

From Oracles to Productive Friction: Designing for Meaningful Human Oversight in Clinical AI

Elisa Rubegni^{1 5} Chiara Natali⁴, Clara Valsecchi⁶, Stefania Rizzo⁶, Omran Ayoub⁷, Alessandro Facchini^{2 3}

¹Department of Computer Science, Lancaster University, UK

²SUPSI, Dalle Molle Institute for Artificial Intelligence (IDSIA), Switzerland

³Management in Networked and Digital Societies (MINDS) Department, Kozminski University, Poland

⁴Department of Computer Science, Vrije Universiteit Amsterdam, the Netherlands

⁵Department of Management, Economics and Industrial Engineering, Politecnico di Milano, Italy

⁶Institute of integrated diagnostics of Southern Switzerland, Ente Ospedaliero Cantonale (EOC), Switzerland

⁷SUPSI, ISIN, Switzerland

elisa.rubegni@polimi.it, alessandro.facchini@supsi.ch

Abstract

We propose productive friction as a design principle for clinical AI: the deliberate, graded introduction of interaction cost when accepting machine recommendations, calibrated to case uncertainty so that analytical reasoning is activated without disrupting workflow. We ground this principle in a mixed-method study with eleven practising radiologists interacting with two contrasting prototypes, an oracular system delivering a single diagnosis and an evaluative system presenting multiple differentials with reference exemplars. Results reveal a consistent preference for evaluative support in complex cases, where it enhances diagnostic reasoning, confidence, and reflective engagement, while oracular outputs were favoured for routine tasks. Radiologists described evaluative support as “accompanying” their thinking, preserving professional agency and offering pedagogical value. We argue that productive friction carries direct ethical significance: when agreeing with an AI recommendation requires less effort than disagreeing, the clinician’s role risks becoming procedural rather than a genuine locus of judgement. We operationalise this principle along four dimensions, temporal delay, option cardinality, evidence depth, and commitment structure, and translate it into implementable design mechanisms for adaptive clinical AI.

1 Introduction

Clinical AI has made rapid technical strides, yet its uptake in practice often stalls when systems behave like oracles, issuing definitive answers that neither fit clinicians’ workflows nor respect the inherent uncertainty of diagnosis (Parasuraman and Manzey 2010; Cabitza, Rasoini, and Gensini 2017). Our study asks what happens when we instead design *evaluative* decision support systems (Miller 2023) that entertain multiple hypotheses rather than proposing a single suggestion (see Section 2). We report a proof-of-concept study in which eleven radiologists explored and used two simulated AI prototypes in think-aloud sessions followed by semi-structured interviews. One prototype (‘oracular’) presents

a single top diagnostic suggestion; the other (‘evaluative’) presents up to four differential diagnoses with reference exemplars. The study is based on a mixed-method approach. Grounded in prior research in Human-AI Interaction (Faas et al. 2024; Chen et al. 2023), we addressed three research questions:

RQ1. To what extent are radiologists willing to adopt the ‘oracular’ and the ‘evaluative’ AI prototype, and what factors drive their acceptance of each?

RQ2. In which contexts (e.g. case complexity, time pressure, diagnostic uncertainty) and for which clinician profiles (e.g. level of expertise) do radiologists prefer the ‘oracular’ over the ‘evaluative’ paradigm, and vice versa?

RQ3. How does each interaction paradigm (i.e. ‘oracular’ and ‘evaluative’) support or hinder radiologists’ diagnostic process in terms of cognitive load, confidence, and workflow integration?

RQ1 investigates whether and why radiologists would adopt each prototype, RQ2 addresses context-specific preference patterns across clinical conditions and clinician expertise levels, and RQ3 explores how each paradigm affects the diagnostic process itself, together tracing a progression from acceptance, to contextual fit, to process-level impact.

Specifically, we examine how each design affects reasoning, trust, and workflow fit. This design-probe approach, having each participant evaluate and comment on contrasting interaction paradigms instead of a single model, lets us observe how interface choices, not only model accuracy, shape collaborative performance and user agency (Hutchinson et al. 2003). From this comparison, we formulate one cross-cutting *design conjecture*: the two paradigms will embody a productive tension between efficiency and professional agency, with radiologists simultaneously valuing oracular speed and resisting its prescriptive authority. The implications of this tension run deeper than interaction design, touching on ethics. In high-stakes clinical AI, the ease of accepting a recommendation determines whether the human decision-maker has exercised the judgement on which moral and legal accountability depends. If agreeing with the

AI requires less effort than disagreeing, the decision may be formally human but substantively automated. Introducing productive friction in interaction can therefore be seen as more than a design principle; by preserving the epistemic conditions under which responsibility can be meaningfully attributed to a clinician, it functions as an ethical safeguard.

The concept of productive friction is gaining prominence within a wider call for intentionally leveraged friction in AI (Natali et al. 2025b). This friction is related to *cognitive* friction, a term coined by Cooper (1999) to describe “the resistance encountered by a human intellect when it engages with a complex system of rules”. Applied to design, Pierce (2021) recasts such friction as a resource useful for stimulating, among others, speculation, critique, and reflection; while the “productive” qualifier itself was introduced by Madsen, Munk, and Søltoft (2023) to denote a beneficial, reflection-inducing slowing of interaction. Productive friction in AI has been applied mainly to pedagogy (Dekker and Hermans 2026; Konradt, Boote, and Taub 2026; Ayyappan et al. 2026) and creativity (Imteyaz et al. 2026; Viduchinsky, Erel, and Morag Yaar 2026); friction-based strategies in medicine remain comparatively under-explored, but are addressed by both quantitative (Srinivasan and Thomason 2026; Cabitza et al. 2024) and qualitative (Bach et al. 2023; Anichini, Natali, and Cabitza 2024) studies.

Our contribution is thence threefold. First, we propose productive friction as a design strategy for trustworthy clinical AI, operationalised along the four dimensions of temporal delay, option cardinality, evidence depth, and commitment structure, and distinguished from adjacent constructs in the cognitive forcing (Bućinca, Malaya, and Gajos 2021), desirable difficulty (Bjork and Bjork 2011), and seamful design (Chalmers 2003; Ehsan et al. 2024) literatures. We translate these dimensions into implementable mechanisms, including think-first gates, dual-mode interfaces, progressive disclosure, and disagreement workflows, offering concrete design guidance for human-AI teaming in high-stakes settings. Second, our design-probe study provides qualitative evidence that evaluative, differential-first interfaces activate productive friction in complex cases while oracular interfaces are preferred when friction would impede routine work. Adaptive clinical AI must manage this efficiency-agency tension (Swaroop et al. 2024). Third, we argue that productive friction carries ethical significance for meaningful human oversight (Santoni de Sio and Van den Hoven 2018) and the attribution of moral responsibility in AI-assisted diagnosis: when accepting an AI recommendation requires less effort than rejecting it, clinician oversight risks becoming procedural, not a genuine exercise of judgement (Grote and Berens 2020; Matthias 2004). Productive friction sustains the epistemic conditions (independent judgement formation, active weighing of alternatives, and engagement with uncertainty) under which responsibility is meaningfully attributed to the signing clinician, grounded in participants’ accounts of feeling “totally passive” under oracular support and “being made to reason” under evaluative support.

