographic and professional background questions, prior experience with clinical AI, and the Technology Anxiety Scale. Participants then classified 20 MRI cases without AI assistance, providing diagnoses, confidence ratings, and perceived case complexity.
- Prototype Interaction I (Phase 1): During a video call with the authors (E. R. & C. N.), participants completed 10 diagnostic tasks using either the oracular or evaluative interface under a think-aloud protocol. After interacting with the prototype, they completed an adapted UTAUT/TAM questionnaire measuring performance expectancy, effort expectancy, attitude toward use, facilitating conditions, job relevance, result demonstrability, anxiety, and behavioural intention to use.
- Prototype Interaction II (Phase 2): Participants repeated the tasks with the other interface.

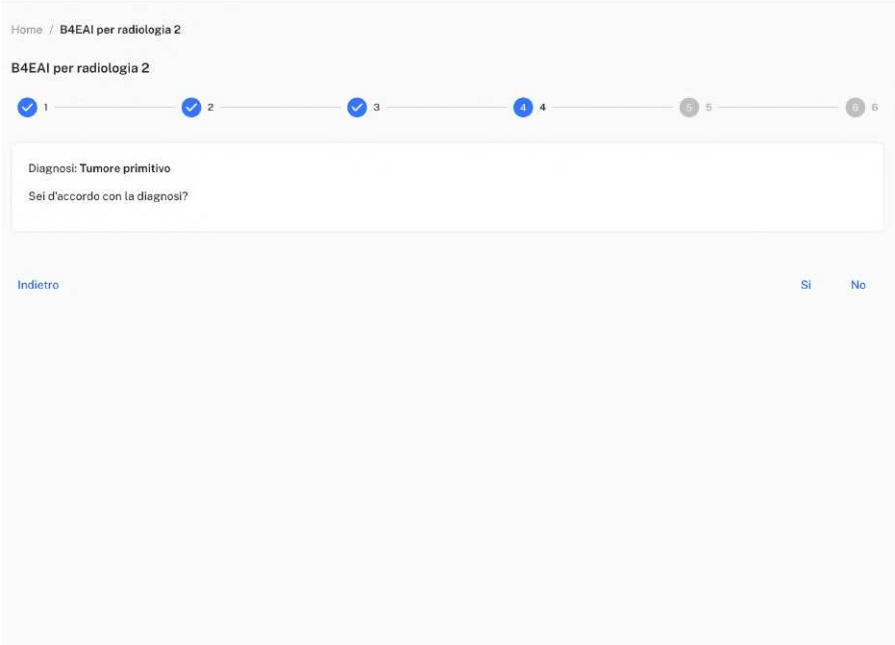


Figure 10.9: A screenshot from the Oracular prototype

- Semi-Structured Interview (Phase 3): Participants reflected on usability, workflow alignment, reasoning support, trust, and perceived value. Interviews and think-aloud utterances were audio-recorded, transcribed, and anonymised.

DATA COLLECTION AND ANALYSIS The study combined quantitative survey data (adapted UTAUT/TAM constructs: performance expectancy, effort expectancy, perceived usefulness, behavioural intention), think-aloud transcripts and post-task semi-structured interview transcripts.

Primary quantitative outcomes were behavioural intention to use and performance and effort expectancy. Within-participant differences between prototypes were analysed using Wilcoxon signed-rank tests with Holm correction. Given the small sample size, statistical analyses were treated as exploratory, with emphasis on effect direction and magnitude.

Think-aloud and interview transcripts were analysed using a hybrid inductive–deductive thematic analysis. An initial deductive codebook was derived from the research questions and prior literature (e.g., usability, workflow integration, trust, adoption). Two researchers independently coded the data via *Atlas.ti*, resolving discrepancies through discussion. A second inductive coding pass identified emergent subthemes within each category, which were iteratively refined to ensure internal coherence and analytic distinction.

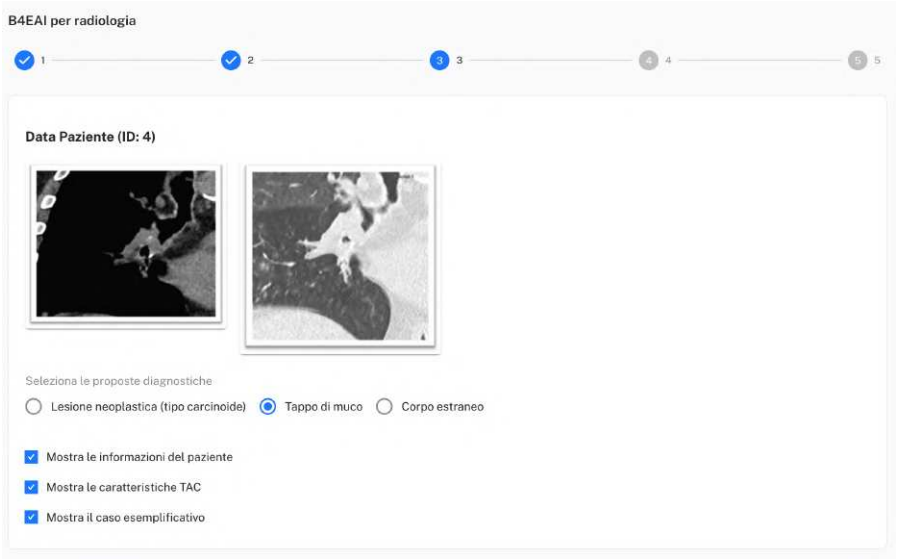


Figure 10.10: A screenshot from the Evaluative prototype: diagnosis selection

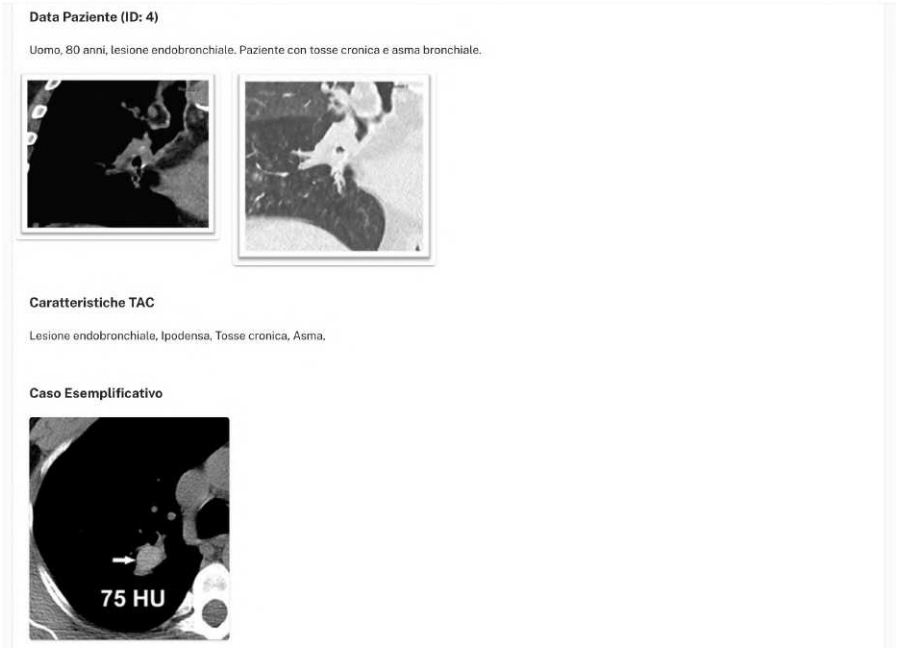


Figure 10.11: A screenshot from the Evaluative prototype

10.2.1.2 *Results*

PARTICIPANTS PROFILE AND BASELINE ATTITUDES The study involved eleven practising radiologists with heterogeneous seniority (ranging from 0–5 years to over 20 years of clinical experience). The cohort was largely familiar with clinical AI: nine participants reported using AI tools at least weekly, and seven reported daily use. Baseline technology anxiety was low (M = 1.37, SD = 0.38 on a 5-point scale). Therefore, participants evaluated the prototypes not as AI-naïve users, but as clinicians with practical exposure to AI-augmented workflows.

QUANTITATIVE RESULTS Acceptance and adoption were assessed using an adapted UTAUT/TAM questionnaire, aggregated into eight constructs. Within-participant comparisons were conducted between the Evaluative and Oracular prototypes using Wilcoxon signed-rank tests. Given the small sample size ($N = 11$), p-values were treated as exploratory; emphasis was placed on median differences and effect sizes.

Across all constructs except *Anxiety*, median scores favoured the Evaluative prototype.

Facilitating Conditions showed the clearest and most robust quantitative difference. Median scores (3.4for the Oracular, to 5.0 for the Hypothesis-drivem), with a large mean paired difference (+1.521.34). This contrast remained statistically significant after Holm correction ($Z = 2.8$, $p = .002$; p Holm $= .016$), with a very large effect size ($r = .85$). This indicates that participants perceived the evaluative interface as substantially easier to integrate into their clinical workflow and organisational context.

Performance Expectancy, *Job Relevance*, and *Result Demonstrability* all showed consistent directional gains in favour of the Evaluative prototype. Median scores for the hypothesis-driven prototype were around 0.7˜1.0 points higher across these constructs. While none of these differences survived Holm correction, effect sizes were medium to large (r ranging from .62 to .80), indicating practically meaningful advantages despite limited statistical power. This suggests a tendency by practitioners to perceive the hypothesis-driven interaction paradigm as more useful for accomplishing diagnostic tasks, more relevant to their professional role, and more effective in making outcomes understandable and communicable.

Attitude Toward Use showed a positive median shift (3.4 for the Oracular, 4.0 for the Evaluative prototype) with a medium effect size ($r = .55$). Behavioural Intention exhibited a larger median increase (3.0, 5.0), but with substantial inter-participant variability (SD = 2.05) and a smaller effect size (r = .42). This suggests that while many par-

ticipants expressed stronger intent to use the evaluative prototype, adoption intentions were more heterogeneous and likely contingent on contextual factors such as case type, time pressure, and system integration.

Effort Expectancy showed minimal differences between prototypes, with both receiving high median scores (4.2). This suggests that the evaluative interface was not perceived as meaningfully more effortful than the oracular one, despite requiring engagement with multiple hypotheses.

Anxiety scores were unchanged at the median (5.0 for both conditions), indicating that neither interaction paradigm differentially increased stress or apprehension.

QUALITATIVE RESULTS Participants' accounts repeatedly framed diagnosis as a *holistic*, *probabilistic* workflow and expressed discomfort when AI behaved as if it could deliver a single definitive answer. They emphasised that radiological judgement typically involves differential reasoning, corroboration across the study, and (often) downstream confirmation (e.g., follow-up tests or multidisciplinary discussion). Consequently, they saw AI as useful only insofar as it fits into this multi-step sensemaking process. Within this framing, participants strongly positioned AI as *complementary decision support*: a second reader, safety net, or hypothesis generator that can flag overlooked findings or broaden differentials, while leaving responsibility and final judgement with the clinician.

Perceptions of value and preferred interaction style depended on expertise, learning, and trust conditions. Many warned that overly frictionless, "answer-giving" support risks undermining learning—especially for trainees—whereas outputs that invite deliberation (e.g., multiple hypotheses and illustrative references) were seen as promoting reflective reasoning and confidence, even if they required an extra minute of effort. Trust was described as conditional and situational: agreement with one's own impression could bolster confidence, while disagreement could disrupt workflow and "block" progress unless accompanied by transparent rationale. Overall, adoption intent was cautiously positive but contingent: participants prioritised improvements in diagnostic quality and reporting support over sheer throughput, and stressed that real-world uptake would depend on workflow integration, reliability, transparency, and practical constraints such as cost.

10.2.1.3 *Limitations*

This study was conceived as a proof-of-concept exploration, intended primarily as a design probe to elicit radiologists' perceptions, expecta-

tions, and reasoning strategies in response to contrasting interaction paradigms. As such, it does not aim to evaluate diagnostic performance or generalisable behavioural outcomes, but rather to surface conceptual insights about workflow alignment and professional meaning.

Several limitations follow from this framing. First, the simulated interaction mode necessarily abstracted from the full complexity of radiological work. MRI studies were presented as static two-dimensional images, without the dynamic navigation, cross-referencing, and contextual information (clinical notes, prior studies) that typically shape real-world interpretation. Consequently, the task environment represented only a partial and idealised subset of actual diagnostic practice.

Second, while participants were encouraged to think aloud and reflect on usability, their evaluations inevitably involved projective reasoning—imagining how such systems might integrate into their real work rather than experiencing them under authentic temporal or organisational constraints. The findings should therefore be read as indicative of conceptual alignment rather than as evidence of actual behavioural change.

Third, the small sample ($N = 11$) drawn from a single regional hospital limits the diversity of institutional contexts and subspecialties represented. Future studies should broaden participation to include other modalities, organisational settings, and levels of AI maturity, and incorporate longitudinal or in-situ deployments to observe evolving appropriation over time.

10.2.1.4 *Summary and design implications*

The implications of this go beyond interface preference. They speak to how clinical AI systems participate in the reproduction or reconfiguration of professional identity. The oracular prototype implicitly re-positions the radiologist as a verifier of machine claims, rather than a generator of diagnostic meaning. The evaluative prototype, when correctly scoped, reinforces the radiologist's role as the principal sense-maker. Our findings suggest that radiologists will not accept systems that flatten diagnostic complexity into prescriptive answers, nor systems that obscure the evidence underlying recommendations. They will, however, adopt systems that help them reason better.

Drawing from these insights, we have translated participants' expectations into practical design implications, of which a more compact version is reported in Table 10.1. These recommendations foreground integration into existing PACS-based reading environments, display of clinical context at the point of interpretation, a differential-first

presentation of hypotheses, adaptive friction mechanisms, support for disagreement resolution, and human-in-the-loop report authoring.

Table 10.1: Summary of key design implications for radiology-oriented decision support (P# = participant).

Design Implication	**What to Build**	**Rationale & Use Context**
Native PACS integration	AI overlays embedded directly in PACS with multi-slice, MPR and zoom; toggleable contours/heatmaps.	Radiologists reason across slices; external viewers break workflow (P1, P5). Suitable for routine reading.
Dual-mode support (Fast / Exploratory)	A toggle offering either succinct suggestion or ranked differentials with exemplars.	Experts favour speed; trainees require depth (P2, P7, P8). Use according to case complexity.
Differential-first presentation	Present 3–5 candidate diagnoses with supporting and opposing cues.	Radiology work is probabilistic (P7, P1). Use in uncertain findings.
Evidence-centred explanations	Highlight lesion attributes; show similar cases and relevant clinical factors behind ranking.	Transparent reasoning improves trust calibration (P4, P8, P10). Use when accepting or challenging AI.
Constructive friction prompts	Gentle "second look" reminders and quick justifications before report acceptance.	Encourages active reasoning and avoids over-reliance (P3, P5, P7). Use in ambiguous or high-risk situations.
Report co-authoring	Editable draft findings and structured insertion of differentials and evidence snippets.	Supports efficiency without removing accountability (P8, P2). Suitable for routine reporting and teaching.

10.2.2 *Further work: Supporting radiological* vision work *through* open, multiple, adjunct *AI support*

Bibliographic reference

Anichini, G., **Natali, C.**, & Cabitza, F. (2024). Invisible to machines: designing AI that supports vision work in radiology. Computer Supported Cooperative Work (CSCW), 33(4), 993-1036.[3]

This article presents an empirically grounded analysis of how AI-based automatic detection tools are appropriated within radiological practice. Its central claim is that dominant evaluation paradigms for medical AI—most notably human–machine performance comparison—systematically overlook the socio-material, tacit, and normative dimensions of radiological work. As a result, they misrepresent both what radiologists do when they "see" medical images and how AI systems intervene in, reshape, or sometimes undermine this work.

