From Error Eradication to Productive Fallibility: A Socio-Technical Perspective on Human–AI Oversight

Federico Cabitza[1,2], Luca Marconi[1,**]

[1]Department of Informatics, Systems and Communication (DISCo), University of Milano-Bicocca, Viale Sarca 336, 20126 Milan, Italy

[2]Digital Health & Wellbeing Center, Fondazione Bruno Kessler (FBK), Trento, Italy

Abstract

Artificial intelligence systems are commonly evaluated through an error-eradication paradigm in which better systems minimize visible failure and support increasingly seamless automation. From a socio-technical perspective, however, the absence of visible error can weaken oversight by concealing the conditions under which AI outputs should be questioned, repaired, or rejected. We introduce *productive fallibility* as a property of human–AI configurations in which residual fallibility becomes detectable, interpretable, actionable, and contestable. We distinguish destructive, hidden, and productive configurations of error and develop a five-level framework spanning cognitive, interactional, organizational, ethical, and learning functions. Illustrative vignettes from healthcare, education, and organizational decision-making show how interfaces, workflows, and governance arrangements can transform otherwise concealed fallibility into opportunities for verification, human repair, escalation, and organizational learning without legitimizing unsafe failure. We conclude with design principles for inspectability, repairability, calibrated reliance, distributed responsibility, human reflection, and learning loops.

Keywords

Human–AI interaction, productive fallibility, socio-technical systems, human oversight, calibrated reliance, contestability, resilience engineering

1. Introduction

Artificial Intelligence (AI) systems are increasingly embedded in socio-technical environments characterized by uncertainty, distributed responsibility, organizational constraints, and continuous human interpretation. In healthcare, education, public administration, and organizational decision-making, AI systems are no longer isolated computational tools; they increasingly participate in processes of classification, recommendation, explanation, prioritization, and judgement [1, 2, 3]. In such contexts, dominant approaches to AI design and evaluation often assume that better systems are those that progressively reduce error, remove friction, and make automated support more fluent, seamless, and authoritative [4, 5, 6].

This assumption is only partially adequate. Error reduction is indispensable, especially in high-stakes domains where incorrect AI outputs may produce harm, discrimination, unsafe action, or institutional irresponsibility. However, the pursuit of increasingly invisible and frictionless automation may also generate socio-technical fragility. When AI outputs appear fluent, confident, institutionally authorized, or practically difficult to contest, users may progressively disengage from verification and rely on the system beyond its actual competence. This is the familiar problem of inappropriate reliance, automation bias, and supervisory complacency [7, 8, 9, 10]. From this perspective, the absence of visible error does not necessarily indicate reliability at the system level. It may instead indicate that fallibility has become harder to detect, question, and govern.

This paper proposes a socio-technical reconceptualization of error in human–AI systems. We do not argue that AI systems should be made less accurate, nor that harmful failures should be tolerated.

The 12th International Conference on Socio-Technical Perspectives in IS (STPIS'26), September 16–17, 2026, L'Aquila, Italy

**Corresponding author.

federico.cabitza@unimib.it (F. Cabitza); luca.marconi@unimib.it (L. Marconi)

0000-0002-4065-3415 (F. Cabitza); 0000-0002-0236-6159 (L. Marconi)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

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

published 2026-11-10

CEUR-WS.org/Vol-4325/paper4.pdf

Rather, we argue that residual fallibility should be made socio-technically available: visible enough to be detected, interpretable enough to be understood, and actionable enough to support repair, escalation, and learning. The central issue is therefore not whether error should be eliminated or accepted, but how different forms of error are produced, represented, interpreted, and acted upon within situated AI-mediated practices.

The paper makes three contributions. First, it reframes residual error in human–AI systems as a socio-technical governance problem, distinguishing destructive, hidden, and productive configurations while identifying visible-but-non-actionable fallibility as an important failure mode of nominal oversight. Second, it develops *productive fallibility* as a property of socio-technical configurations in which residual human or machine fallibility becomes detectable, interpretable, actionable, contestable, repairable, and capable of supporting learning. Third, it translates this account into a five-level framework and design principles spanning inspectability, human repair, calibrated reliance, distributed responsibility, metacognitive reflection, and socio-technical learning loops.