2 Background and Related Work

Supporting Radiological Practice

Clinical decision-making is an exemplary domain for studying AI’s impact on human judgment. In healthcare, AI is positioned to enhance diagnostic accuracy amid workforce shortages (Bienefeld, Keller, and Grote 2025), particularly in under-served settings (Guo and Li 2018). Radiology has served as a flagship case in which decision support systems aim to shift the speciality “from a subjective perceptual skill to a more objective science” (Pesapane, Codari, and Sardanelli 2018, p. 5), while also facing repeated predictions of partial or full automation (Langlotz 2019). This reflects a model of technical rationality (Schön 2017) that treats diagnosis as decomposable problem-solving, sitting uneasily with what radiological work demands in practice. Human radiologists integrate clinical context, prior studies, and tacit knowledge that current AI cannot replicate, what Anichini, Natali, and Cabitza (2024) term ‘vision work’, including the sense-making that determines which findings matter and how they fit cooperative workflows of diagnosis, reporting, and quality assurance. Typical AI applications focus on automating image interpretation tasks such as lesion detection, acting in an ‘oracular’ role by outputting predicted findings for the clinician to accept or reject. While such systems can reduce omission errors, they introduce workflow and decision-making challenges (Parasuraman and Manzey 2010). Radiologists follow a hypothesis-driven process, starting from a clinical question, systematically inspecting images, and generating differential diagnoses, sometimes consulting reference repositories such as *Radiopaedia*¹ (Sorantin et al. 2022), with which a single-output ‘second reader’ may poorly align. It can trigger premature anchoring on the AI’s suggestion (Jussupow et al. 2021) or, more broadly, automation bias affecting both novice and expert operators (Parasuraman and Manzey 2010; Goddard, Roudsari, and Wyatt 2012). An AI acting purely as an oracle can thus inadvertently diminish clinician agency, calling for alternative strategies in the provision of AI support.

Oracular vs Evaluative AI

Already Miller and Masarie Jr (1990) cautioned against ‘oracular’ clinical decision support systems, warning that their inscrutable nature risked generating passive, complacent users, and argued that such systems should instead act as catalysts for physicians’ diagnostic reasoning. *Evaluative AI* can be seen as an heir to this vision, that is a framework emphasising evidence for and against each decision option rather than a single recommendation, aimed at mitigating over-reliance (Miller 2023). Le et al. (2024) define evaluative AI as decision support providing targeted evidence to help users evaluate hypotheses they are already considering, demonstrating that this hypothesis-driven approach improves accuracy and reduces reliance over recommendation-driven baselines. Instead of passively receiving counter-arguments, which risks fixation rather than genuine recon-

¹Radiopaedia website: <https://radiopaedia.org/>

sideration (Aicher et al. 2024; Vasconcelos et al. 2023), evaluative AI engages users in abductive reasoning, iteratively generating and testing their own hypotheses (Miller 2023). Complementary debiasing strategies have been proposed, such as *Judicial AI* (Cabitza et al. 2025; Fregosi, Natali, and Cabitza 2025), where two diagnostic hypotheses are concurrently defended by the AI. Other strategies involve describing uncertainty for both positive and negative outcomes (Bertrand et al. 2022), prompting users to “consider the opposite” (Bach et al. 2023), and requiring classification of evidence as supporting, opposing, or neutral (Lau and Coiera 2009). Each introduces friction at specific decision points. However, these have been studied as isolated interventions rather than instances of a broader design principle. Bućinca, Malaya, and Gajos (2021) demonstrated that cognitive forcing functions can significantly reduce over-reliance, though at the cost of user satisfaction. Empirically, Miller (2023) proposed the evaluative AI framework conceptually without testing it; Le et al. (2024) tested it behaviourally with domain-adjacent participants in skin cancer diagnosis; Cabitza et al. (2025) compared oracular and non-oracular protocols with practising surgeons on fracture classification. All measured accuracy and reliance under controlled experimental conditions. However, this body of work leaves two questions unaddressed. The first is whether the deliberation benefits of cognitive forcing can be preserved while making the added effort graded, context-sensitive, and acceptable in routine clinical work. The second is and how practising clinicians reason about, resist, and selectively adopt contrasting interaction paradigms in ecologically valid settings.

Towards Productive Friction in Medical AI

We address both questions by proposing *productive friction* as a design-level concept that reframes isolated debiasing interventions as instances of a coherent principle. We define it as a deliberate increase in the interaction cost of accepting a machine recommendation, designed so that (a) added cognitive effort activates analytical reasoning proportionate to case uncertainty, (b) the effort remains below the threshold at which clinicians abandon or circumvent the system, and (c) friction degrades gracefully, becoming minimal for routine, low-uncertainty cases. This distinguishes productive friction from three adjacent constructs. Unlike cognitive forcing functions (Bućinca, Malaya, and Gajos 2021), which impose a uniform deliberation barrier before any AI suggestion, productive friction is graded and context-sensitive, varying with case complexity and user expertise.² Unlike desirable difficulty (Bjork and Bjork 2011), its primary goal is in-the-moment decision quality rather than long-term retention, though our data suggest pedagogical benefits as a secondary effect. And unlike seamful design (Chalmers 2003; Inman and Ribes 2019), which reveals infrastructural boundaries, productive friction reveals epistemic ones, the limits of

²Recent evidence shows that deliberation interventions reduce overreliance only when tasks are sufficiently complex that engaging with the explanation is worth the cognitive cost (Vasconcelos et al. 2023), suggesting uniform friction imposes avoidable cost on routine cases.

what the AI (and the clinician) can know about a given case.

We operationalise productive friction along four dimensions that structure our empirical analysis and ground our design implications. *Temporal delay*: when AI output is disclosed relative to the clinician’s own reasoning.³ *Option cardinality*: how many diagnostic alternatives the interface presents. *Evidence depth*: how much supporting and disconfirming evidence is shown per hypothesis. *Commitment structure*: whether the interface requires the clinician to actively select, rank, or justify a choice before proceeding. These dimensions capture both the design variables our two prototypes manipulate and the themes participants spontaneously raised. The evaluative prototype scores higher on option cardinality and evidence depth, while the oracular prototype imposes commitment friction through its disagree-and-justify mechanism. We revisit them as design implications in Section 5.

3 Methodology

Study Design and Protocol

To address our research questions, we conducted a mixed-methods, within-subject proof-of-concept study with practising radiologists. Two simulated AI decision-support prototypes were developed within a radiology viewer environment: one oracular (single top-diagnosis output) and one evaluative (multiple differential hypotheses with supporting evidence). In the study, participants completed 10 diagnostic tasks with each prototype, drawn from a balanced pool of 20 MRI cases curated by an experienced radiologist on the research team. This dual-prototype design operationalises Miller’s evaluative AI framework (2023) and complements prior work on human–AI complementarity (Hemmer et al. 2025, 2021; Ferrario, Facchini, and Durán 2026). Our prototypes function as design probes (Hutchinson et al. 2003): artefacts intentionally situated between concept and deployed product, designed to elicit participants’ situated judgements about contrasting interaction paradigms rather than to measure task performance with a functioning system (Boehner et al. 2007). Each participant engaged with both systems in randomised order during think-aloud sessions, followed by an adapted UTAUT survey and a semi-structured interview. The approach combined quantitative attitudinal measures with qualitative thematic analysis to understand when and why evaluative vs oracular support better fits radiologists’ workflows and reasoning. The study was articulated in four main phases.

Phase 0 – Baseline & Washout. Participants completed a preliminary survey. The first part collected demographic data, professional profiles, and information on participants’ prior use of AI in their clinical practice. They also completed the Technology Anxiety Scale (Wilson et al. 2023). Finally, participants were asked to classify 20 MRI cases without AI assistance to establish baseline accuracy, confidence, and perceived case complexity.⁴ This enabled case-complexity

³Withholding AI assistance until after the user forms an independent judgment reduces overreliance compared to presenting it upfront, though at the cost of decision time (Swaroop et al. 2024).

⁴See full text in (Rubegni et al. 2026, Appendix).

matching across prototypes. This phase was conducted remotely and independently by participants two weeks prior to the main study, helping to mitigate anchoring or recall effects (Cockburn, Quinn, and Gutwin 2017).

Phase 1 – Prototype Interaction I. This phase happened remotely via a Teams call. After a brief interface tutorial and two warm-up cases, participants used the first assigned prototype (randomised order) to complete 10 diagnostic tasks under a think-aloud protocol. Immediately afterwards, they filled out an adapted UTAUT/TAM questionnaire measuring Performance Expectancy, Effort Expectancy, Attitude toward Use, Facilitating Conditions, Job Relevance, Result Demonstrability, Anxiety, and Behavioural Intention (Venkatesh and Bala 2008; Panigutti et al. 2022).⁵

Phase 2 – Prototype Interaction II. Participants repeated the procedure with the other prototype and completed the same survey.⁶

Phase 3 – Post-Study Interview. A semi-structured interview probed perceptions of trust, usability, reasoning support, and preferred contexts for each paradigm. Sessions were audio-recorded, transcribed verbatim, and anonymised.⁷

Participants

In the study, we involved 11 radiologists with a wide range of expertise and seniority in the field of radiology. All participants were fully qualified specialists at the time of the study, each practising and performing diagnostic tasks daily in a public regional hospital in Switzerland. Their professional experience ranged as follows: 2 reported 0–5 years, 5 reported 5–10 years, 1 reported 15–20 years, and 3 senior radiologists reported over 20 years of practice. Participants were recruited through the university network using a snowball sampling approach. Ethical approval was granted by the university's ethics committee, and informed consent, including permission for audio and video recording, was obtained from all participants before the study.