We challenged the widespread techno-scientific narrative that frames AI as a neutral, objective, and accuracy-enhancing replacement for human perception in radiology. Instead, we argue that the production of the "visible" in medical imaging is inseparable from a large body of invisible work: interpretive, contextual, ethical, emotional, and organisational activities that cannot be reduced to image-level pattern recognition. This invisible work is not incidental but constitutive of safe and meaningful clinical decision-making.

Against this background, the paper asks a precise and consequential question: *under what conditions do the norms embedded in AI systems converge with, or diverge from, the professional norms that structure radiological vision work?*

This study is the result of an an 18-month ethnographic account of how two AI-based detection tools are received, negotiated, and appropriated in radiological practice. Through interviews and in-situ observations with 17 radiologists/residents and two start-up staff across multiple institutions performed by a sociologist, we traced the friction between automation's promise of standardised image classification and the situated, socially organised "vision work" of radiologists.

The two AI systems examined are AIT, an automatic chest X-ray anomaly detection system (Figure 10.13), and AIM, a mammography support tool for breast cancer screening (Figure 10.12).

3 Open access available at https://link.springer.com/article/10.1007/s10606-024-09491-0 or https://www.boa.unimib.it/handle/10281/487660

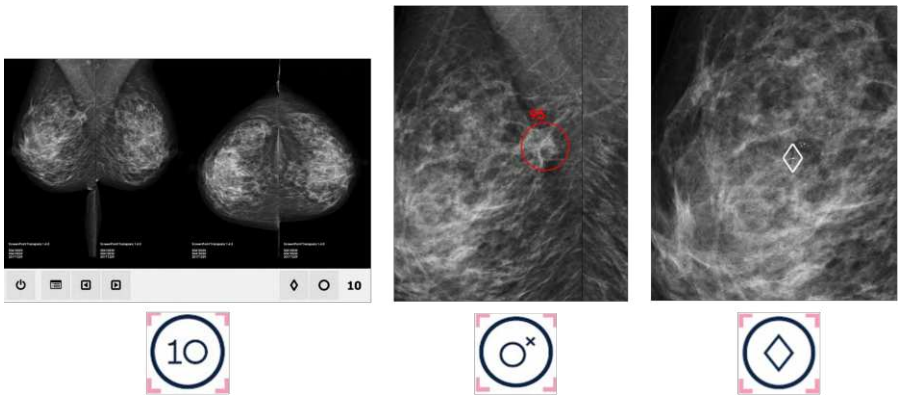


Figure 10.12: A screenshot from the breast cancer detection system AIM

THE UNDATAFIABLE DIMENSION OF RADIOLOGICAL WORK A central empirical finding is that radiologists routinely rely on information that remains inaccessible to AI systems. This includes patient history, family background, prior imaging exams, physical examination, and patient-reported sensations. The pathological meaning of a visible anomaly emerges only through its relation to this broader context.

AI systems may arrive at the same classification as the radiologist, but often for fundamentally different reasons. This mismatch generates both epistemic tension and professional dissatisfaction, as the AI's "reasoning" fails to align with clinicians' situated understanding of the case.

WRITING PRACTICES AS NORMATIVE WORK Medical reporting is shown to be a key site where vision work is stabilised and made actionable. Radiologists deliberately omit, emphasise, or contextualise findings to guide downstream clinical decisions, manage risk, and maintain professional credibility. AI-generated reports, by contrast, tend to over-describe, pathologise benign findings, or fail to account for communicative intent. Therefore, diagnosis is not merely about detection, but about responsible representation.

SUPER-NORMALITY AND ALARM FATIGUE We identified a structural bias in AI systems towards a form of "super-normality", whereby bodies that deviate from training-set norms—due to surgery, treatment, or medical devices—are repeatedly flagged as pathological. This leads to false positives, alarm fatigue, and, paradoxically, to the risk of errors of omission when clinicians begin to trust AI silence more than their own concerns.

TACIT KNOWLEDGE, INTUITION, AND OVER-STRUCTURING Perhaps most critically, we reported how AI systems can disrupt the

Observations sur radiographies de thorax et estimation

IMPRESSION

Examen anormal

OBSERVATIONS	PRÉSENCE	LOCALISATION
Anormal	☑	
Parenchyme pulmonaire		
Opacité	☑	
Cavité	☑	MG, IG, ID
Consolidation	☑	MG, IG, ID
Nodules		
Fibrose		
Autres opacités		
Médiastin		
Cardiomégalie		
Proéminence des régions hilaires		
Plèvre		
Épanchement pleural		
Emoussement de l'angle costophrénique		
Pneumothorax	☑	

Figure 10.13: A screenshot from the pneumothorax detection system AIT

acquisition and exercise of tacit knowledge. Junior radiologists fear that early exposure to automated classifications will short-circuit learning. Senior clinicians report occasions where AI reassurance dampens intuitive concern. In both cases, automation introduces a subtle overstructuring of vision work, privileging machine-detectable patterns over experiential judgment.

FRICTION AS ALIGNMENT WITH VISION WORK Beyond deconstructing dominant narratives of seamless AI integration, the study contributes design implications under a Frictional AI lens: openness to undatafiable context and uncertainty, multiplicity of outputs over single verdicts, and auxiliarity that positions AI as a supportive colleague.

Together, these principles align decision support with the epistemic, social, and pedagogical realities of radiology, aiming to preserve professional judgement and enable durable, context-sensitive adoption.

The paper identifies three core principles of Frictional AI—openness, multiplicity, and auxiliarity—each directly traceable to the empirical findings.

Design principle	Design requirements	Related themes	Interview excerpts
Openness	AI systems should be transparent and sensitive to surrounding context, including its complexities, ambiguities, and uncertainties; allow human input and feedback; encourage users to integrate outputs with professional judgement while seeking additional contextual information beyond the digital output.	*The Undatafiable Dimension*, *From tacit knowledge to over-structuring*, and *Purposeful Omission for Useful Representation*: AI systems should be transparent about what they can and cannot capture, nudging users beyond clear-cut categories to seek additional information beyond the computer screen, and allowing omission of irrelevant elements (depending on context). It also touches on *Super-normality* through the proposed introduction of iatrogenic anomalies in training data.	"if I see an opacity on the mammogram, I'll look for it, I'll palpate it in that area, and I'll see that there's a mass, is it soft, is it hard?..."; "Finally, what should be integrated into AI is the patient's history ... family history... We will also take into account her age, of course..."; "There are a lot of recommendations and indications to follow. But in the end there is a patient and a doctor in front of him or her ... here are a whole bunch of things to keep the human being in check, but a lot of the time it's still experience..."; "... you don't have to describe them... no need to because it's irrelevant, it has no influence, it won't change anything..."

Table 10.4: Summary table presenting a short definition for the *Openness* principle, the related themes, and examples of interview excerpts that informed the design implications.

Design principle	Design requirements	Related themes	Interview excerpts
Multiplicity	AI outputs should be presented in multiple, varied, and complementary forms; support interpretation and integration in medical decision-making on the basis of the inherent complexity of the medical phenomenon; enable more nuanced decisions and reduce reliance on singular algorithmic recommendations.	*Purposeful Omission for Useful Representation* and *The Undatafiable Dimension*: emphasis on the value of complementary AI outputs, taking contextual elements into account, and applying comparison-based reasoning. Multiplicity is connected to *Alarm Fatigue* through the provision of multiple cognitive-activating outputs as a countermeasure to heuristic disruption and false alarms that frustrate professionals and hinder novices from effective learning.	"Above all, it is the comparison that will help us"; "Don't give me decisions! give me probabilities! . . . I like more the AIM approach, where they give me a score, and it keeps the decision to a certain extent to the human being. And he comes to help me by saying, he sees that there is a 60% or 80% or 95% probability, and that alerts me! But the decision is mine"; "If we are in the initial assessment of a cancer and the patient's liver, for example, shows a metastasis and other lesions that are cysts, we will describe them . . . We will describe them even if they are of no interest because they are present with something that is . . . malignant"

Table 10.6: Summary table presenting a short definition for the *Openness* principle, the related themes, and examples of interview excerpts that informed the design implications.

Design principle	Design requirements	Related themes	Interview excerpts
Auxiliarity	AI should serve as a supportive tool that augments and complements human expertise without replacing it; perform ancillary tasks that facilitate diagnostic reasoning and empower professionals to make informed decisions; prioritise professional insights and intuitions.	*From Tacit Knowledge to Over-structuring* and *Super-normality and Alarm Fatigue*: AI is positioned as supportive, enhancing rather than disrupting human expertise. It also touches on *Purposeful Omission for Useful Representation* insofar as auxiliarity protects professional discretion and heuristics in the preparation of the radiological report.	"When I see that AIM hasn't analysed it or hasn't noticed it, I say to myself that there's no need to do it . . . I can skip it without any problems. . . "; "Sometimes I can see opacities that it (AIM) does not detect, so it is reassuring for us to know that it is . . . not suspicious"; "(the medical report) shouldn't be a stereotypical thing . . . the surgeons who receive it will know whether they can potentially operate the patient or not! They know that you are a specialist"; "(AIM) sometimes detects masses and assumes that they are pathological, but I often reinterpret them and say, 'No, it's scar tissue'. It doesn't know the patient's history, it can't make that assessment . . . Except that we actually know that this opacity is scar tissue and so we're able to correct the diagnosis"; "There are plenty of things that come into play, it's even intuition actually, when discussing with the patients"; "The machine, the AI must stay in its place. It shouldn't tell us to do this, do that. . . "

Table 10.8: Summary table presenting a short definition for the *Openness* principle, the related themes, and examples of interview excerpts that informed the design implications.

Openness is reconceptualised beyond transparency or explainability. It refers to openness to context, uncertainty, and human intervention (Table 10.4). Given that radiological judgment depends on undatafiable elements, Frictional AI systems should represent uncertainty explicitly, including through vague or non-deterministic visualisations; Support bidirectional interaction, allowing clinicians to provide feedback and correct outputs; Act as case-mining tools rather than authoritative classifiers. This form of openness introduces epistemic friction by preventing premature closure around a single "correct" interpretation.

Radiological decision-making is inherently collective and iterative. The Frictional AI principle of *Multiplicity* therefore supports multiple hypotheses rather than singular outputs, mirroring collegial practices such as joint interpretation and postponed decision-making (Table 10.6). By resisting the drive towards singular optimisation, multiplicity keeps diagnostic reasoning open and contestable.

Finally, Frictional AI is positioned as *auxiliary* rather than substitutive. The system's role is not to replace human judgment but to provoke it, support it, and sometimes challenge it. This auxiliary stance directly counters the observed over-structuring of vision work and reframes AI as a partner in deliberation rather than an oracle (Table 10.8).

Part III

CONCLUSION

11

CONCLUSION

> *Tolerating some congestion, some friction, some inefficiency, even some transaction costs may be necessary to sustain an underdetermined environment conducive to human flourishing.*
> — Brett Frischmann & Evan Selinger,
>
> *Re-Engineering Humanity* (Frischmann and Selinger, 2019)

11.1 SUMMARY OF RESEARCH CONTRIBUTIONS

This thesis advances an interaction-centred account of Artificial Intelligence in high-stakes decision-making, arguing that the benefits and risks of AI in clinical practice cannot be understood at the level of algorithms alone, but emerge from the protocols of interaction through which human judgement and machine output are coordinated. Across conceptual, methodological, and empirical contributions, the thesis demonstrates that how AI participates in decision-making is at least as consequential as how accurate it is.

HUMAN–AI INTERACTION AS AN EMERGENT PHENOMENON The first contribution of this thesis is conceptual. It reconceptualises Human–AI collaboration through an emergence-oriented lens, treating intelligence, competence, and responsibility as properties of sociotechnical configurations rather than of isolated humans or machines.

Across the thesis, Human–AI Interaction Protocols are established as a central analytical and design unit. Rather than framing AI as a tool that either assists or replaces human cognition, the work shows that different protocols—such as adjunct, oracular, human-first, evaluative, and collective configurations—produce qualitatively different forms of hybrid intelligence. This perspective is articulated most explicitly in the comparative study of individual and collective AI use, where AI is shown to act as a performance leveler rather than a dominant authority, particularly when embedded in group deliberation rather than individual decision-making.

A second conceptual contribution concerns the risk of AI-induced deskilling and epistemic sclerosis. Through a mixed-method literature review, the thesis synthesises empirical evidence showing how AI systems can erode clinical skills not only through direct substitution, but through upskilling inhibition: the gradual removal of opportunities for judgment formation, uncertainty management, and learning. By

situating these effects within medical competency frameworks, the thesis moves deskilling from an abstract ethical concern to a concrete socio-technical risk.

Together, these contributions shift the focus of Human–AI research away from optimisation narratives and toward the sustainability of professional expertise, positioning hybrid intelligence as something that must be actively designed and maintained rather than assumed to emerge automatically from better models.

EVALUATING HUMAN–AI INTERACTION BEYOND ACCURACY Building on empirical studies in radiology and logic-based tasks, the thesis argues for a beyond-accuracy evaluation framework that captures how AI reshapes human reasoning rather than merely outcomes. This brings together metrics and analytical consturcts presented in the literature, such as reliance patterns, appropriate reliance, and overall technology impact.

This framework is applied across multiple studies, including the analysis of eXplainable AI: findings show that explanations can both improve accuracy and, under certain conditions, mislead users. The identification and formalisation of phenomena such as the white-box paradox and the XAI halo effect provide concrete evidence that explanations are interactional interventions with persuasive power, not neutral transparency devices.

Methodologically, the thesis therefore contributes a shift in evaluative stance: from asking whether AI systems are correct, to asking how they influence human judgement, confidence, and learning trajectories over time.

FRICTIONAL AI IN CLINICAL DECISION-MAKING The third contribution of the thesis is empirical and design-oriented. Through controlled studies and mixed-method investigations in radiology, the thesis introduces and evaluates Frictional AI as a principled interaction design approach.

Frictional AI deliberately departs from dominant design ideals of seamlessness and cognitive ease. Instead, it introduces desirable difficulties—such as delayed disclosure, comparative evidence, multiple hypotheses, and second-opinion protocols—to sustain reflective engagement and mitigate inappropriate reliance. This design philosophy is empirically instantiated in several studies. As *Adjunct AI* or *AI-first* protocols, AI is marginalised to a second-opinion, ancillary role in the decision-making process (Natali et al., 2024). In investigations of pro-hoc explanations, AI systems present alternative outcomes and

exemplar cases (Cabitza et al., 2024b) rather than a single definitive recommendation, encouraging clinicians to reason comparatively rather than deferentially. As *Reflective XAI*, the frictional AI support aimed at inducing doubt in the user, showing examples of similar cases that had been previously misclassified by the system (Cabitza et al., 2023b).

In comparative studies of oracular versus evaluative interfaces, radiologists consistently prefer evaluative designs that align with hypothesis-driven diagnostic reasoning and preserve professional agency, particularly in ambiguous cases. Here, qualitative findings further show that clinicians appropriate friction strategically, valuing it as a safeguard for autonomy and learning rather than as a usability defect (Rubegni et al., 2025).

Crucially, these studies demonstrate that frictional designs can achieve non-inferior performance while reducing automation bias and supporting confidence calibration. This challenges the assumption that efficiency and cognitive ease are always desirable in high-stakes settings, and provides empirical grounding for friction as a design resource rather than a usability failure.

Taken together, the thesis makes an integrative contribution to CSCW, HCI, and medical AI by reframing what it means for a Human–AI system to be successful. Success, in this account, is not defined solely by predictive accuracy or user satisfaction, but by whether a system preserves human judgement and responsibility, distributes competence equitably across individuals and groups, supports learning rather than eroding skill, and remains resilient under uncertainty and failure.