The argument draws on connected traditions in human factors, safety science, HCI, and information systems. Foundational socio-technical work emphasizes that system performance depends on the joint configuration of technical capabilities, human roles, work practices, organizational structures, and opportunities for adaptation [11, 12, 13, 14]. Recent STPIS scholarship has extended this socio-technical perspective explicitly to AI applications and conversational systems, emphasizing the importance of examining AI within concrete work systems and situated organizational contexts rather than as an isolated technological capability [15, 16]. Contemporary IS research similarly treats increasingly agentic AI artifacts as participants in delegation and organizational coordination rather than passive tools [17, 18]. This perspective complements research on automation bias, resilience, seamful interaction, and designed friction [19, 20, 21, 22, 23, 24, 25]. Reliability is therefore treated here not as a property of model outputs alone, but as an achievement of human–machine configurations that must remain inspectable, contestable, and governable.

2. Background: Fallibility, Friction, and Oversight in AI

2.1. From machine infallibility to fallible intelligence

The contemporary tendency to evaluate AI systems primarily through accuracy, performance, and error reduction is rooted in a longstanding cultural imagination of machines as instruments of precision and regularity. In this imagination, the machine is valuable precisely because it appears to operate without hesitation, fatigue, ambiguity, or error. Yet this ideal becomes problematic when AI systems are embedded in domains where interpretation, responsibility, adaptation, and situated judgement are central.

Edgar Allan Poe's analysis of Maelzel's chess-playing automaton offers an early critique of the fascination with apparent machine infallibility [26]. Poe's concern was not only that the automaton was deceptive, but that its apparent autonomy concealed the human and material arrangements that made its performance possible; the relevance here is therefore not the device as an instance of AI, but the socio-technical concealment produced by an appearance of autonomous competence. This insight remains relevant to contemporary AI: systems that appear autonomous, objective, or authoritative are always sustained by human labour, design choices, data practices, maintenance work, institutional framing, and situated interpretation. The appearance of infallibility should therefore become an object of inquiry rather than a reason for epistemic surrender.

Alan Turing provides a complementary perspective. In his discussion of computing machinery and intelligence, intelligence is not equated with perfect correctness [27]. The point is not that error is desirable, but that intelligence need not be equated with perfect correctness: imperfect, unexpected, or contestable answers can remain part of the relational and interpretive process through which intelligence is attributed and tested. Turing and Poe thus help open a conceptual space between two inadequate positions: that trustworthy machines must appear error-free, and that error merely indicates failure. In

human–AI systems, visible fallibility may reveal limits, dependencies, and contextual assumptions that would otherwise remain hidden.

2.2. Automation bias, calibrated reliance, and meaningful oversight

Human factors research has long shown that automation changes how people allocate attention, responsibility, and effort. Automated support can improve performance, but it can also produce misuse, disuse, and abuse when trust is poorly calibrated [7, 20]. Automation bias occurs when users accept automated suggestions despite contradictory evidence, missing information, or contextual reasons for doubt [8, 9, 10]. This risk is particularly acute when outputs are presented with fluent confidence or embedded within institutional workflows that make disagreement costly, exceptional, or undocumented.

For this reason, human oversight cannot be reduced to the formal presence of a human actor after an automated output has been produced. Meaningful control requires that human reasons and relevant moral considerations can shape system behaviour and that responsibility can be traced to actors with an appropriate capacity to intervene [28]. In operational terms, oversight requires time, authority, information, competence, and organizational support. A person who is nominally "in the loop" but lacks the ability to inspect, contest, override, or escalate an AI output is not exercising robust control. This concern is central to critiques of procedural human oversight [29] and to the warning that humans may become "moral crumple zones" when responsibility for system failure is displaced onto operators who had limited control over the automated arrangement [30, 31]. In the European regulatory context, Article 14 of the AI Act likewise requires high-risk AI systems to be designed and developed so that they can be effectively overseen by natural persons during use [32].

2.3. Friction, resilience, and socio-technical reliability

The ideal of seamless interaction is not universally desirable. In many settings, reduced effort, speed, and fluency are genuine design virtues. Yet in high-stakes, ambiguous, or value-laden decisions, excessive seamlessness may suppress the very interruptions through which users recognize uncertainty and exercise judgement. Work on seamful interaction, frictional AI, and designed friction challenges the assumption that all interactional resistance is a defect [23, 24, 25, 33]. Properly placed friction can slow down consequential action, prompt verification, expose uncertainty, and support more appropriate reliance.