The Prototypes Design, Materials and Tasks

Two web-based interfaces, mimicking decision-support systems, were developed to isolate interaction style rather than algorithmic performance.

Prototype 1 – Evaluative AI. Displays 3–4 plausible diagnostic hypotheses for each case. Each hypothesis is accompanied by concise textual evidence and representative exemplar images. Participants can freely inspect each option before committing to a final decision. This design corresponds to Sheridan's Level 2 autonomy, the computer offers a complete set of decision alternatives; the human decides and acts (Sheridan, Verplank, and Brooks 1978), and embodies a deductive, user-led reasoning model (Figure 1).

Prototype 2 – Oracular AI. Outputs a single top-ranked diagnostic recommendation. Participants may agree or disagree; if they disagree, they must articulate their alternative hypothesis. This aligns with Level 4 autonomy, in which the

computer suggests one alternative and the human acts, representing a prescriptive, system-led interaction (Figure 2).

Both prototypes used the same visual layout, patient metadata, and typology of TC slices to ensure parity of visual input. Only the explanation and interaction structure of the cases differed.

A balanced dataset of 23 TC slices representing varying diagnostic complexity was curated. The cases were provided by an experienced radiologist who is also a member of the project team. Complexity ratings obtained during baseline assessment ensured that both prototypes presented comparable task difficulty. Each prototype displayed 10 cases drawn from this pool (20 unique instances per participant). Cases were assigned to the two prototypes using a fixed, balanced design, in which each prototype was tested on the same 10 cases across all participants, with both sets matched for baseline complexity. The order in which participants used the prototypes was counterbalanced. This fixed assignment may introduce a confound between case content and interaction type; although matching cases by complexity reduces this risk, it does not fully remove it. The same cases were also used in Phase 0.

Data Analysis

This study adopted a concurrent embedded mixed-methods design. The designs allowed survey trends to be interpreted in light of participants' experiential accounts, making it possible to examine how interface choices, and not model accuracy alone, shape diagnostic confidence, trust calibration, and workflow integration. Quantitative and qualitative data were integrated to identify both complementarities and tensions between the two sources of evidence. Given the small sample size ($N = 11$), the quantitative data were used to corroborate the findings emerging from the qualitative analysis rather than to support independent inferential claims.

Quantitative. Primary quantitative outcomes were Behavioural Intention to Use and Performance/Effort Expectancy scores from the UTAUT instrument. Secondary metrics included Result Demonstrability and Facilitating Conditions. Optional task-timing logs were collected when available. Survey items were scored on 5-point Likert scales, reverse-coded as needed, and aggregated per construct. Within-participant contrasts (Evaluative—Oracular) are summarised using medians, interquartile ranges, mean paired differences, and rank-biserial effect sizes (r) to capture the direction and magnitude of attitudinal differences. These patterns are reported in (Rubegni et al. 2026, Appendix), to support interpretation of the qualitative findings that follow.

Qualitative. We analysed the interview and think-aloud transcripts using a hybrid inductive–deductive thematic analysis, following Braun and Clarke's guidelines (Braun and Clarke 2021). This approach allowed us to integrate a priori concerns (derived from our research questions and the concepts from prior literature investigated in Section 2) with themes that emerged directly from the data. *Atlas.ti* was used to support coding and maintain an audit trail. An initial codebook of 16 broad deductive themes (e.g., usability, workflow integration, trust, adoption) was drafted by

⁵See full text in (Rubegni et al. 2026, Appendix).

⁶See full text in (Rubegni et al. 2026, Appendix).

⁷See full text in (Rubegni et al. 2026, Appendix).

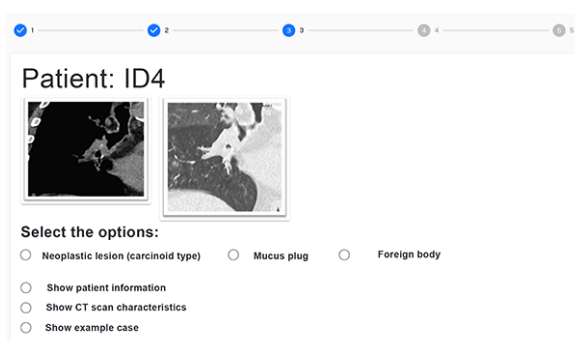


Figure 1: The Evaluative prototype.

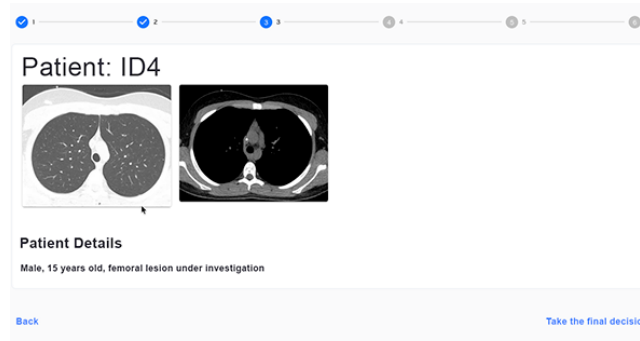


Figure 2: The Oracular prototype.

the research team and refined through discussion. Two researchers applied these categories independently, comparing interpretations and resolving discrepancies. To capture unanticipated insights, we conducted a second round of inductive coding on the single quotes categorised within each broad theme, from which sub-themes were named and refined. For example, within the theme of radiological practice, inductive analysis surfaced specific concerns such as “*Radiology as probabilistic, not absolute*” and “*Discomfort with definitive answers under uncertainty*”. Thematic structures were reviewed across the dataset to ensure internal coherence and distinct boundaries between themes. Because our study does not aim for statistical representativeness of findings (Cornish, Gillespie, and Zittoun 2013), we did not pursue inter-coder reliability in the conventional sense. Instead, we adopted a collaborative analytic approach to enhance rigour and reflexivity. The first and second authors reviewed, refined, and, where appropriate, combined the codes to construct a coherent and meaningful narrative.

The final codebook comprised of 19 codes and 4 high-level deductive themes, each elaborated by inductive sub-themes illustrated with participant quotations. This approach enabled us to address targeted research questions while remaining open to new, data-driven insights into radiologists’ interaction with AI decision support.

Limitations

Our mixed-method study privileges depth over breadth. The single-speciality, single-country sample, likely shaped by self-selection of research-engaged clinicians, may limit generalisability to broader radiology populations or different institutional settings. The design-probe approach captures preferences and reasoning patterns rather than behavioural outcomes, and ecological validity is necessarily limited for this proof-of-concept. Future work should test whether productive friction effects hold under the time pressure and cognitive load of live clinical reading. As noted above, within-subject quantitative contrasts are hypothesis-generating; while effect sizes are informative, they remain imprecise at this scale, and the quantitative strand serves a descriptive and triangulating function throughout. Finally, fixed case-to-prototype assignment introduces a potential confound between case content and in-

teraction paradigm; baseline complexity matching mitigates but does not eliminate this threat.

4 Results

Profile and Professional Expertise with AI

Among the members of the study cohort, ten participants had prior professional AI experience (4 for less than 1 year, 5 for 1–3 years, and 1 for over 5 years) with 7 reporting daily AI use and 2 using it 2–3 times per week. In terms of case volume, 7 had applied AI in more than 100 cases, 2 in 20–100, and 1 in 10–20, reflecting a group largely familiar with clinical AI tools though with varied depth of adoption. Participants described a broad range of everyday AI applications, including detection and recognition (fractures, pulmonary embolism, infiltrates), quantification and measurement (volumes, bone age), image quality optimisation (MR sequences, CT dose reduction), diagnostic and reporting support (draft reports, double-check software), and screening quality assessment (e.g., mammography projection parameters). Since participants entered the study with meaningful AI exposure, enabling them to assess the prototypes’ fit relative to established practices of AI-augmented work rather than reacting to novelty alone, this baseline matters for interpreting subsequent findings.