Beyond its individual publications, the original contribution of the thesis lies in integrating these strands into a coherent, emergence-oriented framework for Human–AI interaction—one that is attentive not only to what AI can do, but to what it does to practice. This synthesis positions Frictional AI as a transferable design paradigm for any domain in which preserving human judgement, accountability, and skill is essential.

At the thesis level, the central claim is that effective human–AI collaboration in high-stakes domains requires a paradigm shift away from frictionless automation and toward designs that deliberately sustain and amplify human expertise. Rather than treating AI as an infallible oracle or merely a tool to be made ever more seamless, this view positions AI as a thoughtful partner that provokes critical thinking, preserves professional skill, and supports human agency. This stance echoes a long tradition in HCI and CSCW warning against "underrating the skills and competencies required in even the most

routine of tasks" (Hartswood et al., 2003, p. 249). Indeed, early work by Winograd and Flores argued that technology design is ontological: new systems quietly reshape how we work and even who we become (Winograd and Flores, 1986). If AI systems are designed for maximum convenience at the cost of human involvement, they risk deskilling users and altering professional practice in undesirable ways. Therefore, the thesis claims that incorporating friction or "seamful" interactions is a viable design strategy that compels users to engage their judgement, rather than passively accept algorithmic outputs. This perspective aligns with Licklider's original vision of "man–computer symbiosis" (Licklider, 2008) where tightly coupled human–machine systems would augment human problem-solving instead of replacing it. By introducing calibrated "cognitive friction" into AI tools – such as prompts that question an AI's suggestion or interfaces that mandate human input – we aim to create a synergy in which human intuition and tacit knowledge are preserved alongside algorithmic efficiency. In sum, the thesis-level claim is that intentionally designing friction into AI–human interactions is a necessary intervention to ensure that augmented intelligence is truly augmentative: sustaining human expertise, responsibility, and learning over the long run, rather than undermining them in the pursuit of short-term optimization. In this Discussion, I will provide more detail on how the thesis contributes to bridging disciplines, the implications of emergence-oriented design, a taxonomy of frictional interventions, and close with a presentation of limits and open challenges.

11.2 DISCUSSION

Bridging disciplines

A key insight underlying this work is the importance of emergence in human–AI systems – the idea that when humans and AI interact closely, the outcomes can be more than the sum of their parts. By bridging research on Hybrid Intelligence (HI) with the socio-technical tradition of Computer-Supported Cooperative Work (CSCW), we can better understand and design for these emergent outcomes. Hybrid Intelligence research, as defined by Dellermann et al. (2019), views human and machine intelligence as complementary, seeking "something gained in the process" of combining their strengths. This implies that well-designed human–AI partnerships can yield synergistic benefits – novel solutions, insights, or capabilities that neither the human nor AI could achieve alone. For example, in medical diagnosis a clinician's intuitions about a complex case might, in combination with an AI's pattern recognition, produce a more accurate assessment than either could independently. Realizing such synergy demands moving beyond simple task allocation toward interaction mechanisms that let human

insight and machine computation continually inform one another. However, achieving productive synergy also requires acknowledging the situated, social nature of human expertise – a core focus of CSCW. The CSCW community emphasizes that work is inherently cooperative and contextually shaped: people coordinate through "articulation work" and subtle forms of communication to get things done (Schmidt and Bannon, 1992).

Bridging HI with CSCW brings this perspective into the design of AI systems. It reminds us that a doctor, lawyer, or pilot does not perform tasks in isolation or by rigid protocol; rather, they rely on tacit knowledge, teamwork, and continuous adaptation to context (Polanyi, 1966). Much of an expert's knowledge is tacit, developed through experience and apprenticeship, and cannot be fully captured in explicit rules or code (Beane, 2019). This means that effective AI support must respect and incorporate the unspoken, nuanced aspects of human work instead of steamrolling them. By designing AI that works with human operators in a collaborative, context-aware manner, we enable emergent intelligence to arise from the partnership – for instance, an AI system prompting a clinician with relevant uncertainties may trigger the clinician's memory or insight in a way neither could have achieved alone.

In summary, emphasizing emergence shifts our goal from merely inserting AI into workflows to cultivating new joint capabilities and insights that arise from human–AI engagement. And bridging HI with CSCW provides the necessary theoretical and practical foundation: HI contributes the goal of combining strengths of humans and AI for mutual learning (Akata et al., 2020), while CSCW contributes an understanding of collaboration, context, and the tacit dimensions of expertise that must be respected (Schmidt and Bannon, 1992). Together, they point toward systems enable human–AI teams to achieve outcomes neither could attain alone.

11.2.1 *The implication of centering frictional protocols*

Considered together, the evidence from HCI, CSCW, and AI ethics suggests that effective human–AI interaction is not determined by model accuracy alone, but emerges from the collaboration protocol—the specific choreography of how human and machine contributions are sequenced, weighted, and presented (Cabitza et al., 2023e). This perspective shifts the focus from algorithm performance to the quality of the configuration linking human and AI: then, optimising for speed, seamlessness, or directive clarity is not always desirable. On the contrary, deliberately frictional interventions—such as slower interaction tempos, less prescriptive outputs, or protocols that require prior hu-

man commitment—become a tenable and theoretically grounded design choice insofar as it maintains human awareness and judgment. While such designs may appear less efficient or more cumbersome at the interface level, they aim to preserve critical forms of human engagement that are essential in high-stakes decision-making.

What Is Lost

SEAMLESSNESS AND COGNITIVE EASE The primary "loss" is the traditional ideal of the frictionless, invisible interface. Designing for robust human–AI assemblages often requires introducing Frictional AI elements—deliberate "desirable difficulties" or programmed inefficiencies—to keep humans actively engaged in critical thinking. This runs counter to the prevailing ethos of software design that prizes minimal effort and "don't make me think" (Krug, 2000) usability. Instead of seamless automation, users may be forced into slower, more effortful analytical reasoning (System 2 thinking, Kahneman 2011) in order to scrutinize AI outputs. While such friction can improve decision quality, it undeniably sacrifices the immediate cognitive ease and speed that busy professionals (e.g. overworked clinicians) often desire; thus, an assemblage approach may trade some user convenience and efficiency for enhanced cognitive vigilance.

APPARENT EFFICIENCY By foregoing seamless automation, frictional design may also seem less efficient in the short term. Traditional evaluations often highlight time saved or workload reduced by AI (Hart and Staveland, 1988); in contrast, a frictional workflow might initially take longer or use more resources (e.g. requiring double readings, additional documentation of reasoning). For example, a clinical decision support that slows down the user's decision process deliberately might appear to undermine efficiency gains. Economic pressures and productivity metrics in healthcare can therefore conflict with these design principles. It is important to note, however, that these losses—speed, simplicity, and seamlessness—are often necessary trade-offs to achieve long-term safety and human expertise preservation. As (Winograd and Flores, 1986) presciently observed, every tool we design is ontological, shaping our practices and defining what we value. Designing frictionless AI may maximize immediate convenience at the cost of eroding the very human skills and awareness that make the system effective over time.

What Is Gained

RESILIENCE AND EXPERTISE PRESERVATION An emergence-oriented approach explicitly values the preservation of human expertise and long-term system resilience. By keeping humans in the loop in a substantive way, we guard against the risks of deskilling and cognitive

complacency. Instead of humans gradually atrophying into passive overseers, they continue to practice and refine their domain skills (Rafner et al., 2022).

Over time, such configurations can counteract the tendency of AI to induce complacency. They foster antifragility (Taleb, 2012) – the system gets stronger under stress because the human operators continue learning and can adapt when the algorithm encounters novel situations. In sum, we gain a more adaptive and robust partnership. The system is not brittlely reliant on AI; it retains a human in the loop who is competent to take over or correct course when needed

HOLISTIC AND CONTEXTUAL AWARENESS Another benefit of the emergence-oriented paradigm is improved situated awareness and interpretability. Rather than focusing on narrow predictions, the interaction protocol can be designed to contextualize AI outputs, cross-check them against human knowledge, and expose why disagreements occur. This addresses the issue that raw model accuracy often ignores context. As Winograd & Flores argued in the 1980s, each design embodies assumptions about cognition and action (Winograd and Flores, 1986). Centering the protocol allows those assumptions to be surfaced and negotiated in practice. For example, a clinical AI that always outputs a single probability may narrow a clinician's framing of the problem ("premature closure"). But a protocol that invites the clinician to articulate their own hypothesis first, or that presents evidence for multiple possible conclusions, maintains a broader view of the problem space. Such practices yield rich interpretive dialogues rather than one-shot answers. Over time, this can expand the clinician's expertise (as the AI exposes them to patterns they hadn't considered) while also keeping the AI aligned with evolving human insights (as developers observe where the AI confuses or disagrees with experts). In essence, we gain a system that is not only about answer accuracy but about collective sense-making.

TECHNOVIGILANCE Finally, centering interaction protocols equips us to practice technovigilance – a continual monitoring of not just the AI's performance, but the entire human–AI system's behavior and impact. This concept will be expanded in the next section, but it is worth noting here as a "gain" of the assemblage view. By treating the human and AI as a coupled unit, we become attuned to metrics beyond accuracy (e.g. how often the human overrides the AI, how decision time or confidence is affected, how skills change over months of usage). We start to see success in terms of appropriate reliance and maintenance of human agency, rather than just raw correctness at a single point in time. The benefit is a culture of proactive vigilance: issues like creeping deskilling, knowledge ossification, or emergent biases can be detected early and addressed through design adjustments or

policy interventions. In a traditional approach, these problems might go unnoticed until a major failure occurs. Thus, the protocol-centered paradigm inherently promotes a safer, more reflective mode of innovation.

In summary, moving to an emergence-oriented focus means relinquishing some beloved ideals of frictionless design and simplicity. In return, we gain emergent team intelligence, system resilience, and a richer, human-centered notion of success. As Kasparov's Law and real-world studies attest, a mediocre AI embedded in a good workflow can outperform a great AI in a bad workflow (Cabitza, Campagner, and Sconfienza, 2021; Kasparov and Greengard, 2017). The emergence-oriented view compels us to evaluate AI in context, focusing on team performance, trust calibration, skill sustainability, and safety under real-use conditions.

11.3 LIMITS AND OPEN CHALLENGES

While the argument for friction and protocol-centered design is compelling, it is not a panacea. There are important limits and open challenges to acknowledge.

First, not all tasks or contexts can accommodate friction. What works in radiology (where decisions are image-based and time pressure is moderate) may not directly work in settings such as anesthesiology (where decisions are second-to-second). In time-critical scenarios, even small delays or extra steps can be costly or unacceptable. The design challenge is to achieve calibration without impeding necessary speed. We must recognize there are domains where frictionless automation is valued for survival, and the burden shifts to robust training and fail-safe mechanisms outside the moment of operation.
Second, poorly executed friction can backfire. As noted, users may develop workarounds or resentment if they feel the system is cumbersome without clear benefit. There is a risk of alert fatigue or "crying wolf"– if the system forces frequent checks on the AI and 99% of the time the AI was right, the user can become desensitized and might start rubber-stamping anyway. Designing friction thus requires careful tuning and iteration with user feedback, and possibly personalizing the friction to each user's needs (an expert might bypass some steps that a novice cannot). It is an open challenge to find the right metrics for when friction is helping versus when it's just adding noise.

Lastly, organizational and policy constraints can limit friction-oriented designs. Adding AI-related friction could be seen as yet another burden unless it is well-integrated. Regulatory environments sometimes mandate explicit procedures for using decision aids – those rules

might lag behind and even inadvertently discourage optimal friction. For instance, if a hospital policy said "always follow the AI for certain screenings to save time", that would conflict with our principle of friction for safety. Advocating for these design changes thus involves convincing not just end-users, but also institutional stakeholders and policymakers about the long-term value.

EVEN "PERFECT AI" NEEDS HUMAN ENGAGEMENT Looking ahead, it might be tempting to assume that as AI models get ever more accurate – approaching "perfect" prediction – the need for human oversight and friction will diminish. On the contrary, the synthesis of current thinking suggests that even a hypothetically near-perfect AI would still require friction and human monitoring to ensure safe and effective use. There are several reasons for this, grounded in the unpredictable and context-rich nature of real-world practice.

First, no AI can be truly perfect across all scenarios; there will always be edge cases and novel situations. Human experts serve as a critical safety net in these uncharted waters. If a system has operated at 99.9% accuracy for years, the human operators might be extremely unprepared for that 0.1% scenario unless they have remained mentally engaged. A future "super-AI" that usually nails the diagnosis might misfire on a once-in-a-decade exotic disease. Only a clinician who has maintained diagnostic curiosity and breadth could catch that. Thus, even as AI performance soars, human vigilance must be maintained.

Moreover, as AI takes over routine tasks, the human role may evolve to focus on higher-level judgment, empathy, and ethical decisions that AI cannot handle. Aslam and Hoyle (2022) describe that clinicians will need to cultivate " their human skills that are beyond the capabilities of the AI system"(Aslam and Hoyle, 2022, p. 71) – for instance, integrating patient preferences, understanding contextual factors, and managing scenarios with sparse data or atypical presentations. In the future, a "perfect" AI diagnostic engine might give an answer, but the clinician of the future might be more like a pilot of the case, evaluating whether that answer fits the messy reality of the patient.

Future policy and governance will likely mandate some of these practices. We are already seeing guidelines emphasizing human oversight as a requirement for high-risk AI systems (European Parliament and the Council of the European Union, 2024). Regulators recognize that accountability ultimately lies with human professionals and organizations, so they may require features like traceability of AI recommendations, second reads on certain AI-driven decisions, and periodic audits of AI impact on outcomes. Policy might also enforce

that professionals using AI maintain competency in core skills. This kind of policy would institutionalize technovigilance: it sends a clear signal that AI is an aid, not a crutch to lean on blindly.

11.4 CLOSING REMARKS

> *By "augmenting human intellect" we mean increasing the capability of a man to approach a complex problem situation, to gain comprehension to suit his particular needs, and to derive solutions to problems. [. . .] We do not speak of isolated clever tricks that help in particular situations. We refer to a way of life in an integrated domain where hunches, cut-and-try, intangibles, and the human "feel for a situation" usefully co-exist with powerful concepts, streamlined terminology and notation, sophisticated methods, and high-powered electronic aids.*
> — Douglas C. Engelbart, *A conceptual framework for the augmentation of man's intellect* (Engelbart, 2023, p. 13)

This thesis set out from a deceptively simple observation: in medicine, as in other high-stakes domains, it is not the predictive power of AI in isolation that matters, but the quality of the human–AI process through which clinical judgements are formed. Across the chapters, I have argued that current AI deployments too often privilege seamlessness, immediacy, and cognitive ease, creating conditions in which automation biases, conservatism biases, and skill erosion can quietly accumulate. Against this background, this thesis advanced Frictional AI as a principled alternative: a design paradigm that treats interaction friction not as a usability flaw but as an epistemic resource for sustaining diagnostic reasoning, preserving professional agency, and enabling hybrid intelligence to emerge.

Throughout the thesis I have insisted that the real design unit in medical AI is not the model, but the interaction protocol: who sees what, in what order, with what opportunity for contestation, and under what evaluative regime. This is closely aligned with long-standing CSCW calls to "avoid underrating the skills and competencies that are required in even the most routine of tasks" and to decide deliberately "what to automate and what to leave to human skill and ingenuity" (Hartswood et al., 2003). The frictional perspective operationalises these calls. It says: we should automate less than we can, and we should do so more slowly and more transparently than current AI rhetoric suggests. This is, in Luciano Floridi's terms, a *festina lente* stance—make haste slowly—towards AI integration (Floridi, 2021).