Safety science and resilience engineering offer a related lesson. Reliability is not merely the absence of failure; it also depends on the capacity to detect, absorb, respond to, and learn from disturbances [21, 22]. Applied to AI, this means that reliability cannot be assessed only through model-level metrics such as accuracy, precision, recall, or calibration. These metrics remain essential, but they do not capture whether users can detect when the system is wrong, understand why an output may be unsafe or inappropriate, intervene before harm occurs, or ensure that recurrent breakdowns inform redesign. Reliability in human–AI systems is therefore a socio-technical achievement. This argument also has clear antecedents. Bainbridge's classic "ironies of automation" showed that increasing automation can leave human operators responsible for the abnormal conditions for which their skills are least exercised [19]; seamful design argued against treating the disappearance of technological seams as an unconditional design goal and later work framed seamful and seamless design as complementary strategies for downstream user agency [23, 24]; and cognitive forcing interventions have tested deliberate interactional mechanisms for reducing overreliance on AI recommendations [34]. Our contribution is therefore not the claim that friction, breakdown, or visible limitations have never been recognized as informative. It is to integrate these insights into a socio-technical account of how residual human and machine fallibility can be made governable across detection, interpretation, action, repair, and learning.

Table 1
Conceptual distinctions relevant to productive fallibility in human–AI systems.

Concept	Working definition	Relevance for design and oversight
Error	An output, action, or recommendation that is incorrect, unsafe, misleading, or inappropriate relative to a task, criterion, context, or normative standard.	Errors must be prevented when harmful, surfaced when concealed, and made actionable when they reveal limits or mismatches.
Hallucination / unsubstantiated generative output	A fluent and plausible-seeming output that appears correct, comprehensive, or contextually adequate, although it is false, inaccurate, unsupported, irrelevant, or otherwise unfit for the task.	Hallucinations are especially harmful when plausibility masks wrongness; design should surface uncertainty, provenance, evidential grounding, and limits rather than allowing fluency to substitute for reliability.
Uncertainty	A condition of incomplete knowledge concerning the correctness, robustness, scope, or consequences of an output.	Uncertainty should be communicated so that users can adjust reliance and decide when verification is required.
Anomaly	A deviation from expected patterns of system behaviour, user expectation, or domain plausibility.	Anomalies can function as early signals that something requires inspection, even before an error is formally established.
Disagreement	A divergence between AI output, human judgement, alternative evidence, or another system.	Disagreement can prompt comparison, second opinions, justification, or escalation without presupposing that either party is already known to be wrong.
Breakdown	A disruption in the ordinary flow of action, interpretation, or coordination.	Breakdowns expose dependencies and can reveal where workflow, interface, or responsibility structures are fragile.
Friction	Interactional or organizational resistance that slows, interrupts, or complicates action.	Friction is harmful when it adds unnecessary burden, but productive when it preserves attention, verification, and responsibility.

3. Conceptual Clarification: Error, Uncertainty, Anomaly, and Friction

A clearer vocabulary is needed because discussions of AI failure often use *error*, *hallucination*, *uncertainty*, *anomaly*, *breakdown*, *disagreement*, and *friction* as if they were interchangeable. They are related, but analytically distinct, as represented in Table 1. Without this distinction, the claim that error may be productive risks becoming normatively ambiguous or empirically vague.

On this basis, *productive fallibility* does not mean that designers should intentionally introduce arbitrary mistakes. We use the term to denote a property of a socio-technical configuration: residual fallibility becomes productive when interfaces, workflows, organizational arrangements, and governance mechanisms make it sufficiently detectable, interpretable, actionable, and contestable to support verification, repair, escalation, or learning. A *productive error event* is therefore a particular episode occurring within such a configuration, not an error that is beneficial by virtue of being wrong. Designed friction is one possible mechanism for supporting productive fallibility, but the concepts are not synonymous: friction is an intentional intervention in interaction or workflow, whereas fallibility is a residual condition of the human–AI system. The concept also differs from Kapur's *productive failure*, a pedagogical design in which unsuccessful problem-solving can prepare learners for subsequent instruction [35]; productive fallibility is not a teaching method and does not require intentionally induced failure.

4. Destructive, Hidden, and Productive Configurations of Error

The first step in moving beyond an error-eradication paradigm is to distinguish different forms of error. Treating all errors as equivalent obscures the fact that their consequences depend on visibility, interpretability, actionability, and organizational response. We distinguish three principal configurations: destructive error, hidden error, and productive fallibility as instantiated in productive error events.

Destructive error refers to error that produces harm, risk, confusion, exclusion, or unsafe action without adequate safeguards. These errors are the traditional focus of safety, reliability, and risk-management approaches. They must be prevented, reduced, mitigated, and monitored as much as possible. In high-stakes contexts such as healthcare, welfare, education, and organizational decision-making, destructive errors may affect diagnoses, treatments, opportunities, rights, or evaluations.