Attitudinal Baseline

Participants reported generally low levels of technology-related anxiety. We administered the Abbreviated Technology Anxiety Scale (Wilson et al. 2023), adapted to eight items on a 5-point Likert scale (1 = not at all anxious, 5 = extremely anxious). The mean score was 1.37 (SD = 0.38), indicating that respondents overall did not feel anxious about engaging with new technology, including AI systems.⁸

For each construct and condition (Oracular, Evaluative), we computed participant-level scores as the mean of positively keyed items, with negatively phrased items reverse-scored. Table 2 in (Rubegni et al. 2026) reports the median and interquartile range across participants, the mean paired difference (Evaluative – Oracular) and its standard deviation, and rank-biserial effect sizes (r) to characterise the direction

⁸The item set is provided in the Appendix of (Rubegni et al. 2026).

and magnitude of within-participant contrasts. Quantitative contrasts are therefore reported descriptively and interpreted alongside the qualitative findings under each research question.

Cross-Cutting Interface Findings

Two prototype-independent observations conditioned how participants engaged with both systems and are therefore reported before the RQ-specific findings. First, the static, single-slice presentation was repeatedly described as fundamentally incompatible with practice, where radiologists examine multiple slices and planes: “*I find it very difficult not to be able to scroll through the images. . .*” (P1); “*single slices is not our usual way of working. . . it’s more like something from a textbook*” (P5). Static presentation prevented verification across views, making the interaction “*very limiting*” and leading participants to stress that AI must integrate with existing viewers or allow full exploration of the imaging study. Second, minimal clinical context (e.g., “*lesion, male, 80 years old. . .*”) was insufficient for interpretation and “*put me in difficulty*” (P1), since in practice radiologists rely on symptoms and prior history. Without this, AI outputs were harder to interpret and potentially misleading. These constraints applied equally to both prototypes; participants therefore framed multi-slice navigation and integrated patient context as preconditions for usability that any subsequent comparison of interaction paradigms presupposes. We return to both points in the design implications (Section 5).

RQ1: Willingness to Adopt and Acceptance Factors

Quantitative measures revealed a conditional but favourable disposition toward the Evaluative prototype.⁹ Effect sizes consistently favoured the Evaluative paradigm. The largest effects appeared for Facilitating Conditions ($Mdn = 5.0$ vs. 3.4 , $r = .85$) and Job Relevance ($r = .80$), indicating that radiologists perceived the evaluative paradigm as substantially better aligned with their work context and professional needs. Medium-to-large effects were also found for Anxiety ($r = .67$), Result Demonstrability ($r = .63$), Performance Expectancy ($r = .62$), and Attitude toward Use ($r = .55$). Behavioural Intention ($r = .42$) and Effort Expectancy ($r = .34$) showed smaller effects, indicating less differentiation between prototypes in adoption intent and perceived ease of use. Qualitative data explains why adoption intent lagged behind perceived fit. Participants recognised clear value in AI assistance, “*both [prototypes] are useful. . . compared to having no help*” (P9), but expressed a characteristic “*yes, but. . .*” stance that conditioned adoption on five factors. First, **seamless PACS integration**: AI must embed in existing viewers and worklists rather than displace them. Second, **quality over throughput**: “*It improves the quality. . . more so than the quantity*” (P1), and “*It can accelerate the work, but that means you are not doing the reasoning*” (P2). Participants preferred slower, confident decisions over fast but unverified ones, and accepted automation for routine tasks but not at the expense of reasoning.

⁹See Table 2 in (Rubegni et al. 2026, Appendix), the extended version of this work.

Third, **preserved reasoning agency**: clinicians stressed that diagnosis is a multi-step inferential process—image examination, correlation of findings, integration of clinical data—that AI should not bypass: “*The system might give me an idea. . . but then I still have to. . . put together the whole picture*” (P5). The final judgement must remain with the clinician, with AI acting as a second opinion or safety net: “*The decision at the end is still taken by you*” (P7). Fourth, **affordability and validated reliability** (P2). Fifth, **transparent recommendations**: “*Not seeing the reasoning. . . makes it much harder to accept*” (P4). Participants wanted insight into the basis of predictions, including visual features, clinical factors, and supporting cases (P8, P10), and judged the perceived provenance of suggestions—outputs grounded in robust data or prior cases—as more credible than opaque inferences. Trust was similarly conditional and bound up with these adoption conditions. Agreement between AI and clinician increased confidence, “*I feel more supported*” (P7), while disagreement without explanation could “*block*” the decision process (P7). Participants were aware of risks such as bias and over-reliance, “*There’s a risk it could be misleading*” (P6), and many emphasised verifying AI outputs rather than relying on them completely, adopting a cautious stance. The combination of a large Facilitating Conditions effect ($r = .85$) with a more modest Behavioural Intention effect ($r = .42$) thus reflects a coherent pattern: the evaluative paradigm was seen as a strong structural fit for clinical reasoning, but adoption remained contingent on conditions that neither prototype yet fully satisfied. The Anxiety result is illuminating in the same light: both prototypes scored at ceiling ($Mdn = 5.0$), yet the medium-to-large effect ($r = .67$) suggests the *source* of residual unease differed. In the oracular condition it stemmed from prescriptive authority without explanation; in the evaluative condition it stemmed from the cognitive cost of comparing multiple options, a qualitatively different and arguably more productive form of friction, whose implications we return to under RQ3.

RQ2: Preferences Across Cases and Clinicians

Clear distinctions emerged between the two prototypes, mapped consistently onto task complexity, time pressure, and user expertise.

The **Oracular prototype was preferred for simple, binary decisions** such as fracture detection, routine straightforward cases, time-critical ER triage, and high-volume repetitive tasks. Experienced radiologists in particular valued fast confirmation (P2). Several explicitly cast the AI in a detection role rather than a diagnostic one, “*The system should just show me where the problem might be. Then I will go investigate it myself*” (P4), and valued AI for catching oversights, especially under fatigue (P5, P6), functioning as a quality-assurance layer. Participants identified concrete integration points consistent with this framing: AI triage in high-volume settings (P2), routine detection tasks analogous to existing fracture-detection systems (P6), and rapid confirmation in clearly straightforward cases (P7). For simple tasks, a single-answer system could provide rapid confirmation without disrupting diagnostic judgement. Instead, the **Evaluative prototype was favoured for complex di-**

agnostic cases involving uncertainty, ambiguous or atypical findings, and learning contexts. *“In doubtful cases it might put forth diagnostic hypotheses I hadn’t thought of”* (P3). Participants explicitly contrasted the two paradigms: rapid clear-cut decisions benefit from the single-answer system, whereas complex CT interpretations require the multiple-hypothesis approach. Some rejected the idea of trusting one-shot answers in ambiguous cases outright, *“that’s like playing roulette”* (P5), and worried that systems presenting a single answer *“as if it’s 100% certain”* (P1) were unrealistic and potentially unsafe given that radiology rarely yields a single certain diagnosis in complex cases. These context-specific preferences resolved into a stance of **selective reliance**, with AI as a guide in unfamiliar domains, but secondary to one’s own judgement in routine cases (P6). Differences across expertise levels reinforced the pattern. Experts preferred fast confirmation (P2), while trainees stood to benefit from exploratory systems that supported learning (P8). Participants were nonetheless concerned about skill erosion, noting that trainees should still develop foundational skills without over-reliance on AI (P1). The large Facilitating Conditions effect ($r = .85$) is consistent with this distribution: the evaluative prototype fit the multi-step reasoning sequence radiologists already follow, whereas the oracular prototype’s single output was experienced as a shortcut that bypassed, rather than supported, established verification steps. Overall, our findings indicate that the appropriate paradigm is contingent on case complexity and clinician experience, motivating decision support that can adaptively offer either mode.