This stance is not only ergonomic or pedagogical; it is ethical. Gosline's notion of "acts of inconvenience" as an antidote to dark patterns provides a vocabulary for this ethics (Gosline, 2022). If manipulative design uses friction strategically to channel users into choices that benefit the platform, then responsible AI can and should use friction strategically to channel users back into reasoning, doubt, and verification—i.e. into the very cognitive activities that sustain clinical safety and professional growth. In other words, if dark patterns weaponise ease, frictional AI weaponises deliberation.

This ethical inflection also connects the thesis to the discourse on Meaningful Human Control (Cavalcante Siebert et al., 2023; Robbins, 2024). As Suchman and Thimm (2024) notes, MHC has been "really important in *slowing things down*", forcing actors to ask what "meaningful" actually means in specific socio-technical settings. The frictional protocols proposed here are one practical way of "slowing things down": they insert micro-moments of reflection, counterfactual comparison, or evidence inspection that prevent the human from becoming a nominal overseer of an opaque process. They turn MHC from a vague governance aspiration into a concrete interaction property.

The thesis also advances *technovigilance* (Harvey, Cabitza, et al., 2018). Clinical AI today is marked by short release cycles, opacity in training data, and performance claims that are not routinely stress-tested at the point of human use. Under such conditions, it is not enough to require explainability or to demand post-market surveillance of models. We must equally monitor *human* trajectories of competence, vigilance, and reliance once AI has been introduced. Technovigilance, then, is dual: it keeps an eye on models and on humans. It asks, continuously: is this system still helping clinicians to think? Is it preserving a diversity of interpretive pathways? Or is it creating epistemic sclerosis—the gradual hardening of categories and routines that makes a socio-technical system brittle in the face of novelty? (Natali and Cabitza, 2025) Frictional AI is one way of keeping the arteries open

A sceptical reader might object that friction risks slowing care or burdening already overworked clinicians. The studies reported here do not dismiss this concern; they delimit it. What emerges is that friction must be *structured*, *proportional*, and *professionally legible*. Friction that merely adds clicks will rightly be rejected. Friction that mirrors the interpretive steps clinicians already recognise—commit-before-reveal, comparative views, just-in-time evidence for dispute—was judged acceptable, even desirable. This is why I have insisted throughout on workflow alignment and on the co-design of protocols with practitioners. Friction is a powerful instrument, but like all instruments of

control it must be human-centred, contestable, and revisable.

What, then, are the implications of this thesis beyond radiology and even beyond medicine?

For AI design and HCI, we need to broaden our definition of "good interaction" beyond speed and subjective ease. In high-stakes domains, "good" may mean "epistemically effortful"—supporting sense-making, not just task completion. Friction should become a first-class design parameter, specified, justified, and evaluated.

For evaluation and regulation, the beyond-accuracy instruments proposed here offer a concrete starting point for regulators, hospitals, and AI procurers who must ascertain not only whether a system is safe, but whether it sustains human competence over time. A system that achieves short-term gains at the cost of long-term deskilling should not be considered high-quality.

For clinical organisations, AI adoption needs to be coupled with pedagogical and organisational measures that keep human expertise in active use—rotations without AI, peer review of AI-assisted reports, reflective sessions on AI errors, and local logging of reliance patterns. This is consistent with the literature on automation ironies and with the CSCW tradition of making invisible work visible.

For AI ethics and philosophy of technology, frictional AI gives operational content to calls for preserving human agency in an age of "smart" infrastructures. It shows that agency is not preserved by exhortation but by designing environments that stay underdetermined, in which the path of least resistance is not always the path taken (Frischmann and Selinger, 2019).

If we take seriously the view of intelligence as emergent from socio-technical practice, then we must design *for* that emergence. This means creating interaction protocols that invite, sometimes even oblige, clinicians to think; that refuse to make invisibly easy what ought to remain visible and effortful; that keep open, in Frischmann and Selinger's sense, the space in which human flourishing can occur (Frischmann and Selinger, 2019). It also means adopting Floridi's *festina lente* as a normative tempo for AI in medicine (Floridi, 2021): advance, but deliberately; automate, but reversibly; support, but never supplant.

BIBLIOGRAPHY

Agre, Philip (1997). *Computation and human experience*. Cambridge University Press.

Aicher, Annalena, Yuki Matsuda, Keichii Yasumoto, Wolfgang Minker, Elisabeth André, and Stefan Ultes (2024). "Enhancing reflective and conversational user engagement in argumentative dialogues with virtual agents." In: *Multimodal Technologies and Interaction* 8.8, p. 71.

Akata, Zeynep, Dan Balliet, Maarten De Rijke, Frank Dignum, Virginia Dignum, Guszti Eiben, Antske Fokkens, Davide Grossi, Koen Hindriks, Holger Hoos, et al. (2020). "A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence." In: *Computer* 53.8, pp. 18–28.

Akudjedu, Theophilus N, Sofia Torre, Ricardo Khine, Dimitris Katsifarakis, Donna Newman, and Christina Malamateniou (2023). "Knowledge, perceptions, and expectations of Artificial intelligence in radiography practice: A global radiography workforce survey." In: *Journal of Medical Imaging and Radiation Sciences* 54.1, pp. 104–116.

Ala-Luopa, Saara, Sami Koivunen, Thomas Olsson, and Kaisa Väänänen (2024). "Considerations on human-AI collaboration in knowledge work–recruitment experts' needs and expectations." In: *Proceedings of the 57th Hawaii International Conference on System Sciences*, pp. 197–206.

Alon-Barkat, Saar and Madalina Busuioc (2023). "Human–AI interactions in public sector decision making:"automation bias" and "selective adherence" to algorithmic advice." In: *Journal of Public Administration Research and Theory* 33.1, pp. 153–169.

Alter, Adam L, Daniel M Oppenheimer, Nicholas Epley, and Rebecca N Eyre (2007). "Overcoming intuition: metacognitive difficulty activates analytic reasoning." In: *Journal of experimental psychology: General* 136.4, p. 569.

Alter, Steven (2010). "Designing and engineering for emergence: A challenge for HCI practice and research." In: *AIS Transactions on Human-Computer Interaction* 2.4, pp. 127–140.

Amer, Mounia, Yassine Hilmi, and Hamza El Kezazy (2024). "Big Data and Artificial Intelligence at the Heart of Management Control: Towards an Era of Renewed Strategic Steering." In: *The International Workshop on Big Data and Business Intelligence*. Springer, pp. 303–316.

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 inter-

action." In: *Proceedings of the 2019 chi conference on human factors in computing systems*, pp. 1–13.

Ang, Beng Heng, Sujatha Das Gollapalli, and See Kiong Ng (2023). "Socratic question generation: A novel dataset, models, and evaluation." In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 147–165.

Anichini, Giulia, Chiara Natali, and Federico Cabitza (2024). "Invisible to machines: designing AI that supports vision work in radiology." In: *Computer Supported Cooperative Work (CSCW)* 33.4, pp. 993–1036.

Aquino, Yves Saint James, Wendy A Rogers, Annette Braunack-Mayer, Helen Frazer, Khin Than Win, Nehmat Houssami, Christopher Degeling, Christopher Semsarian, and Stacy M Carter (2023). "Utopia versus dystopia: Professional perspectives on the impact of healthcare artificial intelligence on clinical roles and skills." In: *International Journal of Medical Informatics* 169, p. 104903.

Aristidou, Angela, Rajesh Jena, and Eric J Topol (2022). "Bridging the chasm between AI and clinical implementation." In: *The Lancet* 399.10325, p. 620.

Aroyo, Lora and Chris Welty (2014). "The three sides of crowdtruth." In: *Human Computation* 1.1.

Aslam, Tariq M and David C Hoyle (2022). "Translating the machine: skills that human clinicians must develop in the era of artificial intelligence." In: *Ophthalmology and therapy* 11.1, pp. 69–80.

Assale, Michela, Silvia Bordogna, and Federico Cabitza (2020). "Vague Visualizations to Reduce Quantification Bias in Shared Medical Decision Making." In: *VISIGRAPP (3: IVAPP)*, pp. 209–216.

Bach, Anne Kathrine Petersen, Trine Munch Nørgaard, Jens Christian Brok, and Niels Van Berkel (2023). ""If I had all the time in the world": Ophthalmologists' perceptions of anchoring bias mitigation in clinical AI support." In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–14.

Bainbridge, Lisanne (1983). "Ironies of automation." In: *Analysis, design and evaluation of man–machine systems*. Elsevier, pp. 129–135.

Banerjee, M., D. Chiew, K. T. Patel, I. Johns, D. Chappell, N. Linton, and S. Zaman (2021). "The impact of artificial intelligence on clinical education: perceptions of postgraduate trainee doctors in London (UK) and recommendations for trainers." In: *BMC Medical Education* 21.1, pp. 1–10.

Bansal, Gagan, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld (2021). "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*, pp. 1–16.

Bardzell, Shaowen, Jeffrey Bardzell, Jodi Forlizzi, John Zimmerman, and John Antanitis (2012). "Critical design and critical theory: the

challenge of designing for provocation." In: *Proceedings of the designing interactive systems conference*, pp. 288–297.

Barrett, Teanna, Quanze Chen, and Amy Zhang (2023). "Skin deep: Investigating subjectivity in skin tone annotations for computer vision benchmark datasets." In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1757–1771.

Baselli, Giuseppe, Marina Codari, and Francesco Sardanelli (2020). "Opening the black box of machine learning in radiology: can the proximity of annotated cases be a way?" In: *European Radiology Experimental* 4.1, pp. 1–7.

Beane, Matt (2019). "Learning to work with intelligent machines." In: *Harvard Business Review* 97.5, pp. 140–148.

Beane, Matthew (2024). *The Skill Code: How to Save Human Ability in an Age of Intelligent Machines*. 1st ed. New York: HarperBusiness.

Beaulieu, Anne (2001). "Voxels in the brain: Neuroscience, informatics and changing notions of objectivity." In: *Social studies of Science* 31.5, pp. 635–680.

Benedetti, Andrea and Michele Mauri (2023). "A literature review on "friction" as a method for reflection in design interventions." In: *Convergências: Revista de Investigação e Ensino das Artes.* 16.31, pp. 139–146.

Benford, Steve, Chris Greenhalgh, Gabriella Giannachi, Brendan Walker, Joe Marshall, and Tom Rodden (2012). "Uncomfortable interactions." In: *Proceedings of the sigchi conference on human factors in computing systems*, pp. 2005–2014.

Benford, Steve, Chris Greenhalgh, Gabriella Giannachi, Brendan Walker, Joe Marshall, Paul Tennent, and Tom Rodden (2018). "Discomfort—the dark side of fun." In: *Funology 2: From Usability to Enjoyment*. Springer, pp. 209–224.

Berg, Marc (1999). "Accumulating and coordinating: occasions for information technologies in medical work." In: *Computer Supported Cooperative Work (CSCW)* 8.4, pp. 373–401.

Bertalanffy, Ludwig von (1968). *General Systems Theory: Foundations, Development, Applications*. New York: George Braziller.

Bertrand, Astrid, Rafik Belloum, James R Eagan, and Winston Maxwell (2022). "How cognitive biases affect XAI-assisted decision-making: A systematic review." In: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 78–91.

Bertrand, Astrid, Tiphaine Viard, Rafik Belloum, James R Eagan, and Winston Maxwell (2023). "On selective, mutable and dialogic xai: A review of what users say about different types of interactive explanations." In: *Proceedings of the 2023 CHI conference on human factors in computing systems*, pp. 1–21.

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*, pp. 401–413.

Bienefeld, Nadine, Emanuela Keller, and Gudela Grote (2025). "AI interventions to alleviate healthcare shortages and enhance work conditions in critical care: qualitative analysis." In: *Journal of Medical Internet Research* 27, e50852.

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*, pp. 173–184.

Bjork, Elizabeth L, Robert A Bjork, et al. (2011). "Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning." In: *Psychology and the real world: Essays illustrating fundamental contributions to society* 2.59-68, pp. 56–64.

Bjork, Robert A (1994). "Memory and metamemory considerations in the training of human beings." In: *Metacognition: Knowing about knowing* 185.7.2, pp. 185–205.

Botvinick, Matthew M, Todd S Braver, CS Carter, DM Barch, and JD Cohen (2001). "Evaluating the demand for control: Anterior cingulate cortex and crosstalk monitoring." In: *Psychological Review* 108, pp. 624–652.

Brachman, Michelle, Zahra Ashktorab, Michael Desmond, Evelyn Duesterwald, Casey Dugan, Narendra Nath Joshi, Qian Pan, and Aabhas Sharma (2022). "Reliance and Automation for Human-AI Collaborative Data Labeling Conflict Resolution." In: *Proceedings of the ACM on Human-Computer Interaction* 6.CSCW2, pp. 1–27.

Brdnik, Saša, Ivona Colakovic, and Sašo Karakatič (2025). "Non-experts' Trust in XAI is Unreasonably High." In: *World Conference on Explainable Artificial Intelligence*. Springer, pp. 184–197.

Breckner, Karin, Thomas Neumayr, Martina Mara, Marc Streit, and Mirjam Augstein (2025). "The Changing Nature of Human-AI Relations: A Scoping Review on Terminology and Evolvement in the Scientific Literature." In: *International Journal of Human–Computer Interaction*, pp. 1–58.

Brouwer, Thomas, Roberta Ferrario, and Daniele Porello (2021). "Hybrid collective intentionality." In: *Synthese* 199.1, pp. 3367–3403.

Browning, John G (2024). "No "Robot Lawyers" Just Yet: The Role of Continuing Legal Education in Fulfilling the Duty of Technological Competence." In: *Journal of Legal Education* 72.3, p. 11.

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." In: *Proceedings of the ACM on Human-computer Interaction* 5.CSCW1, pp. 1–21.

Bunch, Jacinda, Daryl Jones, and Alex Psirides (2023). "Are we deskilling or reskilling our hospital ward clinicians?" In: *Internal Medicine Journal*.

Burkart, Nadia and Marco F Huber (2021). "A survey on the explainability of supervised machine learning." In: *Journal of Artificial Intelligence Research* 70, pp. 245–317.

Bussone, Adrian, Simone Stumpf, and Dympna O'Sullivan (2015). "The role of explanations on trust and reliance in clinical decision support systems." In: *2015 international conference on healthcare informatics*. IEEE, pp. 160–169.

Byrne, David (2017). "Eliminating the human." In: *MIT Technol. Rev* 120.5, pp. 8–10.

Cabitza, F., R. Rasoini, and G. F. Gensini (2017). "Unintended consequences of machine learning in medicine." In: *JAMA* 318.6, pp. 517–518.

Cabitza, Federico (Aug. 2021b). "Cobra AI: Exploring Some Unintended Consequences of Our Most Powerful Technology." In: *Machines We Trust: Perspectives on Dependable AI*. The MIT Press. ISBN: 9780262366212. DOI: 10.7551/mitpress/12186.003.0011. eprint: https://direct.mit.edu/book/chapter-pdf/2249912/c004800_9780262366212.pdf. URL: https://doi.org/10.7551/mitpress/12186.003.0011.

– (2021a). "Cobra AI: Exploring Some Unintended Consequences." In: *Machines We Trust: Perspectives on Dependable AI* 87.

Cabitza, Federico, Andrea Campagner, Domenico Albano, Alberto Aliprandi, Alberto Bruno, Vito Chianca, Angelo Corazza, Francesco Di Pietto, Angelo Gambino, Salvatore Gitto, et al. (2020). "The elephant in the machine: Proposing a new metric of data reliability and its application to a medical case to assess classification reliability." In: *Applied Sciences* 10.11, p. 4014.

Cabitza, Federico, Andrea Campagner, Riccardo Angius, Chiara Natali, and Carlo Reverberi (2023a). "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*, pp. 1–20.

Cabitza, Federico, Andrea Campagner, Davide Ciucci, and Andrea Seveso (2019). "Programmed inefficiencies in DSS-supported human decision making." In: *International Conference on Modeling Decisions for Artificial Intelligence*. Springer, pp. 201–212.