Hidden error is error that remains concealed by smooth, fluent, authoritative, or procedurally legitimate system behaviour. Hidden error is especially dangerous because it preserves the experience of reliability while reducing opportunities for critique and correction. A natural-language answer may be fluent but unsupported; a risk score may be precise-looking but poorly transported to the local population; a ranking may appear neutral while reproducing historical bias. In such cases, the failure is not merely technical. It is interactionally and institutionally masked. This notion differs from Reason's "latent conditions," which are dormant system-level vulnerabilities that can combine with local circumstances to contribute to failure [21]. Hidden error refers here to an already instantiated erroneous or inappropriate output whose problematic status is concealed by presentation, workflow, or institutional conditions; latent conditions may help produce such concealment, but the concepts operate at different analytical levels.

A *productive error event* is an error that, within a productive-fallibility configuration, becomes visible, interpretable, and actionable. It is not valuable because it is wrong, but because its wrongness becomes available for human and organizational response. Such an event can interrupt passive reliance, reveal boundaries of model competence, expose mismatches between computational representation and situated context, trigger verification, and inform redesign. The productive property therefore belongs primarily to the surrounding socio-technical configuration, not intrinsically to the error.

Figure 1 summarizes three principal configurations. Detectability and actionability are enabling conditions rather than exhaustive binary axes from which the triad is mechanically derived. Importantly, visibility alone is insufficient: fallibility may be visible but non-actionable when users receive warnings yet lack time, authority, usable response mechanisms, or escalation paths. Such cases constitute a failure mode of nominal oversight rather than a fourth mutually exclusive type of error.

This triad also clarifies the normative position of the paper. The problem is not the presence or absence of error in the abstract, but whether fallibility is visible enough to be governed. An AI system with a low error rate may still be socio-technically fragile if its remaining errors are concealed, overtrusted, or difficult to contest. Conversely, an imperfect system may be safer and more accountable if it communicates uncertainty, exposes anomalies, enables repair, and supports escalation.

5. A Multi-Level Framework of Productive Fallibility

Productive fallibility operates across five levels of human–AI systems: cognitive, interactional, organizational, ethical, and learning. These levels capture how fallibility can influence attention and judgement, shape verification and repair, expose responsibilities and dependencies, support appropriate reliance and contestability, and connect local breakdowns to longer-term adaptation.

Table 2 presents the proposed framework. It connects each level with the role of error, the failure mode that emerges when productive fallibility is absent, and design questions that can guide system development and deployment.

At the cognitive level, productive fallibility interrupts automatic acceptance. This is important because reliance on AI is shaped not only by performance, but also by workload, time pressure, professional hierarchy, prior experience, interface cues, and organizational expectations [20, 36]. An implausible,