RQ3: Support for the Diagnostic Process

The Evaluative prototype demonstrated superior overall alignment with radiological practice, supported by consistent positive quantitative contrasts across Performance Expectancy ($r = .62$), Result Demonstrability ($r = .63$), and Facilitating Conditions ($r = .85$). Three interrelated qualitative properties account for this superiority.

First, respect for radiology’s probabilistic nature. Participants framed their work as offering possible answers rather than verdicts, *“Our role... is to offer possible answers”* (P7), and saw the oracular framing as a category mismatch with how diagnosis actually proceeds. Inductive analysis surfaced specific concerns about *“radiology as probabilistic, not absolute”* and *“discomfort with definitive answers under uncertainty”*, both of which the evaluative interface accommodated by design. Systems presenting a single answer *“as if it’s 100% certain”* (P1) were experienced as misaligned with established practice.

Second, preservation of clinical reasoning. Participants stressed that diagnosis is an investigative process that should not be short-circuited. Providing answers without reasoning risked rendering clinicians *“totally passive”* (P5), and over-simplified systems were viewed as detrimental: *“It’s too easy... we’d study medicine in one year”* (P5). Multi-option interfaces were valued because they *“accompany you in making the decision”* (P7), a phrase repeatedly contrasted with the binary acceptance posture invited by the single-answer system. Single-answer outputs in ambiguous cases

were rejected outright: *“that’s like playing roulette”* (P5). Several participants described **constructive friction** in concrete terms: spending *“one extra minute”* considering alternatives increased confidence (P3), suggesting that some deliberation can improve decision quality without unacceptable cost. The investigative process, as much as its output, was experienced as constitutive of professional judgement.

Third, pedagogical value. The evaluative prototype was seen as a learning aid that encouraged comparison and deeper engagement with cases, particularly through reference images and the breadth of differential hypotheses offered. It was repeatedly described as *“making me reason”* (P8, P3, P9). Trainees in particular benefited from the exploratory structure in ways that experienced practitioners did not need; experienced radiologists, by contrast, sometimes found multiple options redundant for routine tasks (P2), reinforcing the case for adaptivity rather than uniform depth. The Oracular prototype retains merit for streamlining routine confirmations and quick verification in straightforward cases, but its prescriptive framing was consistently resisted under uncertainty. Returning to the Anxiety result reported under RQ1, the asymmetry now reads as substantive rather than incidental, with residual unease under oracular support tracking the prescriptive authority of an unexplained verdict, and under evaluative support tracking the cognitive cost of weighing alternatives. The first form of friction obstructs reasoning, whereas the second activates it.

Taken together, these findings confirm our design conjecture. Beyond differences in preference rankings, the two paradigms stage a productive tension between efficiency and professional agency. Radiologists simultaneously valued oracular speed, particularly for routine cases and under time pressure, and resisted its prescriptive framing when uncertainty was present. We therefore approach this tension as a persistent feature of the design space, one that adaptive, context-sensitive support must address by calibrating deliberation cost to diagnostic uncertainty. The design and ethical implications of this position are developed next.

5 Discussion and Implications for Design

Productive Friction as a Design Principle

The empirical contrasts reported in Section 4 do more than rank two prototypes. They reveal a structural feature of clinical AI design, one in which efficiency and professional agency form a persistent design condition irreducible to a choice between them. The efficiency–agency tension formalised in our design conjecture maps directly onto the four productive friction dimensions introduced in Section 2 and summarised in Table 1. Each dimension constitutes a designable lever through which deliberation cost can be calibrated to diagnostic uncertainty, such that it remains high enough to activate analytical reasoning in complex cases, and low enough to avoid disrupting routine workflow. Crucially, participants did not experience these dimensions as abstract interface properties but as felt differences in agency. *Temporal delay* surfaced when P3 and P8 described forming an impression before AI output and gaining reflective distance from the suggestion. *Option cardinality* surfaced in the

Dimension	Definition	Evidence from data	Key participants
Temporal delay	Withholding AI output until the clinician forms an initial impression	Forming a hypothesis before seeing AI reduced anchoring and increased reflection	P3, P8
Option cardinality	Number of diagnostic alternatives displayed by the system	3-4 options prompted active comparison; a single option invited passive acceptance	P2, P5, P7
Evidence depth	Amount of supporting/disconfirming evidence shown per hypothesis	Clinicians wanted progressive disclosure, not uniformly imposed detail	P4, P8, P10
Commitment structure	Whether the interface requires active selection, ranking, or justification	Free inspection engaged more than agree/disagree; forced articulation productive only with a competing hypothesis	P7

Table 1: Productive friction dimensions mapped to empirical evidence.

contrast between being “*accompanied*” in reasoning (P7) and “*playing roulette*” (P5) on a single output. *Evidence depth* surfaced as the rejection of opaque outputs (“*not seeing the reasoning... makes it much harder to accept*”, P4) coupled with explicit demand for clinician-controlled disclosure rather than uniform detail. *Commitment structure* surfaced asymmetrically: forced articulation was productive when participants held a competing hypothesis but “*blocking*” (P7) when they did not. Table 1 consolidates these mappings between dimension, definition, and empirical evidence. Read this way, the prototype contrast is evidence that the four dimensions can each be tuned independently and must be tuned, since their optimal settings vary with case complexity, time pressure, and clinician expertise. This reframing carries weight beyond interaction design, bearing directly on the ethical standing of clinician oversight, to which we turn next.

Productive Friction as Ethical Infrastructure

Frictionless compliance creates a gap between formal responsibility, where the clinician signs the report, and substantive responsibility, where the clinician has exercised judgement. This transposes to the human-AI decision dyad the ‘responsibility gap’ that Matthias (2004) identified for learning automata. Where accepting AI output is easier than rejecting it, the clinician retains legal liability without having performed the cognitive work that liability presupposes. In such cases, the clinician’s signature continues to function as formal sign-off, even though the deliberation it is meant to certify has not occurred. The concern has been taken up in both philosophical and regulatory discourse. The concept of meaningful human control (Santoni de Sio and Van den Hoven 2018) has become central to both, most concretely in the EU AI Act (Article 14), which requires “meaningful human oversight” for high-risk AI systems including medical diagnostic support. However, neither the philosophical account nor the regulation specifies how meaningfulness should be evidenced at the interface level. The four productive friction dimensions we propose (temporal delay, option cardinality, evidence depth, and commitment structure) can be read as designable operationalisations of what meaningful oversight demands in practice, translating a normative aspiration into concrete interface proper-

ties that determine whether a clinician’s engagement with an AI recommendation is substantive or ceremonial. Grote and Berens (2020) have argued that optimising for compliance efficiency in high-stakes automated systems may systematically undermine the human judgement that regulation assumes is present. Under oracular support, several reported “*feeling blocked*” (P7) or “*totally passive*” (P5), corresponding to the erosion of judgement conditions that Grote and Berens theorise. Under evaluative support, the same participants described “*being accompanied*” in their reasoning (P7) and “*made to reason*” (P8), pointing to the preservation of precisely those conditions. Read through an ethical lens, these statements constitute phenomenological evidence of the conditions under which clinicians experience themselves as responsible agents versus ratifiers of machine output. When P5 described the oracular prototype as making diagnosis “*too easy*”, the concern was not aesthetic but about the felt disappearance of the investigative process that grounds professional accountability. When P3 described “*spending one extra minute*” considering alternatives and gaining confidence, the gain was not merely psychological comfort but the epistemic standing to defend and be held responsible for the resulting diagnosis. Productive friction, on this account, is thus an ethical intervention that operates through usability.