Cabitza, Federico, Andrea Campagner, Lorenzo Famiglini, Chiara Natali, Valerio Caccavella, and Enrico Gallazzi (2023b). "Let me think! investigating the effect of explanations feeding doubts about the AI advice." In: *International cross-domain conference for machine learning and knowledge extraction*. Springer, pp. 155–169.

Cabitza, Federico, Andrea Campagner, Caterina Fregosi, Matteo Cameli, Enrico Gallazzi, Luca Maria Sconfienza, and Gian Eugenio Tontini

(2025a). "Five Degrees of Separation: Investigating the Unexpected Potential of Displaced Human-AI Collaboration Protocols for Apter AI Support." In: *Proceedings of the ACM on Human-Computer Interaction* 9.7, pp. 1–28.

Cabitza, Federico, Andrea Campagner, Gianclaudio Malgieri, Chiara Natali, David Schneeberger, Karl Stoeger, and Andreas Holzinger (2023c). "Quod erat demonstrandum?-Towards a typology of the concept of explanation for the design of explainable AI." In: *Expert systems with Applications* 213, p. 118888.

Cabitza, Federico, Andrea Campagner, Chiara Natali, Enea Parimbelli, Luca Ronzio, and Matteo Cameli (2023d). "Painting the black box white: experimental findings from applying XAI to an ECG reading setting." In: *Machine Learning and Knowledge Extraction* 5.1, pp. 269–286.

Cabitza, Federico, Andrea Campagner, Luca Ronzio, Matteo Cameli, Giulia Elena Mandoli, Maria Concetta Pastore, Luca Maria Sconfienza, Duarte Folgado, Marília Barandas, and Hugo Gamboa (2023e). "Rams, hounds and white boxes: Investigating human–AI collaboration protocols in medical diagnosis." In: *Artificial Intelligence in Medicine* 138, p. 102506.

Cabitza, Federico, Andrea Campagner, and Luca Maria Sconfienza (2021). "Studying human-AI collaboration protocols: the case of the Kasparov's law in radiological double reading." In: *Health information science and systems* 9, pp. 1–20.

Cabitza, Federico, Andrea Campagner, and Carla Simone (2021). "The need to move away from agential-AI: Empirical investigations, useful concepts and open issues." In: *International Journal of Human-Computer Studies* 155, p. 102696.

Cabitza, Federico, Lorenzo Famiglini, Caterina Fregosi, Samuele Pe, Enea Parimbelli, Giovanni Andrea La Maida, and Enrico Gallazzi (2025b). "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*, pp. 745–754.

Cabitza, Federico, Caterina Fregosi, Andrea Campagner, and Chiara Natali (2024a). "Explanations considered harmful: the impact of misleading explanations on accuracy in hybrid human-ai decision making." In: *World conference on explainable artificial intelligence*. Springer, pp. 255–269.

Cabitza, Federico, Angela Locoro, and Aurelio Ravarini (2020). "Trading off between control and autonomy: a narrative review around de-design." In: *Behaviour & Information Technology* 39.1, pp. 5–26.

Cabitza, Federico and Chiara Natali (2022). "Open, multiple, adjunct. decision support at the time of relational AI." In: *Frontiers in Artificial Intelligence and Applications* 354, pp. 243–245.

Cabitza, Federico, Chiara Natali, Lorenzo Famiglini, Andrea Campagner, Valerio Caccavella, and Enrico Gallazzi (2024b). "Never tell me the odds: Investigating pro-hoc explanations in medical decision making." In: *Artificial intelligence in medicine* 150, p. 102819.

Cabitza, Federico, Chiara Natali, Francesco Varanini, and David Gunkel (2025c). "Beyond cyborgs: the cybork idea for the de-individuation of (artificial) intelligence and an emergence-oriented design." In: *AI & SOCIETY* 40.5, pp. 3333–3348.

Cacioppo, John T and Richard E Petty (1982). "The need for cognition." In: *Journal of personality and social psychology* 42.1, p. 116.

Cai, Alice, Ian Arawjo, and Elena L Glassman (2024). "Antagonistic ai." In: *arXiv preprint arXiv:2402.07350*.

Cambo, Scott Allen and Darren Gergle (2022). "Model positionality and computational reflexivity: Promoting reflexivity in data science." In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–19.

Cambria, Erik, Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, and Navid Nobani (2023). "A survey on XAI and natural language explanations." In: *Information Processing & Management* 60.1, p. 103111.

Campagner, Andrea, Federico Cabitza, and Davide Ciucci (2019). "Three–way classification: Ambiguity and abstention in machine learning." In: *International Joint Conference on Rough Sets*. Springer, pp. 280–294.

Campagner, Andrea, Davide Ciucci, Carl-Magnus Svensson, Marc Thilo Figge, and Federico Cabitza (2021). "Ground truthing from multi-rater labeling with three-way decision and possibility theory." In: *Information Sciences* 545, pp. 771–790.

Campagner, Andrea, Frida Milella, Davide Ciucci, and Federico Cabitza (2024). "Three-way decision in machine learning tasks: a systematic review." In: *Artificial Intelligence Review* 57.9, p. 228.

Campbell, Cara G, Daniel SW Ting, Pearse A Keane, and Paul J Foster (2020). "The potential application of artificial intelligence for diagnosis and management of glaucoma in adults." In: *British Medical Bulletin* 134.1, pp. 21–33.

Carayannis, Elias and John Coleman (2005). "Creative system design methodologies: the case of complex technical systems." In: *Technovation* 25.8, pp. 831–840.

Carr, Nicholas (2015). *The glass cage: Where automation is taking us*. Random House.

Carrel, Alyson (2018). "Legal intelligence through artificial intelligence requires emotional intelligence: a new competency model for the 21st century legal professional." In: *Ga. St. UL Rev.* 35, p. 1153.

Cavalcante Siebert, Luciano, Maria Luce Lupetti, Evgeni Aizenberg, Niek Beckers, Arkady Zgonnikov, Herman Veluwenkamp, David Abbink, Elisa Giaccardi, Geert-Jan Houben, Catholijn M Jonker, et

al. (2023). "Meaningful human control: actionable properties for AI system development." In: *AI and Ethics* 3.1, pp. 241–255.

Chae, Bongsug, James F Courtney, and John D Haynes (2005). "Information Technology and Hegelian Inquiring Organizations." In: *Inquiring Organizations: Moving from Knowledge Management to Wisdom*. IGI Global, pp. 22–45.

Chalmers, Matthew (2003). "Seamful design and ubicomp infrastructure." In: *Proceedings of Ubicomp 2003 workshop at the crossroads: The interaction of HCI and systems issues in Ubicomp*, pp. 577–584.

Checkland, Peter (1999). "Systems thinking." In: *Rethinking management information systems*, pp. 45–56.

Chen, Jessie YC and Michael J Barnes (2014). "Human–agent teaming for multirobot control: A review of human factors issues." In: *IEEE Transactions on Human-Machine Systems* 44.1, pp. 13–29.

Chen, Quan Ze, Daniel S Weld, and Amy X Zhang (2021). "Goldilocks: Consistent crowdsourced scalar annotations with relative uncertainty." In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW2, pp. 1–25.

Chen, Yaru, Charitini Stavropoulou, Radhika Narasinkan, Adrian Baker, and Harry Scarbrough (2021). "Professionals' responses to the introduction of AI innovations in radiology and their implications for future adoption: a qualitative study." In: *BMC health services research* 21, pp. 1–9.

Chiriatti, Massimo, Marianna Ganapini, Enrico Panai, Mario Ubiali, and Giuseppe Riva (2024). "The case for human–AI interaction as system 0 thinking." In: *Nature Human Behaviour* 8.10, pp. 1829–1830.

Choudhury, Avishek and Zaira Chaudhry (2024). "Large language models and user trust: consequence of self-referential learning loop and the deskilling of health care professionals." In: *Journal of Medical Internet Research* 26, e56764.

Christin, Angèle (2020). "The ethnographer and the algorithm: Beyond the black box." In: *Theory and Society* 49.5, pp. 897–918.

Chromik, Michael, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz (2021). "I think i get your point, AI! the illusion of explanatory depth in explainable AI." In: *Proceedings of the 26th International Conference on Intelligent User Interfaces*, pp. 307–317.

Clark, Andy and David Chalmers (1998). "The extended mind." In: *analysis* 58.1, pp. 7–19.

Collins, Harry M (2005). "What is tacit knowledge?" In: *The practice turn in contemporary theory*. Routledge, pp. 115–128.

Cox, Anna L, Sandy JJ Gould, Marta E Cecchinato, Ioanna Iacovides, and Ian Renfree (2016). "Design frictions for mindful interactions: The case for microboundaries." In: *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems*, pp. 1389–1397.

Craik, Fergus IM and Endel Tulving (1975). "Depth of processing and the retention of words in episodic memory." In: *Journal of experimental Psychology: general* 104.3, p. 268.

Croskerry, Pat (2013). "From mindless to mindful practice—cognitive bias and clinical decision making." In: *New England Journal of Medicine* 368.26, pp. 2445–2448.

Da Silva, Michael, Tanya Horsley, Devin Singh, Emily Da Silva, Valentina Ly, Bryan Thomas, Ryan C Daniel, Karni A Chagal-Feferkorn, Samantha Iantomasi, Kelli White, et al. (2022). "Legal concerns in health-related artificial intelligence: a scoping review protocol." In: *Systematic Reviews* 11.1, pp. 1–8.

Dalim, Siti (Oct. 2022). "Promoting Students' Critical Thinking Through Socratic Method: The Views and Challenges." In: *Asian Journal of University Education* 18.4, pp. 1034–1047. DOI: 10.24191/ajue.v18i4.20012.

Danry, Valdemar, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes (2023). "Don't just tell me, ask me: Ai systems that intelligently frame explanations as questions improve human logical discernment accuracy over causal ai explanations." In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–13.

Davani, Aida Mostafazadeh, Mark Díaz, and Vinodkumar Prabhakaran (2022). "Dealing with disagreements: Looking beyond the majority vote in subjective annotations." In: *Transactions of the Association for Computational Linguistics* 10, pp. 92–110.

Davis, Fred D et al. (1989). "Technology acceptance model: TAM." In: *Al-Suqri, MN, Al-Aufi, AS: Information Seeking Behavior and Technology Adoption* 205.219, p. 5.

Delibasic, Boris, Milan Vukicevic, and MILO Jovanovic (2013). "White-box decision tree algorithms: A pilot study on perceived usefulness, perceived ease of use, and perceived understanding." In: *International Journal of Engineering Education* 29.3, pp. 674–687.

Dell'Anna, Davide, Pradeep K Murukannaiah, Bernd Dudzik, Davide Grossi, Catholijn M Jonker, Catharine Oertel, and Pınar Yolum (2024). "Toward a quality model for hybrid intelligence teams." In: *23rd International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2024*. ACM Press Digital Library, pp. 434–443.

Dellermann, Dominik, Philipp Ebel, Matthias Söllner, and Jan Marco Leimeister (2019). "Hybrid intelligence." In: *Business & Information Systems Engineering* 61.5, pp. 637–643.

Dewey, John (1933). *How We Think: A Restatement of the Relation of Reflective Thinking to the Educative Process*. 2nd ed. Lexington, Massachusetts: D.C. Heath and Company.

Dias Duran, Leslye Denisse (2021). "Deskilling of medical professionals: an unintended consequence of AI implementation?" In: *Giornale di filosofia* 2.2.

Diaz Alfaro, G., S. M. Fiore, and K. Oden (2024). "Externalized and extended cognition: Cognitive offloading for human-machine teaming." In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. SAGE Publications Sage CA: Los Angeles, CA, p. 10711813241275506.

Dickersin, Kay, SS Chan, TC Chalmersx, HS Sacks, and H Smith Jr (1987). "Publication bias and clinical trials." In: *Controlled clinical trials* 8.4, pp. 343–353.

Dietvorst, Berkeley J, Joseph Simmons, and Cade Massey (2014). "Understanding algorithm aversion: forecasters erroneously avoid algorithms after seeing them err." In: *Academy of management proceedings*. Vol. 2014. 1. Academy of management Briarcliff Manor, NY 10510, p. 12227.

Dix, Alan (2007). "Designing for appropriation." In: *Proceedings of HCI 2007 The 21st British HCI Group Annual Conference University of Lancaster, UK*. BCS Learning & Development.

Dourish, Paul (2001). *Where the action is: the foundations of embodied interaction*. MIT press.

Drabiak, K., S. Kyzer, V. Nemov, and I. El Naqa (2023). "AI and machine learning ethics, law, diversity, and global impact." In: *The British Journal of Radiology* 96, p. 20220934.

Dragicevic, Pierre, Yvonne Jansen, Abhraneel Sarma, Matthew Kay, and Fanny Chevalier (2019). "Increasing the transparency of research papers with explorable multiverse analyses." In: *proceedings of the 2019 chi conference on human factors in computing systems*, pp. 1–15.

Drosos, Ian, Advait Sarkar, Neil Toronto, et al. (2025). ""' It makes you think": Provocations Help Restore Critical Thinking to AI-Assisted Knowledge Work." In: *arXiv preprint arXiv:2501.17247*.

Dunne, Anthony (2008). *Hertzian tales: Electronic products, aesthetic experience, and critical design*. MIT press.

Duran, Huong-Tram, Meredith Kingeter, Carrie Reale, Matthew B Weinger, and Megan E Salwei (2023). "Decision-making in anesthesiology: will artificial intelligence make intraoperative care safer?" In: *Current Opinion in Anesthesiology* 36.6, pp. 691–697.

Eckhardt, Sven, Niklas Kühl, Mateusz Dolata, and Gerhard Schwabe (2024). "A Survey of AI Reliance." In: *arXiv preprint arXiv:2408.03948*.

Ehsan, Upol and Mark O Riedl (2020). "Human-centered explainable ai: Towards a reflective sociotechnical approach." In: *International Conference on Human-Computer Interaction*. Springer, pp. 449–466.

Eiband, Malin, Daniel Buschek, Alexander Kremer, and Heinrich Hussmann (2019). "The impact of placebic explanations on trust in intelligent systems." In: *Extended abstracts of the 2019 CHI conference on human factors in computing systems*, pp. 1–6.

Elish, M. C. (2019). "Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction." In: *Engaging Science,Technology, and So-*

ciety. DOI: 10.17351/ests2019.260. URL: https://doi.org/10.17351/ests2019.260.

Engelbart, Douglas C (1963). "A conceptual framework for the augmentation of man's intellect." In: *Vistas in information handling* 1, pp. 1–29.

– (2023). "Augmenting human intellect: A conceptual framework." In: *Augmented education in the global age*. Routledge, pp. 13–29.

Ericson, Jonathan (2022). "Reimagining the Role of Friction in Experience Design." In: *Journal of User Experience* 17.4.

European Parliament and the Council of the European Union (June 2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)*. Official Journal of the European Union. Article 14. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689.

Ferrario, Andrea, Alessandro Facchini, and Alberto Termine (2024). "Experts or authorities? The strange case of the presumed epistemic superiority of artificial intelligence systems." In: *Minds and Machines* 34.3, p. 30.

Fischer, Gerhard (2003). "Meta-design: Beyond user-centered and participatory design." In: *Proceedings of HCI international*. Vol. 4, pp. 88–92.

Fischer, Gerhard and Thomas Herrmann (2011). "Socio-technical systems: a meta-design perspective." In: *International Journal of Sociotechnology and Knowledge Development (IJSKD)* 3.1, pp. 1–33.

Fischer, S. W., H. Schraffenberger, S. Thill, and P. Haselager (2025). "A Taxonomy of Questions for Critical Reflection in Machine-Assisted Decision-Making." In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. Vol. 8. 1, pp. 940–954. DOI: 10.1609/aies.v8i1.36602.