Contextualising the Findings

Radiological diagnosis is a distributed, collaborative process embedded in a professional infrastructure of routines, norms, and tacit expectations that shape both *vision work* (Anichini, Natali, and Cabitza 2024) and how expertise is recognised. Radiologists produce diagnoses while also enacting a role that carries authority and accountability for patient care. Adding AI support requires alignment with existing routines, roles, and responsibilities. A system efficient at the point of interaction may still be organisationally disruptive if it displaces interpretive labour, obscures uncertainty, or shifts accountability to the wrong actors (cfr. *articulation work* by Schmidt and Bannon 1992). This is especially true of the oracular model. A single ‘answer’ risks recasting the AI as the primary source of truth, diminishing the radiologist’s visible contribution and eroding agency, skill display, and professional recognition (Natali et al. 2025a). Re-

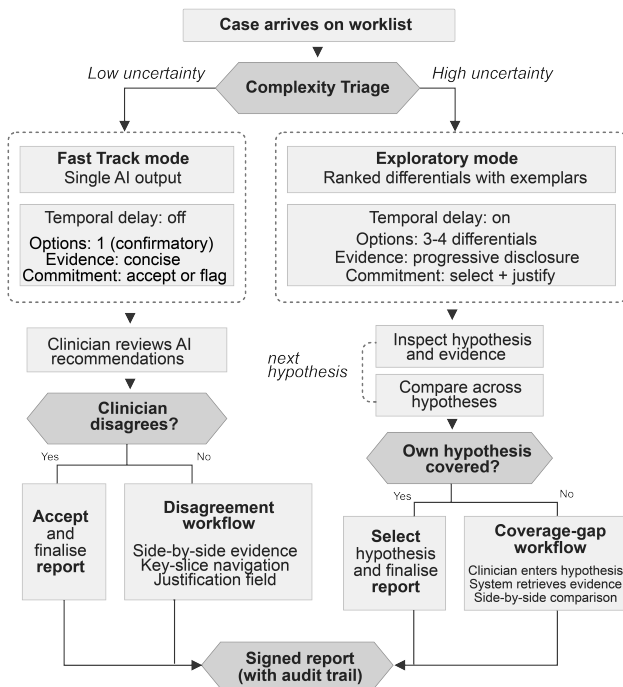


Figure 3: Cases are routed by complexity to Fast Track (single AI diagnosis, minimal friction) or Exploratory mode (multiple differentials, graded friction).

sponsibility may remain formally human, yet the tool still shapes the decision space, since the easier it is to accept a recommendation, the more likely automation bias becomes (Skitka, Mosier, and Burdick 1999), and the less visible the clinician’s independent reasoning is to colleagues, institutions, and the clinician herself. This is the structural basis of the efficiency–agency tension our findings document, where frictionless compliance optimises throughput at the cost of reflective engagement, while deliberative friction preserves judgement at the cost of speed. These observations connect to a counter-tradition in HCI that treats friction as a resource rather than a defect, from reflective design (Sengers et al. 2005) to slow technology (Hallnäs and Redström 2001) and seamful design (Chalmers 2003; Inman and Ribes 2019; Ehsan et al. 2024). However, these traditions have addressed aesthetic, temporal, and infrastructural friction without being applied to the specific problem of preserving the epistemic conditions for accountability in AI-assisted clinical decisions, a gap the next subsection addresses.

Implications for Design

Our findings suggest that clinical AI should be designed around radiologists’ work practices rather than AI capabilities. Participants asked for tools that fit into their existing environment, expose their reasoning, preserve agency, and support both fast confirmation and deeper learning. We translate these needs into the following design implications. AI should be embedded where work already happens. Static, single-slice mock-ups were repeatedly described as “very

limiting” because radiologists need to traverse slices and planes to verify the diagnosis (P1, P5). AI should therefore be PACS-native, scrollable, zoomable, and multi-planar, with overlays that can be toggled on and off within the normal reading flow. Importantly, interfaces should reveal essential patient context (e.g. age, symptoms, and prior findings) alongside the image and the AI output, and make clear which fields the model used. This supports interpretation and helps clinicians judge whether a suggestion is plausible for the specific case. To further support their workflow, participants proposed the inclusion of an AI triage lane to the worklist. Participants saw clear value in prioritising urgent findings and emergency cases (P2, P6). If studies are reordered, the criteria should remain transparent and the evidence immediately visible, so that prioritisation supports human verification. They also saw potential in AI for co-authoring, not auto-authoring, the clinical report. Although auto-generated text can save time, radiologists emphasised that outputs remain editable and accountable (P8, P2). Draft findings and impressions should therefore be presented with provenance and links back to evidential slices, so that editing is faster than rewriting without removing responsibility from the clinician. This suggests a dual-mode interface, namely a *Fast Track* presenting a single confirmatory output for low-uncertainty cases, and an *Exploratory* mode presenting ranked differentials with exemplars for complex or ambiguous ones. The threshold between modes could be set by case-level uncertainty scores, user preference, or both. Swaroop et al. (2025) showed that adapting the type of AI assistance based on each user’s observed overreliance rate (providing higher-friction support to users who tend to accept AI suggestions uncritically, and lower-friction support to well-calibrated users) improves human-AI team performance compared to uniform assistance. Applied to the radiological context, this suggests that the friction dimensions we propose need not be set solely by case complexity or professional seniority. A system that monitors how a radiologist interacts with AI over time, tracking agreement rates, time-to-decision, and frequency of independent verification, could dynamically calibrate the level of friction to match the individual’s current reliance patterns, offering a more precise and responsive adaptation than role-based presets alone. AI outputs should make uncertainty legible, not hide it. Since definitive claims without evidence clash with how radiologists reason and report (P7, P1), participants explicitly rejected outputs that hide the basis of a suggestion (P4, P8, P10). Explanations should therefore be concise by default but allow progressive disclosure into salient image regions, key factors, and reference cases, so that explanation functions as practical verification rather than decorative justification. A differential-first interface presenting a small set of likely diagnoses with discriminative cues better mirrors how clinicians reason and report, while encouraging breadth before commitment. Figure 3 synthesises these mechanisms into an adaptive workflow that routes cases through context-sensitive friction. We now detail the design mechanism for each of the four friction dimensions identified in Section 2 and empirically grounded in Table 1.

Temporal delay. The reflective engagement participants

described when forming an impression before AI output (P3, P8) translates into a ‘think-first’ gate that withholds AI output until the clinician has recorded an initial impression. This gate could be toggleable by the user or triggered automatically for high-uncertainty cases, defaulting to off for urgent triage and on for complex diagnostic workups, consistent with experimental evidence that ‘AI-after’ conditions reduce overreliance on hard tasks at a modest time cost (Swaroop et al. 2024). By ensuring the clinician forms an independent impression before AI output appears, the gate transforms the subsequent decision into a comparison between two judgements.

Option cardinality. The contrast between being “accompanied” in reasoning (P7) and “playing roulette” (P5) translates into a requirement that cardinality be adaptive rather than fixed, with experienced radiologists finding multiple options redundant for routine tasks (P2), while complex cases warrant the full differential set. A dual-mode interface could present a single confirmatory output for low-uncertainty cases and ranked differentials with exemplars for complex ones, with the threshold set by case-level uncertainty scores or user preference. This preserves the clinician’s role as the agent who selects among alternatives.

Evidence depth. Participants’ rejection of opaque outputs (P4, P8, P10), together with the observation that no participant asked for all evidence simultaneously, supports progressive disclosure, with a concise default view showing the hypothesis and its confidence, that can be expanded on demand to salient image regions, discriminative cues, and case-linked exemplars. Clinician-controlled disclosure ensures that engagement with evidence is proportionate to perceived uncertainty rather than imposed uniformly, directly addressing the satisfaction cost that Bućinca, Malaya, and Gajos (2021) identified for uniform cognitive forcing.