Fischer, Simon WS (2025). "A Reflection Machine to Support Critical Reflection During Decision-Making." In: *Companion Proceedings of the 30th International Conference on Intelligent User Interfaces*, pp. 199–201.

Fiske, Susan T Tufts and Shelley E Taylor (2020). "Social cognition: From brains to culture." In.

Floridi, Luciano (2021). "Establishing the rules for building trustworthy AI." In: *Ethics, governance, and policies in artificial intelligence*. Springer, pp. 41–45.

Fregosi, Caterina and Federico Cabitza (2024). "A Frictional Design Approach: Towards Judicial AI and its Possible Applications." In: *Proceedings of the Workshops at the Third International Conference on Hybrid Human-Artificial Intelligence (HHAI-WS 2024)*.

Frischmann, Brett and Susan Benesch (2023). "Friction-in-design regulation as 21st century time, place, and manner restriction." In: *Yale JL & Tech.* 25, p. 376.

Frischmann, Brett and Evan Selinger (2019). *Re-engineering humanity*. Cambridge University Press.

Gajos, Krzysztof Z and Lena Mamykina (2022). "Do people engage cognitively with AI? Impact of AI assistance on incidental learning." In: *Proceedings of the 27th International Conference on Intelligent User Interfaces*, pp. 794–806.

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." In: *NPJ digital medicine* 4.1, p. 31.

Gaver, William W, Jacob Beaver, and Steve Benford (2003). "Ambiguity as a resource for design." In: *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 233–240.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford (2021). "Datasheets for datasets." In: *Communications of the ACM* 64.12, pp. 86–92.

Gerke, S., T. Minssen, and G. Cohen (2020). "Ethical and legal challenges of artificial intelligence-driven healthcare." In: pp. 295–336.

Ghafur, Saira, Parminder K Judge, Richard Kitchen, and Samuel Blows (2017). *The MRCP PACES Handbook*. CRC Press.

Ghassemi, Marzyeh, Luke Oakden-Rayner, and Andrew L Beam (2021). "The false hope of current approaches to explainable artificial intelligence in health care." In: *The Lancet Digital Health* 3.11, e745–e750.

Glickman, Moshe and Tali Sharot (2025). "How human–AI feedback loops alter human perceptual, emotional and social judgements." In: *Nature Human Behaviour* 9.2, pp. 345–359.

Goddard, Kate, Abdul Roudsari, and Jeremy C Wyatt (2011). "Automation bias–a hidden issue for clinical decision support system use." In: *International perspectives in health informatics*, pp. 17–22.

Golfetti, Alessia, Linda Napoletano, and Katarzyna Cichomska (2021). "A Framework to Understand Current and Future Competences and Occupations in the Aviation Sector." In: *Transformation of Transportation*, pp. 213–226.

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." In: *Frontiers in Computer Science* 6, p. 1521066.

Gordon, Mitchell L, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein (2022). "Jury learning: Integrating dissenting voices into machine learning models." In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–19.

Gosline, Renee (2022). "Why ai customer journeys need more friction." In: *Harvard Business Review*.

Graziani, Mara, Lidia Dutkiewicz, Davide Calvaresi, José Pereira Amorim, Katerina Yordanova, Mor Vered, Rahul Nair, Pedro Henriques Abreu, Tobias Blanke, Valeria Pulignano, et al. (2023). "A global taxonomy of interpretable AI: unifying the terminology for the technical and social sciences." In: *Artificial intelligence review* 56.4, pp. 3473–3504.

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

Green, Brian Patrick (2019). "Artificial Intelligence, Decision-Making, and Moral Deskilling." In: *Markkula Center for Applied Ethics*.

Grissinger, Matthew (2019). "Understanding human over-reliance on technology." In: *Pharmacy and Therapeutics* 44.6, p. 320.

Gunning, David and David Aha (2019). "DARPA's explainable artificial intelligence (XAI) program." In: *AI magazine* 40.2, pp. 44–58.

Guo, Jonathan and Bin Li (2018). "The application of medical artificial intelligence technology in rural areas of developing countries." In: *Health equity* 2.1, pp. 174–181.

Guo, Ziyang, Yifan Wu, Jason D Hartline, and Jessica Hullman (2024). "A Decision Theoretic Framework for Measuring AI Reliance." In: *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 221–236.

Hallnäs, Lars and Johan Redström (2001). "Slow technology–designing for reflection." In: *Personal and ubiquitous computing* 5.3, pp. 201–212.

Hallowell, Nina, Shirlene Badger, Francis McKay, Angeliki Kerasidou, and Christoffer Nellåker (2023). "Democratising or disrupting diagnosis? Ethical issues raised by the use of AI tools for rare disease diagnosis." In: *SSM - Qualitative Research in Health* 3, p. 100240. ISSN: 2667-3215. DOI: https://doi.org/10.1016/j.ssmqr.2023.100240. URL: https://www.sciencedirect.com/science/article/pii/S2667321523000240.

Hao, Karen (2022). "When algorithms mess up, the nearest human gets the blame." In: *Ethics of Data and Analytics*. Auerbach Publications, pp. 418–419.

Harrison, Steve and Paul Dourish (1996). "Re-place-ing space: the roles of place and space in collaborative systems." In: *Proceedings of the 1996 ACM conference on Computer supported cooperative work*, pp. 67–76.

Hart, Sandra G and Lowell E Staveland (1988). "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research." In: *Advances in psychology*. Vol. 52. Elsevier, pp. 139–183.

Hartswood, Mark, Rob Procter, Mark Rouncefield, and Roger Slack (2003). "Making a case in medical work: implications for the elec-

tronic medical record." In: *Computer Supported Cooperative Work (CSCW)* 12.3, pp. 241–266.

Harvey, Hugh, Federico Cabitza, et al. (2018). "Algorithms are the new drugs? Reflections for a culture of impact assessment and vigilance." In: *IADIS International Conference ICT, Society and Human Beings 2018*.

Haselager, Pim, Hanna Schraffenberger, Serge Thill, Simon Fischer, Pablo Lanillos, Sebastiaan Van De Groes, and Miranda Van Hooff (2024). "Reflection machines: Supporting effective human oversight over medical decision support systems." In: *Cambridge Quarterly of Healthcare Ethics* 33.3, pp. 380–389.

He, Gaole, Lucie Kuiper, and Ujwal Gadiraju (2023). "Knowing about knowing: An illusion of human competence can hinder appropriate reliance on AI systems." In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–18.

Helander, Martin G (2014). *Handbook of human-computer interaction*. Elsevier.

Hemmer, Patrick, Max Schemmer, Michael Vössing, and Niklas Kühl (2021). "Human-AI Complementarity in Hybrid Intelligence Systems: A Structured Literature Review." In: *PACIS*, p. 78.

Hemmer, Patrick, Lukas Thede, Michael Vössing, Johannes Jakubik, and Niklas Kühl (2023). "Learning to defer with limited expert predictions." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 5, pp. 6002–6011.

Hernández-Orallo, José and Karina Vold (2019). "AI extenders: the ethical and societal implications of humans cognitively extended by AI." In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 507–513.

Hildebrandt, Mireille (2018). "Algorithmic regulation and the rule of law." In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376.2128, p. 20170355.

– (2019). "Privacy as protection of the incomputable self: From agnostic to agonistic machine learning." In: *Theoretical Inquiries in Law* 20.1, pp. 83–121.

Hinduja, Mohneesh, Philipp Ebel, Edona Elshan, Dominik Siemon, Triparna de Vreede, and Gert-Jan de Vreede (2025). "Confirmation Bias in Human-AI Interactions." In: *ICIS 2025 Proceedings* 16. URL: https://aisel.aisnet.org/icis2025/gen_ai/gen_ai/16.

Ho, Yueh-Ren, Bao-Yu Chen, and Chien-Ming Li (Mar. 2023). "Thinking More Wisely: Using the Socratic Method to Develop Critical Thinking Skills amongst Healthcare Students." In: *BMC Medical Education* 23.1, p. 173. DOI: 10.1186/s12909-023-04134-2.

Hoff, Timothy (2011). "Deskilling and adaptation among primary care physicians using two work innovations." In: *Health Care Management Review* 36.4, pp. 338–348.

Hollan, James, Edwin Hutchins, and David Kirsh (2000). "Distributed cognition: toward a new foundation for human-computer interac-

tion research." In: *ACM Transactions on Computer-Human Interaction (TOCHI)* 7.2, pp. 174–196.

Holzinger, Andreas (2016). "Interactive machine learning for health informatics: when do we need the human-in-the-loop?" In: *Brain Informatics* 3.2, pp. 119–131.

Horvitz, Eric (1999). "Principles of mixed-initiative user interfaces." In: *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pp. 159–166.

Huang, Hsieh-Hong, Jack Shih-Chieh Hsu, and Cheng-Yuan Ku (2012). "Understanding the role of computer-mediated counter-argument in countering confirmation bias." In: *Decision Support Systems* 53.3, pp. 438–447.

Hutchins, Edwin (2000). "Distributed cognition." In: *International encyclopedia of the social and behavioral sciences* 138.1, pp. 1–10.

ISO/IEC (2021). *Information Technology — Artificial Intelligence (AI) — Bias in AI Systems and AI Aided Decision-Making*. Tech. rep. ISO/IEC TR 24027:2021. Geneva, Switzerland: International Organization for Standardization.

Inman, Sarah and David Ribes (2019). "" Beautiful Seams" Strategic Revelations and Concealments." In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–14.

Inuwa-Dutse, Isa, Alice Toniolo, Adrian Weller, and Umang Bhatt (2023). "Algorithmic loafing and mitigation strategies in Human-AI teams." In: *Computers in Human Behavior: Artificial Humans* 1.2, p. 100024.

Iqbal, J., K. Jahangir, Y. Mashkoor, N. Sultana, D. Mehmood, M. Ashraf, and M. H. Hafeez (2022). "The future of artificial intelligence in neurosurgery: a narrative review." In: *Surgical Neurology International* 13.

Jarupathirun, Suprasith et al. (2007). "Dialectic decision support systems: System design and empirical evaluation." In: *Decision Support Systems* 43.4, pp. 1553–1570.

Jentzsch, Sophie F, Sviatlana Höhn, and Nico Hochgeschwender (2019). "Conversational interfaces for explainable AI: a human-centred approach." In: *International workshop on explainable, transparent autonomous agents and multi-agent systems*. Springer, pp. 77–92.

Johnson, Matthew, Jeffrey M Bradshaw, Paul J Feltovich, Catholijn M Jonker, M Birna Van Riemsdijk, and Maarten Sierhuis (2014). "Coactive design: Designing support for interdependence in joint activity." In: *Journal of Human-Robot Interaction* 3.1, pp. 43–69.

Jong, Sander de, Ville Paananen, Benjamin Tag, and Niels van Berkel (2025). "Cognitive Forcing for Better Decision-Making: Reducing Overreliance on AI Systems Through Partial Explanations." In: *Proceedings of the ACM on Human-Computer Interaction* 9.2, pp. 1–30.

Jussupow, Ekaterina, Kai Spohrer, Armin Heinzl, and Joshua Gawlitza (2021). "Augmenting medical diagnosis decisions? An investigation

into physicians' decision-making process with artificial intelligence." In: *Information Systems Research* 32.3, pp. 713–735.

Kahneman, Daniel (2011). "Thinking, fast and slow." In: *Farrar, Straus and Giroux*.

– (2013). *Thinking, Fast and Slow*. New York, NY: Farrar, Straus and Giroux. ISBN: 9780374533557.

Kahneman, Daniel, Olivier Sibony, and Cass R Sunstein (2021). *Noise: A flaw in human judgment*. Hachette UK.

Kapoor, R., S. P. Walters, and L. A. Al-Aswad (2019). "The current state of artificial intelligence in ophthalmology." In: *Survey of Ophthalmology* 64.2, pp. 233–240.

Kashou, Anthony H, Peter A Noseworthy, Nandan S Anavekar, Ian Rowlandson, and Adam M May (2024). "Bridging ECG learning with emerging technologies: Advancing clinical excellence." In: *Journal of Electrocardiology* 86, p. 153765.

Kasparov, Garry and Mig Greengard (2017). *Deep Thinking: Where Machine Intelligence Ends and Human Creativity Begins*. PublicAffairs, p. 304. ISBN: 9781610397872. URL: https://books.google.it/books?id=0Ub6DAAAQBAJ.

Kaur, Harmanpreet, Matthew R Conrad, Davis Rule, Cliff Lampe, and Eric Gilbert (2024). "Interpretability gone bad: The role of bounded rationality in how practitioners understand machine learning." In: *Proceedings of the ACM on Human-Computer Interaction* 8.CSCW1, pp. 1–34.

Keane, Mark T and Barry Smyth (2020). "Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable AI (XAI)." In: *International Conference on Case-Based Reasoning*. Springer, pp. 163–178.

Keane, Pearse A and Eric J Topol (2018). "With an eye to AI and autonomous diagnosis." In: *NPJ digital medicine* 1.1, p. 40.

Kliegr, Tomáš, Štěpán Bahník, and Johannes Fürnkranz (2021). "A review of possible effects of cognitive biases on interpretation of rule-based machine learning models." In: *Artificial Intelligence* 295, p. 103458.

Koplin, Julian J, Molly Johnston, Amy NS Webb, Andrea Whittaker, and Catherine Mills (2025). "Ethics of artificial intelligence in embryo assessment: mapping the terrain." In: *Human Reproduction* 40.2, pp. 179–185.

Krug, Steve (2000). *Don't make me think!: a common sense approach to Web usability*. Pearson Education India.

Kundu, Shinjini (2021). "How will artificial intelligence change medical training?" In: *Communications Medicine* 1.1, p. 8.

Kwong, Jethro CC, David-Dan Nguyen, Adree Khondker, Jin Kyu Kim, Alistair EW Johnson, Melissa M McCradden, Girish S Kulkarni, Armando Lorenzo, Lauren Erdman, and Mandy Rickard (2024). "When the model trains you: induced belief revision and its impli-

cations on artificial intelligence research and patient care—a case study on predicting obstructive hydronephrosis in children." In: *NEJM AI* 1.2, AIcs2300004.

Lai, Vivian, Samuel Carton, Rajat Bhatnagar, Q Vera Liao, Yunfeng Zhang, and Chenhao Tan (2022). "Human-ai collaboration via conditional delegation: A case study of content moderation." In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–18.

Lai, Vivian and Chenhao Tan (2019). "On human predictions with explanations and predictions of machine learning models: A case study on deception detection." In: *Proceedings of the conference on fairness, accountability, and transparency*, pp. 29–38.

Langer, Markus, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum (2021). "What do we want from Explainable Artificial Intelligence (XAI)?–A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research." In: *Artificial Intelligence* 296, p. 103473.

Langlotz, Curtis P (2019). *Will artificial intelligence replace radiologists?*

Lara, Francisco (2021). "Why a virtual assistant for moral enhancement when we could have a socrates?" In: *Science and engineering ethics* 27.4, p. 42.

Lebiere, Christian, Leslie M Blaha, Corey K Fallon, and Brett Jefferson (2021). "Adaptive cognitive mechanisms to maintain calibrated trust and reliance in automation." In: *Frontiers in Robotics and AI* 8, p. 652776.

Lebovitz, Sarah, Natalia Levina, and Hila Lifshitz-Assaf (2021). "Is AI ground truth really true? The dangers of training and evaluating AI tools based on experts' know-what." In: *MIS quarterly* 45.3, pp. 1501–1526.

Lee, John D and Katrina A See (2004). "Trust in automation: Designing for appropriate reliance." In: *Human factors* 46.1, pp. 50–80.

Lennartz, Simon, Thomas Dratsch, David Zopfs, Thorsten Persigehl, David Maintz, Nils Große Hokamp, and Daniel Pinto dos Santos (2021). "Use and control of artificial intelligence in patients across the medical workflow: single-center questionnaire study of patient perspectives." In: *Journal of Medical Internet Research* 23.2, e24221.

Levy, Jaime, Alan Jotkowitz, and Itay Chowers (2019). "Deskilling in ophthalmology is the inevitable controllable?" In: *Eye* 33.3, pp. 347–348.

Li, Matthew D and Brent P Little (2023). "Appropriate reliance on artificial intelligence in radiology education." In: *Journal of the American College of Radiology* 20.11, pp. 1126–1130.

Licklider, Joseph CR (2008). "Man-computer symbiosis." In: *IRE transactions on human factors in electronics* 1, pp. 4–11.

Lighthall, Geoffrey K and Cristina Vazquez-Guillamet (2016). "Understanding decision making in critical care." In: *Clinical medicine & research* 13.3-4, pp. 156–168.

Liu, Jiayu, Zhenya Huang, Tong Xiao, Jing Sha, Jinze Wu, Qi Liu, Shijin Wang, and Enhong Chen (2024). "SocraticLM: Exploring socratic personalized teaching with large language models." In: *Advances in Neural Information Processing Systems* 37, pp. 85693–85721.

Logg, Jennifer M, Julia A Minson, and Don A Moore (2019). "Algorithm appreciation: People prefer algorithmic to human judgment." In: *Organizational Behavior and Human Decision Processes* 151, pp. 90–103.

López-Pernas, Sonsoles, Eduardo Oliveira, Yige Song, and Mohammed Saqr (2025). "AI, explainable AI and evaluative AI: Informed data-driven decision-making in education." In: *Advanced learning analytics methods: AI, precision and complexity*. Springer, pp. 17–39.

Loyola-Gonzalez, Octavio (2019). "Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view." In: *IEEE access* 7, pp. 154096–154113.

Lu, J. (2016). "Will medical technology deskill doctors?" In: *International Education Studies*.

Luff, Paul, Christian Heath, and David Greatbatch (1992). "Tasks-in-interaction: paper and screen based documentation in collaborative activity." In: *Proceedings of the 1992 ACM conference on Computer-supported cooperative work*, pp. 163–170.

Lundberg, Scott M and Su-In Lee (2017). "A unified approach to interpreting model predictions." In: *Proceedings of NeurIPS 2017*, pp. 4768–4777.

Lyell, David and Enrico Coiera (2017). "Automation bias and verification complexity: a systematic review." In: *Journal of the American Medical Informatics Association* 24.2, pp. 423–431.

Lyons, Joseph B, Katia Sycara, Michael Lewis, and August Capiola (2021). "Human–autonomy teaming: Definitions, debates, and directions." In: *Frontiers in psychology* 12, p. 589585.

Ma, Shuai, Qiaoyi Chen, Xinru Wang, Chengbo Zheng, Zhenhui Peng, Ming Yin, and Xiaojuan Ma (2025). "Towards human-ai deliberation: Design and evaluation of llm-empowered deliberative ai for ai-assisted decision-making." In: *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–23.

Maassen, Oliver, Sebastian Fritsch, Julia Palm, Saskia Deffge, Julian Kunze, Gernot Marx, Morris Riedel, Andreas Schuppert, and Johannes Bickenbach (2021). "Future medical artificial intelligence application requirements and expectations of physicians in German university hospitals: web-based survey." In: *Journal of medical Internet research* 23.3, e26646.

Madras, David, Toni Pitassi, and Richard Zemel (2018). "Predict responsibly: improving fairness and accuracy by learning to defer." In: *Advances in neural information processing systems* 31.

Malaguti, Pauline, Alexander J Karran, Di Le, Hayley Mortin, Constantinos K Coursaris, Sylvain Sénécal, and Pierre-Majorique Léger (2025). "Investigating Interaction Friction in Generative AI: Improving User Experience and Decision-Making." In: *SIGHCI 2024 Proceedings*.

Malone, T.W. (2018). *Superminds: The Surprising Power of People and Computers Thinking Together*. Little, Brown. ISBN: 9780316349109. URL: https://books.google.it/books?id=Qe0zDwAAQBAJ.

Marano, Giuseppe, Georgios D Kotzalidis, Francesco Maria Lisci, Maria Benedetta Anesini, Sara Rossi, Sara Barbonetti, Andrea Cangini, Alice Ronsisvalle, Laura Artuso, Cecilia Falsini, et al. (2025). "The Neuroscience Behind Writing: Handwriting vs. Typing—Who Wins the Battle?" In: *Life* 15.3, p. 345.

McCradden, Melissa D, James A Anderson, Elizabeth A. Stephenson, Erik Drysdale, Lauren Erdman, Anna Goldenberg, and Randi Zlotnik Shaul (2022). "A research ethics framework for the clinical translation of healthcare machine learning." In: *The American Journal of Bioethics* 22.5, pp. 8–22.

Mejtoft, Thomas, Emma Parsjö, Ole Norberg, and Ulrik Söderström (2023). "Design Friction and Digital Nudging: Impact on the Human Decision-Making Process." In: *Proceedings of the 2023 5th International Conference on Image, Video and Signal Processing*, pp. 183–190.

Merritt, Stephanie M, Deborah Lee, Jennifer L Unnerstall, and Kelli Huber (2015). "Are well-calibrated users effective users? Associations between calibration of trust and performance on an automation-aided task." In: *Human Factors* 57.1, pp. 34–47.

Meza Martínez, Miguel Angel, Mario Nadj, and Alexander Maedche (2019). "Towards an integrative theoretical framework of interactive machine learning systems." In.

Miceli, Milagros and Julian Posada (2022). "The data-production dispositif." In: *Proceedings of the ACM on human-computer interaction* 6.CSCW2, pp. 1–37.

Miceli, Milagros, Tianling Yang, Laurens Naudts, Martin Schuessler, Diana Serbanescu, and Alex Hanna (2021). "Documenting computer vision datasets: An invitation to reflexive data practices." In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 161–172.

Miller, Randolph A and FE Masarie Jr (1990). "The demise of the "Greek Oracle" model for medical diagnostic systems." In: *Methods of information in medicine* 29.01, pp. 1–2.

Miller, Tim (2019). ""But why?" Understanding explainable artificial intelligence." In: *XRDS: Crossroads, The ACM Magazine for Students* 25.3, pp. 20–25.

Miller, Tim (2023). "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*, pp. 333–342.

Mofatteh, M. (2021). "Neurosurgery and artificial intelligence." In: *AIMS Neuroscience* 8.4, p. 477.

Mol, Annemarie (2002). *The body multiple: Ontology in medical practice*. Duke University Press.

Monteith, Scott, Tasha Glenn, John Geddes, Peter C Whybrow, Eric Achtyes, and Michael Bauer (2022). "Expectations for artificial intelligence (AI) in psychiatry." In: *Current Psychiatry Reports* 24.11, pp. 709–721.

Morley, J., C. C. Machado, C. Burr, J. Cowls, I. Joshi, M. Taddeo, and L. Floridi (2020). "The ethics of AI in health care: a mapping review." In: *Social Science & Medicine* 260, p. 113172.

Morrison, Katelyn, Philipp Spitzer, Violet Turri, Michelle Feng, Niklas Kühl, and Adam Perer (2024). "The impact of imperfect XAI on human-AI decision-making." In: *Proceedings of the ACM on human-computer interaction* 8.CSCW1, pp. 1–39.

Mosch, L., D. Fürstenau, J. Brandt, J. Wagnitz, S. A. Klopfenstein, A. S. Poncette, and F. Balzer (2022). "The medical profession transformed by artificial intelligence: Qualitative study." In: *Digital Health* 8, p. 20552076221143903.

Mosier, Kathleen L and Linda J Skitka (1999). "Automation use and automation bias." In: *Proceedings of the human factors and ergonomics society annual meeting*. Vol. 43. 3. SAGE Publications Sage CA: Los Angeles, CA, pp. 344–348.

Muller, Michael, Christine T Wolf, Josh Andres, Michael Desmond, Narendra Nath Joshi, Zahra Ashktorab, Aabhas Sharma, Kristina Brimijoin, Qian Pan, Evelyn Duesterwald, et al. (2021). "Designing ground truth and the social life of labels." In: *Proceedings of the 2021 CHI conference on human factors in computing systems*, pp. 1–16.

Müller, Rainer, Matthias Vette, and Aaron Geenen (2017). "Skill-based dynamic task allocation in human-robot-cooperation with the example of welding application." In: *Procedia Manufacturing* 11, pp. 13–21.

Naiseh, Mohammad, Reem S Al-Mansoori, Dena Al-Thani, Nan Jiang, and Raian Ali (2021). "Nudging through friction: an approach for calibrating trust in explainable AI." In: *2021 8th International Conference on Behavioral and Social Computing (BESC)*. IEEE, pp. 1–5.

Nakagawa, Keisuke, Lama Moukheiber, Leo A Celi, Malhar Patel, Faisal Mahmood, Dibson Gondim, Michael Hogarth, and Richard Levenson (2023). "AI in pathology: what could possibly go wrong?" In: *Seminars in Diagnostic Pathology*. Vol. 40. Elsevier, pp. 100–108.

Nass, Clifford, Jonathan Steuer, Ellen Tauber, and Heidi Reeder (1993). "Anthropomorphism, agency, and ethopoeia: computers as social

actors." In: *INTERACT'93 and CHI'93 conference companion on Human factors in computing systems*, pp. 111–112.

Natale, Simone and Emiliano Treré (2024). "Dreaming of seamless interfaces: media and friction from the feuilleton to personal computing." In: *Information, Communication & Society* 27.10, pp. 1945–1963.

Natali, Chiara (2023). "Per aspera ad astra, or flourishing via friction: Stimulating cognitive activation by design through frictional decision support systems." In: *CEUR workshop proceedings*. Vol. 3481. CEUR-WS, pp. 15–19.

Natali, Chiara and Federico Cabitza (2025). "Make Some Noise for Ground Truthing! Frictional design against epistemic sclerosis in Decision Support Systems." In.

Natali, Chiara, Andrea Campagner, and Federico Cabitza (2024). "Answering the Call to Go Beyond Accuracy: An Online Tool for the Multidimensional Assessment of Decision Support Systems." In: *BIOSTEC (2)*, pp. 219–229.

Natali, Chiara, Lorenzo Famiglini, Andrea Campagner, Giovanni Andrea La Maida, Enrico Gallazzi, and Federico Cabitza (2023). "Color shadows 2: Assessing the impact of xai on diagnostic decision-making." In: *World Conference on Explainable Artificial Intelligence*. Springer, pp. 618–629.

Natali, Chiara, Luca Marconi, Leslye Denisse Dias Duran, and Federico Cabitza (2025). "AI-induced Deskilling in Medicine: A Mixed-Method Review and Research Agenda for Healthcare and Beyond." In: *Artificial Intelligence Review* 58.11, pp. 1–40.

Natali, Chiara, Luca Marconi, Caterina Fregosi, and Federico Cabitza (2024). "Humans in the Group, Computers in the Coop. Comparison of Individual and Collective Improvement in Cognitive Tasks in Adjunct AI Settings." In: *IFIP Working Conference on Human Work Interaction Design*. Springer, pp. 174–191.

Nelson, Caroline A, Lourdes Maria Pérez-Chada, Andrew Creadore, Sara Jiayang Li, Kelly Lo, Priya Manjaly, Ashley Bahareh Pournamdari, Elizabeth Tkachenko, John S Barbieri, Justin M Ko, et al. (2020). "Patient perspectives on the use of artificial intelligence for skin cancer screening: a qualitative study." In: *JAMA dermatology* 156.5, pp. 501–512.

Nielsen, Jakob (1994). "Enhancing the explanatory power of usability heuristics." In: *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pp. 152–158.

Nonaka, Ikujiro and Noboru Konno (1998). "The concept of "Ba": Building a foundation for knowledge creation." In: *California management review* 40.3, pp. 40–54.

Norman, Don (2007). *Emotional design: Why we love (or hate) everyday things*. Basic books.

Norman, Don (2013). *The design of everyday things: Revised and expanded edition*. Basic books.

Odonkor, Beryl, Simon Kaggwa, Prisca Ugomma Uwaoma, Azeez Olanipekun Hassan, and Oluwatoyin Ajoke Farayola (2024). "The impact of AI on accounting practices: A review: Exploring how artificial intelligence is transforming traditional accounting methods and financial reporting." In: *World Journal of Advanced Research and Reviews* 21.1, pp. 172–188.

Ohm, Paul and Jonathan Frankle (2018). "Desirable inefficiency." In: *Fla. L. Rev.* 70, p. 777.

Page, Matthew J, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, et al. (2021). "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews." In: *International journal of surgery* 88, p. 105906.

Panesar, Sandip S, Michel Kliot, Rob Parrish, Juan Fernandez-Miranda, Yvonne Cagle, and Gavin W Britz (2020). "Promises and perils of artificial intelligence in neurosurgery." In: *Neurosurgery* 87.1, pp. 33–44.

Parasuraman, Raja and Dietrich H Manzey (2010). "Complacency and bias in human use of automation: An attentional integration." In: *Human factors* 52.3, pp. 381–410.

Parasuraman, Raja, Thomas B Sheridan, and Christopher D Wickens (2000). "A model for types and levels of human interaction with automation." In: *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans* 30.3, pp. 286–297.

Parchmann, Nina, David Hansen, Marcin Orzechowski, and Florian Steger (2024). "An ethical assessment of professional opinions on concerns, chances, and limitations of the implementation of an artificial intelligence-based technology into the geriatric patient treatment and continuity of care." In: *GeroScience* 46.6, pp. 6269–6282.

Paul, Richard and Linda Elder (2019). *The thinker's guide to Socratic questioning*. Bloomsbury Publishing PLC.

Pesapane, Filippo, Marina Codari, and Francesco Sardanelli (2018). "Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine." In: *European radiology experimental* 2.1, p. 35.

Pierce, James (2012). "Undesigning technology: considering the negation of design by design." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 957–966.

– (2021). "In tension with progression: Grasping the frictional tendencies of speculative, critical, and other alternative designs." In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–19.

Plass, Markus, Michaela Kargl, Patrick Nitsche, Emilian Jungwirth, Andreas Holzinger, and Heimo Müller (2022). "Understanding and explaining diagnostic paths: toward augmented decision making." In: *IEEE Computer Graphics and Applications* 42.6, pp. 47–57.

Polanyi, Michael (1966). "The logic of tacit inference." In: *Philosophy* 41.155, pp. 1–18.

Pugnana, Andrea, Giovanni De Toni, Cesare Barbera, Roberto Pellungrini, Bruno Lepri, and Andrea Passerini (2025). "To Ask or Not to Ask: Learning to Require Human Feedback." In: *arXiv preprint arXiv:2510.08314*.

Pustejovsky, James and Jessica Moszkowicz (2012). "The Role of Model Testing in Standards Development: The Case of ISO-Space." In: *LREC*, pp. 3060–3063.

Rafner, Janet, Dominik Dellermann, Arthur Hjorth, Dora Veraszto, Constance Kampf, Wendy MacKay, and Jacob Sherson (2022). "Deskilling, upskilling, and reskilling: a case for hybrid intelligence." In: *Morals & Machines* 1.2, pp. 24–39.

Ramchurn, Sarvapali D, Sebastian Stein, and Nicholas R Jennings (2021). "Trustworthy human-AI partnerships." In: *Iscience* 24.8.

Rao, Divya (2023). "The Urgent Need for Healthcare Workforce Upskilling and Ethical Considerations in the Era of AI-Assisted Medicine." In: *Indian Journal of Otolaryngology and Head & Neck Surgery*, pp. 1–2.