Commitment structure. The asymmetry between productive articulation when a competing hypothesis is already held and “blocking” articulation when it is not (P7) translates into commitment mechanisms that are adaptive rather than mandatory. A lightweight implementation could record the clinician’s initial impression before displaying AI output and flag disagreements for structured resolution, presenting side-by-side evidence and a short justification field, without forcing articulation when no competing hypothesis exists. In the evaluative mode, the analogous situation is a coverage gap, according to which when the clinician’s hypothesis is absent from the presented set, the appropriate response is to help retrieve supporting evidence and compare it against listed alternatives, maintaining the collaborative stance participants valued. In both modes, adaptive commitment ensures that finalising a diagnosis feels like a deliberate choice. Together, these four dimensions constitute a design space for future work, for instance, varying temporal delay and option cardinality factorially across case complexity levels. Several participants described a productive ‘extra minute’ that improved confidence when the system prompted consideration of alternatives (P3, P5, P7), suggesting that even modest increases along these dimensions yield disproportionate gains in reflective engagement. The evaluative prototype was also seen as valuable for trainees precisely because it

“made me reason” (P8, P3, P9), pointing to optional micro-explanations and case-linked exemplars as a means of turning routine AI use into learning opportunity without disrupting workflow. As Figure 3 illustrates, both modes include resolution mechanisms, a disagreement workflow for Fast Track and a coverage-gap workflow for Exploratory that scaffolds clinician reasoning. The result is a clinical AI that is context-aware, differential-first, and adaptable; one that alternates between fast confirmation and deeper exploration, shows its working, and introduces gentle contextual friction to protect judgement without usurping the clinician’s role.

6 Conclusion

Through a design-probe study with practising radiologists, we contrasted oracular and evaluative paradigms of AI decision support, finding that interaction design plays as critical role as model accuracy in shaping whether AI supports or supplants clinical reasoning. Evaluative, differential-first interfaces activated reflection, preserved professional agency, and aligned with the probabilistic logic of radiological diagnosis, while oracular outputs were valued for speed in routine cases but resisted when they foreclosed deliberation under uncertainty. The ease of accepting an AI recommendation influences the clinician’s exercise of their judgement, on which moral and legal accountability depends. Productive friction, operationalised through temporal delay, option cardinality, evidence depth, and commitment structure, preserves the epistemic conditions under which a signed report represents a judgement rather than a default acceptance. Efficiency and agency thus constitute a design tension to be managed through adaptive, context-sensitive support that calibrates deliberation cost to diagnostic uncertainty. Taking this tension seriously, clinical AI should operate as a collaborator that keeps reasoning visible, uncertainty legible, and the final decision the clinician’s. **Future work** should replicate these findings with larger multi-site samples and higher-fidelity prototypes in live PACS. Beyond such consolidation, three directions seem consequential. First, the friction mechanisms we propose generate auditable interaction traces, pre-AI impressions, evidence inspection patterns, justification entries, that could serve as accountability infrastructure, providing regulators with designable evidence of meaningful oversight. The EU AI Act (Article 14) requires meaningful human oversight for high-risk AI without specifying how it should be demonstrated at the interface level; productive friction may offer a concrete answer. Second, the question of who controls friction intensity is itself ethically charged. System-set thresholds introduce meta-level automation risk, while clinician-controlled settings may incentivise self-serving minimisation under time pressure. How friction authority should be allocated among system, clinician, and institution remains an open design-governance problem. Third, frictionless support over months and years may cumulatively erode the very expertise on which meaningful oversight depends (Natali et al. 2025a). Longitudinal studies comparing skill trajectories under frictional and frictionless conditions would test whether productive friction serves as both a per-decision safeguard and a structural defence against deskilling.

Acknowledgments

The work presented in the paper was part of the exploratory research programme *Best4EthicalAI*, financed by SUPSI. We would like to thank SUPSI for supporting our research, as well as all the individuals who have been involved in the study. Participants gave their consent to participate voluntarily in the case study that was approved by the SUPSI ethical committee. The work of AF was also partly conducted within the framework of the EUonAIR Centre of Excellence in Responsible AI and Education, and partially supported by a grant from Movetia, funded by the Swiss Confederation. CN acknowledges the financial support provided by the Federal Commission for Scholarships for Foreign Students in the form of the Swiss Government Excellence Scholarship (ESKAS No. 2024.0002) for the academic year 2024–25. We are grateful to all three AIES reviewers for their helpful comments and insightful feedbacks on an earlier version of this work.

References

- Aicher, A.; Matsuda, Y.; Yasumoto, K.; Minker, W.; André, E.; and Ultes, S. 2024. Enhancing reflective and conversational user engagement in argumentative dialogues with virtual agents. *Multimodal Technologies and Interaction*, 8(8): 71.
- Anichini, G.; Natali, C.; and Cabitza, F. 2024. Invisible to machines: designing AI that supports vision work in radiology. *Computer Supported Cooperative Work (CSCW)*, 33(4): 993–1036.
- Ayyappan, D.; Hernández-Leo, D.; Larson, R.; Elhai, L.; Chatain, J.; and Li, G. 2026. Understanding and Engaging Critical Resistance to AI in Education. In *Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems*, 1–4.
- Bach, A. K. P.; Nørgaard, T. M.; Brok, J. C.; and Van Berkel, N. 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*, 1–14.
- Bertrand, A.; Belloum, R.; Eagan, J. R.; and Maxwell, W. 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*, 78–91.
- Bienefeld, N.; Keller, E.; and Grote, G. 2025. AI interventions to alleviate healthcare shortages and enhance work conditions in critical care: qualitative analysis. *Journal of Medical Internet Research*, 27: e50852.
- Bjork, E. L.; and Bjork, R. A. 2011. Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. *Psychology and the real world: Essays illustrating fundamental contributions to society*, 2(59-68): 56–64.
- Boehner, K.; Vertesi, J.; Sengers, P.; and Dourish, P. 2007. How HCI interprets the probes. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 1077–1086.
- Braun, V.; and Clarke, V. 2021. *Thematic analysis: A practical guide*. SAGE publications Ltd.
- Buçinca, Z.; Malaya, M. B.; and Gajos, K. Z. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): 1–21.
- Cabitza, F.; Famiglini, L.; Fregosi, C.; Pe, S.; Parimbelli, E.; La Maida, G. A.; and Gallazzi, E. 2025. 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, IUI ’25*, 745–754. ACM.
- Cabitza, F.; Natali, C.; Famiglini, L.; Campagner, A.; Caccavella, V.; and Gallazzi, E. 2024. Never tell me the odds: Investigating pro-hoc explanations in medical decision making. *Artificial intelligence in medicine*, 150: 102819.
- Cabitza, F.; Rasoini, R.; and Gensini, G. F. 2017. Unintended consequences of machine learning in medicine. *Jama*, 318(6): 517–518.
- Chalmers, M. 2003. Seamful design and ubicomp infrastructure. In *Proceedings of Ubicomp 2003 workshop at the crossroads: The interaction of HCI and systems issues in Ubicomp*, volume 4, 577–584.
- Chen, V.; Liao, Q. V.; Wortman Vaughan, J.; and Bansal, G. 2023. Understanding the role of human intuition on reliance in human-AI decision-making with explanations. *Proceedings of the ACM on Human-computer Interaction*, 7(CSCW2): 1–32.
- Cockburn, A.; Quinn, P.; and Gutwin, C. 2017. The effects of interaction sequencing on user experience and preference. *International Journal of Human-Computer Studies*, 108: 89–104.
- Cooper, A. 1999. *The Inmates Are Running the Asylum: Why High-Tech Products Drive Us Crazy and How to Restore the Sanity*. Indianapolis, IN: Sams Publishing. ISBN 0-672-31649-8.
- Cornish, F.; Gillespie, A.; and Zittoun, T. 2013. Collaborative analysis of qualitative data. *Handbook of qualitative data analysis*, SAGE, 2013/6//79-93.
- Dekker, I.; and Hermans, F. 2026. Why the User-Friendliness of AI Undermines Education. Commentary on Wieczorek 2025. *Philosophy & Technology*, 39(2): 82.
- Ehsan, U.; Liao, Q. V.; Passi, S.; Riedl, M. O.; and Daumé III, H. 2024. Seamful xai: Operationalizing seamful design in explainable ai. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1): 1–29.
- Faas, C.; Bergs, R.; Sterz, S.; Langer, M.; and Feit, A. M. 2024. Give Me a Choice: The Consequences of Restricting Choices Through AI-Support for Perceived Autonomy, Motivational Variables, and Decision Performance. *arXiv preprint arXiv:2410.07728*.
- Ferrario, A.; Facchini, A.; and Durán, J. M. 2026. Epistemology gives a Future to Complementarity in Human-AI Interactions. *arXiv preprint arXiv:2601.09871*.