Rastogi, Charvi, Yunfeng Zhang, Dennis Wei, Kush R Varshney, Amit Dhurandhar, and Richard Tomsett (2022). "Deciding fast and slow: The role of cognitive biases in ai-assisted decision-making." In: *Proceedings of the ACM on Human-computer Interaction* 6.CSCW1, pp. 1–22.

Reicherts, Leon, Gun Woo Park, and Yvonne Rogers (2022). "Extending Chatbots to probe users: Enhancing complex decision-making through probing conversations." In: *Proceedings of the 4th Conference on Conversational User Interfaces*, pp. 1–10.

Rheingold, Howard and Michael Toms (1991). *Tools for thought*. New Dimensions Foundation.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016). "Model-agnostic interpretability of machine learning." In: *arXiv preprint arXiv:1606.05386*.

Riva, Paolo, Nicolas Aureli, and Federica Silvestrini (2022). "Social influences in the digital era: When do people conform more to a human being or an artificial intelligence?" In: *Acta Psychologica* 229, p. 103681.

Robbins, Scott (2024). "Beyond convenience: the ethical use of AI in everyday life." In: *Proceedings of the Workshops at the Third International Conference on Hybrid Human-Artificial Intelligence (HHAI-WS 2024)*.

Roth, E., D. Klein, C. Sushereba, K. Ernst, and L. Militello (2022). "Methods and measures to evaluate technologies that influence

aviator decision making and situation awareness." In: *Contract Report USAARL-TECHCR-2022-22. Applied Decision Science, Cincinnati, OH, USA*.

Ruan, Nantiya (2020). "Attorney Competence in the Algorithm Age." In: *ABAJ Lab. & Emp. L.* 35, p. 317.

Rubegni, Elisa, Chiara Natali, Omran Ayoub, Stefania Maria Rita Rizzo, Clara Valsecchi, and Alessandro Facchini (2025). "Oracles slip on frictionless marble: The case for productive friction in AI-Supported Radiological Work." In: *Unpublished*.

Ruskin, Keith J., Chase Corvin, Stephen C. Rice, and Scott R. Winter (Sept. 2020). "Autopilots in the Operating Room: Safe Use of Automated Medical Technology." In: *Anesthesiology* 133.3. _eprint: https://pubs.asahq.org/anesthesiology/article-pdf/133/3/653/515306/20200900.0-00026.pdf, pp. 653–665. ISSN: 0003-3022. DOI: 10.1097/ALN.0000000000003385. URL: https://doi.org/10.1097/ALN.0000000000003385.

Sambasivan, N. and R. Veeraraghavan (2022). "The deskilling of domain expertise in AI development." In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–14.

Sarkar, Advait (2023a). "Enough with "human-AI collaboration"." In: *Extended abstracts of the 2023 CHI conference on human factors in computing systems*, pp. 1–8.

– (2023b). "Should computers be easy to use? questioning the doctrine of simplicity in user interface design." In: *Extended abstracts of the 2023 CHI conference on human factors in computing systems*, pp. 1–10.

– (2024a). "AI Should Challenge, Not Obey." In: *Communications of the ACM* 67.10, pp. 18–21.

– (2024b). "Intention Is All You Need." In: *arXiv preprint arXiv:2410.18851*.

– (2024c). "Large language models cannot explain themselves." In: *arXiv preprint arXiv:2405.04382*.

Sarkar, Advait, Neil Toronto, Ian Drosos, Christian Poelitz, et al. (2024). "When Copilot Becomes Autopilot: Generative AI's Critical Risk to Knowledge Work and a Critical Solution." In: *arXiv preprint arXiv:2412.15030*.

Schaschek, Myriam, Niko Spatscheck, and Axel Winkelmann (Sept. 2024). "For Those About to Rely-A Taxonomy of Experimental Studies on AI Reliance AI Reliance." In: *19th International Conference on Wirtschaftsinformatik*.

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*, pp. 410–422.

Schemmer, Max, Niklas Kühl, and Gerhard Satzger (2021). "Intelligent decision assistance versus automated decision-making: Enhancing knowledge work through explainable artificial intelligence." In: *arXiv preprint arXiv:2109.13827*.

Schmidt, Kjeld and Liam Bannon (1992). "Taking CSCW seriously: Supporting articulation work." In: *Computer supported cooperative work (CSCW)* 1.1, pp. 7–40.

Schmitt, Anuschka, Thiemo Wambsganß, Matthias Söllner, and Andreas Janson (2021). "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*.

Schön, Donald A (2017). *The reflective practitioner: How professionals think in action*. Routledge.

Schwenk, Charles and Joseph S Valacich (1994). "Effects of devil s advocacy and dialectical inquiry on individuals versus groups." In: *Organizational behavior and human decision processes* 59.2, pp. 210–222.

Seeber, Isabella, Lena Waizenegger, Stefan Seidel, Stefan Morana, Izak Benbasat, and Paul Benjamin Lowry (2020). "Collaborating with technology-based autonomous agents: Issues and research opportunities." In: *Internet Research* 30.1, pp. 1–18.

Selvaraju, Ramprasaath R, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra (2017). "Grad-cam: Visual explanations from deep networks via gradient-based localization." In: *Proceedings of the IEEE international conference on computer vision*, pp. 618–626.

Sengers, Phoebe, Kirsten Boehner, Shay David, and Joseph'Jofish' Kaye (2005). "Reflective design." In: *Proceedings of the 4th decennial conference on Critical computing: between sense and sensibility*, pp. 49–58.

Shafer, Glenn and Vladimir Vovk (2008). "A tutorial on conformal prediction." In: *Journal of Machine Learning Research* 9.3.

Sheridan, Thomas B and Raja Parasuraman (2005). "Human-automation interaction." In: *Reviews of human factors and ergonomics* 1.1, pp. 89–129.

Shneiderman, Ben (2020). "Human-centered artificial intelligence: three fresh ideas." In: *AIS Transactions on Human-Computer Interaction* 12.3, pp. 109–124.

Shors, Tracey J, Megan L Anderson, DM Curlik Ii, and Miriam S Nokia (2012). "Use it or lose it: how neurogenesis keeps the brain fit for learning." In: *Behavioural brain research* 227.2, pp. 450–458.

Silva, Isabella, Pedro Cardoso, and Bruno Giesteira (2022). "Strategies of intentional friction in the user interface of digital games." In: *International Conference on Design and Digital Communication*. Springer, pp. 49–62.

Simone, Carla and Kjeld Schmidt (1993). "Computational Mechanisms of Interaction for CSCW." In.

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman (2013). "Deep inside convolutional networks: Visualising image classification models and saliency maps." In: *arXiv preprint arXiv:1312.6034*.

Sio, Filippo Santoni de and Jeroen Van den Hoven (2018). "Meaningful human control over autonomous systems: A philosophical account." In: *Frontiers in Robotics and AI* 5, p. 323836.

Skaburskis, Andrejs (2008). "The origin of "wicked problems"." In: *Planning Theory & Practice* 9.2, pp. 277–280.

Skitka, Linda J, Kathleen L Mosier, and Mark Burdick (1999). "Does automation bias decision-making?" In: *International Journal of Human-Computer Studies* 51.5, pp. 991–1006.

Smith, Philip J. and Emily Baumann (2020). "Human-Automation Teaming: Unintended Consequences of Automation on User Performance." In: *2020 AIAA/IEEE 39th Digital Avionics Systems Conference (DASC)*, pp. 1–9. DOI: 10.1109/DASC50938.2020.9256418.

Smuha, Nathalie A (2019). "The EU approach to ethics guidelines for trustworthy artificial intelligence." In: *Computer Law Review International* 20.4, pp. 97–106.

Sparrow, Robert and Joshua James Hatherley (2019). "The Promise and Perils of Ai in Medicine." In: *International Journal of Chinese and Comparative Philosophy of Medicine* 17.2, pp. 79–109.

Springer, Aaron and Steve Whittaker (2019). "Progressive disclosure: empirically motivated approaches to designing effective transparency." In: *Proceedings of the 24th international conference on intelligent user interfaces*, pp. 107–120.

Star, Susan Leigh and Anselm Strauss (1999). "Layers of silence, arenas of voice: The ecology of visible and invisible work." In: *Computer supported cooperative work (CSCW)* 8.1, pp. 9–30.

Steegen, Sara, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel (2016). "Increasing transparency through a multiverse analysis." In: *Perspectives on Psychological Science* 11.5, pp. 702–712.

Stepin, Ilia, Jose M Alonso, Alejandro Catala, and Martín Pereira-Fariña (2021). "A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence." In: *IEEE Access* 9, pp. 11974–12001.

Stogiannos, Nikolaos, Tracy O'Regan, Erica Scurr, Lia Litosseliti, Michael Pogose, Hugh Harvey, Amrita Kumar, Rizwan Malik, Anna Barnes, Mark F McEntee, et al. (2025). "Lessons on AI implementation from senior clinical practitioners: An exploratory qualitative study in medical imaging and radiotherapy in the UK." In: *Journal of Medical Imaging and Radiation Sciences* 56.1, p. 101797.

Suchman, Lucille Alice (1987). *Plans and situated actions: The problem of human-machine communication*. Cambridge university press.

– (2007). *Human-machine reconfigurations: Plans and situated actions*. Cambridge university press.

Suchman, Lucy and Caja Thimm (2024). "There is no such thing as a machine that acts outside of relations with humans." In: *Human-Machine Communication* 9, pp. 25–35.

Susskind, Richard and Daniel Susskind (2016). "Technology will replace many doctors, lawyers, and other professionals." In: *Harvard Business Review* 11.

Taleb, Nassim Nicholas (2012). *Antifragile: how to live in a world we don't understand*. Vol. 3. Allen Lane London.

Talib, Manar Abu, Qassim Nasir, Fatima Dakalbab, and Homaiza Saud (2025). "Future Aviation Jobs: The Role of Technology in Shaping Skills and Competencies." In: *Journal of Open Innovation: Technology, Market, and Complexity*, p. 100517.

Taylor, Andrea, Mevagh Sanson, Ryan Burnell, Kimberley A Wade, and Maryanne Garry (2020). "Disfluent difficulties are not desirable difficulties: The (lack of) effect of Sans Forgetica on memory." In: *Memory* 28.7, pp. 850–857.

Tejeda, Heliodoro, Aakriti Kumar, Padhraic Smyth, and Mark Steyvers (2022). "AI-assisted decision-making: A cognitive modeling approach to infer latent reliance strategies." In: *Computational Brain & Behavior* 5.4, pp. 491–508.

The Council of Europe's Ad Hoc Committee on AI (2020). *Towards Regulation of AI Systems*. Tech. rep. Council of Europe.

Tiddi, Ilaria, Victor De Boer, Stefan Schlobach, and André Meyer-Vitali (2023). "Knowledge engineering for hybrid intelligence." In: *Proceedings of the 12th Knowledge Capture Conference 2023*, pp. 75–82.

Tsai, T. L., D. B. Fridsma, and G. Gatti (2003). "Computer decision support as a source of interpretation error: the case of electrocardiograms." In: *Journal of the American Medical Informatics Association* 10.5, pp. 478–483.

Tversky, Amos and Daniel Kahneman (1974). "Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty." In: *science* 185.4157, pp. 1124–1131.

Vallor, Shannon (2013). "The future of military virtue: Autonomous systems and the moral deskilling of the military." In: *2013 5th International Conference on Cyber Conflict (CYCON 2013)*. IEEE, pp. 1–15.

– (2015). "Moral deskilling and upskilling in a new machine age: Reflections on the ambiguous future of character." In: *Philosophy & Technology* 28, pp. 107–124.

Verhagen, Ruben S, Mark A Neerincx, X Jessie Yang, and Myrthe L Tielman (2025). "Advancing Human-Machine Teaming: Definitions, Challenges, Future Directions." In: *HHAI 2025*. IOS Press, pp. 49–59.

Wachter, Sandra, Brent Mittelstadt, and Chris Russell (2018). "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR." In: *Harvard Journal of Law & Technology* 31.2, pp. 841–887.

Wang, Danding, Qian Yang, Ashraf Abdul, and Brian Y Lim (2019). "Designing theory-driven user-centric explainable AI." In: *Proceed-*

ings of the 2019 CHI conference on human factors in computing systems, pp. 1–15.

Wang, Xinru and Ming Yin (2021). "Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making." In: *Proceedings of the 26th International Conference on Intelligent User Interfaces*, pp. 318–328.

Wessel, Nadine-Christine (2023). "Decision-Support Systems and Decision Making: Managing Decisional Deskilling in Human-DSS Interactions in Organizations." In: *ICDS 2023: The Seventeenth International Conference on Digital Society*.

Wiethof, Christina and Eva AC Bittner (2021). "Hybrid Intelligence–Combining the Human in the Loop with the Computer in the Loop: A Systematic Literature Review." In.

Wilde, Oscar (1918). *The Soul of Man Under Socialism*. Boston: John W. Luce and Co., p. 74.

Wilson, Ben, Chiara Natali, Matt Roach, Darren Scott, Alma Rahat, David Rawlinson, and Federico Cabitza (2025). "Dimensions of human-machine combination: prompting the development of deployable intelligent decision systems for situated clinical contexts." In: *Computer Supported Cooperative Work (CSCW)*, pp. 1–57.

Winograd, Terry and Fernando Flores (1986). *Understanding computers and cognition: A new foundation for design*. Ablex, Norwood, NJ.

Winter, P. and A. Carusi (2022). "Professional expectations and patient expectations concerning the development of Artificial Intelligence (AI) for the early diagnosis of Pulmonary Hypertension (PH)." In: *Journal of Responsible Technology* 12, p. 100052.

Wolfe, Christopher R and M Anne Britt (2008). "The locus of the myside bias in written argumentation." In: *Thinking & reasoning* 14.1, pp. 1–27.

Woodger, Joseph Henry (2014). *Biological principles: A critical study*. Routledge.

Xie, Heping, Zongkui Zhou, and Qingqi Liu (2018). "Null effects of perceptual disfluency on learning outcomes in a text-based educational context: A meta-analysis." In: *Educational Psychology Review* 30.3, pp. 745–771.

Xu, Wei, Marvin J. Dainoff, Liezhong Ge, and Zaifeng Gao (2021). "From Human-Computer Interaction to Human-AI Interaction: New Challenges and Opportunities for Enabling Human-Centered AI." In: *ArXiv* abs/2105.05424.

Yao, Yiyu (2009). "Three-way decision: an interpretation of rules in rough set theory." In: *International conference on rough sets and knowledge technology*. Springer, pp. 642–649.

Yin, Ming, Jennifer Wortman Vaughan, and Hanna Wallach (2019). "Understanding the effect of accuracy on trust in machine learning models." In: *Proceedings of the 2019 chi conference on human factors in computing systems*, pp. 1–12.

Yu, Kun, Shlomo Berkovsky, Ronnie Taib, Jianlong Zhou, and Fang Chen (2019). "Do i trust my machine teammate? an investigation from perception to decision." In: *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pp. 460–468.

Zajac, Hubert Dariusz (2022). "Designing ground truth for Machine Learning-conceptualisation of a collaborative design process between medical professionals and data scientists." In: *Proceedings of 20th European Conference on Computer-Supported Cooperative Work*. European Society for Socially Embedded Technologies (EUSSET).

Zhang, W., M. Cai, H. J. Lee, R. Evans, C. Zhu, and C. Ming (2023). "Ai in medical education: Global situation, effects and challenges." In: *Education and Information Technologies*, pp. 1–23.

Zulkipli, Ihsan Nazurah, Faiza Alam, and Mei-Ann Lim (2023). "Integrating AI in medical education: embracing ethical usage and critical understanding." In: *Frontiers in Medicine* 10, p. 1279707.

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede and Ivo Pletikosić. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both LaTeX and LyX:

`https://bitbucket.org/amiede/classicthesis/`