- Fregosi, C.; Natali, C.; and Cabitza, F. 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 co-located with the Fourth International Conference on Hybrid Human-Artificial Intelligence (HHAI2025)*. CEUR-WS.
- Goddard, K.; Roudsari, A.; and Wyatt, J. C. 2012. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1): 121–127.
- Grote, T.; and Berens, P. 2020. On the Ethics of Algorithmic Decision-Making in Healthcare. *Journal of Medical Ethics*, 46(3): 205–211.
- Guo, J.; and Li, B. 2018. The application of medical artificial intelligence technology in rural areas of developing countries. *Health equity*, 2(1): 174–181.
- Hallnäs, L.; and Redström, J. 2001. Slow technology—designing for reflection. *Personal and ubiquitous computing*, 5(3): 201–212.
- Hemmer, P.; Schemmer, M.; Kühn, N.; Vössing, M.; and Satzger, G. 2025. Complementarity in human-AI collaboration: Concept, sources, and evidence. *European Journal of Information Systems*, 34(6): 979–1002.
- Hemmer, P.; Schemmer, M.; Vössing, M.; and Kühn, N. 2021. Human-AI Complementarity in Hybrid Intelligence Systems: A Structured Literature Review. *PACIS*, 78.
- Hutchinson, H.; Mackay, W.; Westerlund, B.; Bederson, B. B.; Druin, A.; Plaisant, C.; Beaudouin-Lafon, M.; Conversy, S.; Evans, H.; Hansen, H.; et al. 2003. Technology probes: inspiring design for and with families. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 17–24.
- Imteyaz, K.; Muller, M.; Flores-Saviaga, C.; and Savage, S. 2026. Co-Designing Collaborative Generative AI Tools for Freelancers. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, 1–21.
- Inman, S.; and Ribes, D. 2019. “Beautiful Seams” Strategic Revelations and Concealments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14.
- Jussupow, E.; Spohrer, K.; Heinzl, A.; and Gawlitza, J. 2021. Augmenting medical diagnosis decisions? An investigation into physicians’ decision-making process with artificial intelligence. *Information Systems Research*, 32(3): 713–735.
- Konradt, L.; Boote, D. N.; and Taub, M. 2026. Making AI Annoying on Purpose: When Helpful Tools Don’t Always Help. *Computers and Education Open*, 100343.
- Langlotz, C. P. 2019. Will artificial intelligence replace radiologists? *Radiology: Artificial Intelligence*, 1(3): e190058.
- Lau, A. Y.; and Coiera, E. W. 2009. Can cognitive biases during consumer health information searches be reduced to improve decision making? *Journal of the American Medical Informatics Association*, 16(1): 54–65.
- Le, T.; Miller, T.; Sonenberg, L.; and Singh, R. 2024. Towards the new XAI: A hypothesis-driven approach to decision support using evidence. In *ECAI 2024: 27th European Conference on Artificial Intelligence, 19–24 October 2024, Santiago de Compostela, Spain—including 13th Conference on Prestigious Applications of Intelligent Systems (PAIS 2024)*, 850–857. SAGE Publications Pvt. Ltd 1 Oliver’s Yard, 55 City Road, London, EC1Y 1SP.
- Madsen, A. K.; Munk, A. K.; and Søltoft, J. I. 2023. Friction by machine: How to slow down reasoning with computational methods. *Ethnographic Praxis in Industry Conference Proceedings*, 2023(1): 65–88.
- Matthias, A. 2004. The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata. *Ethics and Information Technology*, 6(3): 175–183.
- Miller, R. A.; and Masarie Jr, F. 1990. The demise of the “Greek Oracle” model for medical diagnostic systems. *Methods of information in medicine*, 29(01): 1–2.
- Miller, T. 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*, 333–342.
- Natali, C.; Marconi, L.; Dias Duran, L. D.; and Cabitza, F. 2025a. AI-induced Deskillling in Medicine: A Mixed-Method Review and Research Agenda for Healthcare and Beyond. *Artificial Intelligence Review*, 58(11): 1–40.
- Natali, C.; Naiseh, M.; Cabitza, F.; and Frischmann, B. 2025b. Better AI with Designed Friction: Theories, Applications and Research Agenda. In *HHAI 2025: Proceedings of the 4th International Conference on Hybrid Human-Artificial Intelligence*, 518–520. SAGE Publications 1 Oliver’s Yard, 55 City Road, London, EC1Y 1SP.
- Panigutti, C.; Beretta, A.; Giannotti, F.; and Pedreschi, D. 2022. Understanding the impact of explanations on advice-taking: a user study for AI-based clinical Decision Support Systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–9.
- Parasuraman, R.; and Manzey, D. H. 2010. Complacency and bias in human use of automation: An attentional integration. *Human factors*, 52(3): 381–410.
- Pesapane, F.; Codari, M.; and Sardanelli, F. 2018. Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine. *European radiology experimental*, 2(1): 35.
- Pierce, J. 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*, 1–19.
- Rubegni, E.; Natali, C.; Valsecchi, C.; Rizzo, S.; Ayoub, O.; and Facchini, A. 2026. From Oracles to Productive Friction: Designing for Meaningful Human Oversight in Clinical AI.
- Santoni de Sio, F.; and Van den Hoven, J. 2018. Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI*, 5: 323836.
- Schmidt, K.; and Bannon, L. 1992. Taking CSCW seriously: Supporting articulation work. *Computer supported cooperative work (CSCW)*, 1(1): 7–40.

- Schön, D. A. 2017. *The reflective practitioner: How professionals think in action*. Routledge.
- Sengers, P.; Boehner, K.; David, S.; and Kaye, J. 2005. Reflective design. In *Proceedings of the 4th decennial conference on Critical computing: between sense and sensibility*, 49–58.
- Sheridan, T. B.; Verplank, W. L.; and Brooks, T. 1978. Human/computer control of undersea teleoperators. In *NASA. Ames Res. Center The 14th Ann. Conf. on Manual Control*.
- Skitka, L. J.; Mosier, K. L.; and Burdick, M. 1999. Does automation bias decision-making? *International Journal of Human-Computer Studies*, 51(5): 991–1006.
- Sorantin, E.; Grasser, M. G.; Hemmelmayr, A.; Tschauner, S.; Hrzic, F.; Weiss, V.; Lacekova, J.; and Holzinger, A. 2022. The augmented radiologist: artificial intelligence in the practice of radiology. *Pediatric radiology*, 52(11): 2074–2086.
- Srinivasan, T.; and Thomason, J. 2026. Adjust for trust: Mitigating trust-induced inappropriate reliance on ai assistance. In *Proceedings of the 31st International Conference on Intelligent User Interfaces*, 1883–1900.
- Swaroop, S.; Buçinca, Z.; Gajos, K. Z.; and Doshi-Velez, F. 2024. Accuracy-Time Tradeoffs in AI-Assisted Decision Making under Time Pressure. In *Proceedings of the 29th International Conference on Intelligent User Interfaces, IUI '24*, 138–154. ACM.
- Swaroop, S.; Buçinca, Z.; Gajos, K. Z.; and Doshi-Velez, F. 2025. Personalising AI Assistance Based on Overreliance Rate in AI-Assisted Decision Making. In *Proceedings of the 30th International Conference on Intelligent User Interfaces, IUI '25*, 1107–1122. ACM.
- Vasconcelos, H.; Jörke, M.; Grunde-McLaughlin, M.; Gerstenberg, T.; Bernstein, M. S.; and Krishna, R. 2023. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1): 1–38.
- Venkatesh, V.; and Bala, H. 2008. Technology acceptance model 3 and a research agenda on interventions. *Decision sciences*, 39(2): 273–315.
- Viduchinsky, N.; Erel, H.; and Morag Yaar, N. 2026. Negotiating Creativity with AI: Effects on Self-Efficacy and Task-Performance. In *Proceedings of the 8th ACM Conference on Conversational User Interfaces*, 1–6.
- Wilson, M. L.; Huggins-Manley, A. C.; Ritzhaupt, A. D.; and Ruggles, K. 2023. Development of the abbreviated technology anxiety scale (ATAS). *Behavior Research Methods*, 55(1): 185–199.