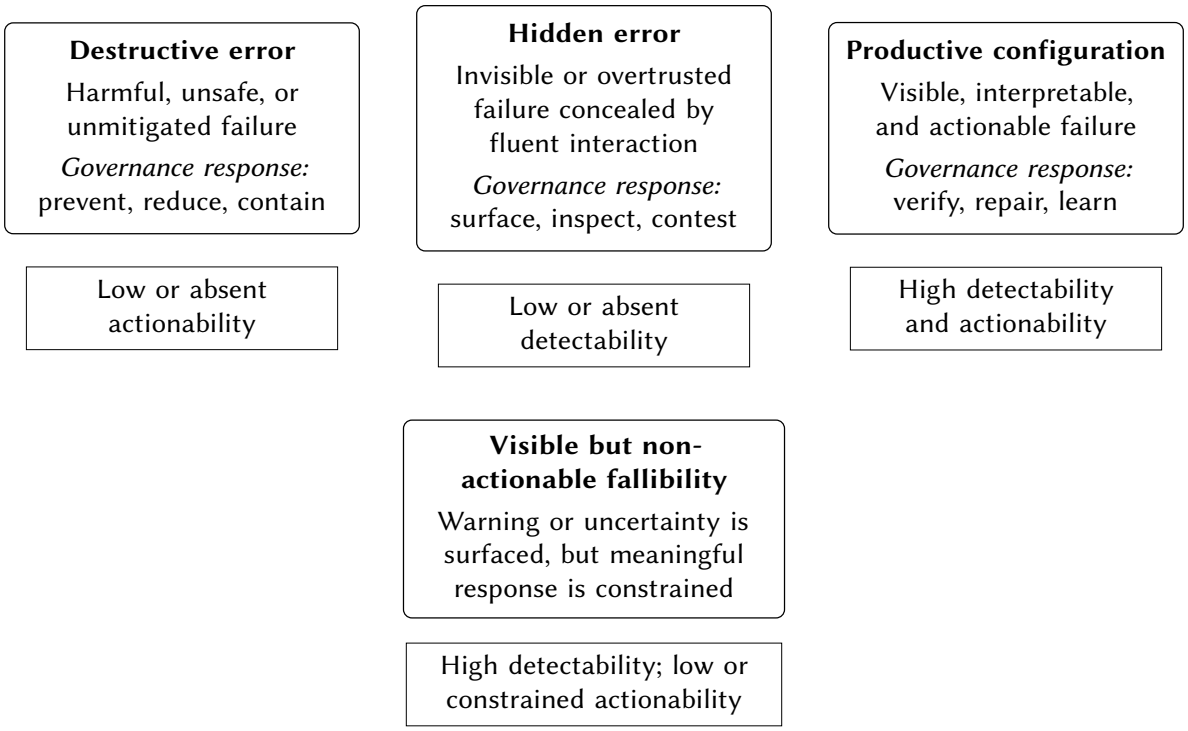


Figure 1: Three principal socio-technical configurations of error. Detectability and actionability are enabling conditions rather than exhaustive binary axes; visible-but-non-actionable fallibility is shown as a failure mode of nominal oversight.

uncertain, or inconsistent output can act as a cognitive alarm, reminding users that the AI system is not an infallible authority but a fallible component of a broader reasoning process.

At the interactional level, fallibility becomes productive only when users can do something with it. Detection without repairability is insufficient. Users must be able to request clarification, compare alternatives, document disagreement, override recommendations, or escalate ambiguous cases. This is particularly important for AI systems that support complex reasoning tasks, where human contribution should not be reduced to final approval after machine recommendation.

At the organizational level, productive fallibility makes visible that AI errors rarely belong only to the model. They may reflect inadequate data practices, unclear deployment procedures, poor training, insufficient monitoring, inappropriate task allocation, or ambiguous responsibility. Treating errors as socio-technical events allows organizations to examine the configuration of people, models, interfaces, policies, and work practices that produced the failure.

At the ethical level, productive fallibility supports calibrated reliance. By calibrated reliance, we mean reliance that is proportionate to the system's actual capabilities and limitations in the current context, rather than globally high or low trust. Trustworthy human–AI collaboration does not require users to trust AI systems as much as possible. It requires them to rely appropriately, in proportion to the system's capabilities, limitations, and context of use [20, 37]. Visible fallibility can help users form more realistic mental models of what the system can and cannot do.

Finally, at the learning level, productive fallibility connects local breakdowns to broader adaptation. A single error may prompt correction; recurrent patterns of error should prompt redesign. This requires feedback loops that treat errors not only as performance data, but as evidence about the fit between system, user, task, organization, and governance arrangement.

Table 2
A multi-level framework of productive fallibility in human–AI systems.

Level	Productive role of fallibility	Failure mode if absent	Design question
Cognitive	Interrupts automatic acceptance and reactivates attention, doubt, and reflective judgement.	Users accept outputs passively, especially under workload, time pressure, or perceived system authority.	Can users notice uncertainty, anomaly, disagreement, or low confidence at the point of decision?
Interactional	Creates opportunities for questioning, comparison, correction, override, and repair.	The interface permits only acceptance or rejection, with little support for interpretation or correction.	Can users inspect, challenge, revise, annotate, or escalate the output?
Organizational	Makes dependencies, procedures, responsibilities, and monitoring needs visible.	Responsibility is displaced onto frontline users while systemic causes remain unexamined.	Are escalation paths, audit trails, review procedures, and accountability structures defined?
Ethical	Challenges the illusion of infallibility and supports calibrated reliance, epistemic humility, and contestability.	AI outputs acquire inappropriate epistemic or institutional authority.	Does the system avoid overconfident presentation and preserve the user's right and capacity to contest?
Learning	Transforms recurrent breakdowns into evidence for model improvement, workflow redesign, training, and governance adaptation.	The same failures recur without structural correction or organizational learning.	Are individual errors connected to feedback loops for technical and socio-organizational improvement?

6. Illustrative Socio-Technical Vignettes

The following vignettes are analytic illustrations rather than empirical case studies. They were selected as contrasting contexts in which productive fallibility performs different primary functions: preserving professional judgement in healthcare, sustaining epistemic and learner agency in education, and enabling institutional contestability in organizational decision-making. The three contexts are therefore illustrative rather than mutually exclusive or ontologically equivalent categories; they are used to test the portability of the same socio-technical concepts across different configurations of users, tasks, authority, and consequences.

6.1. Healthcare AI: error as a signal for professional judgement

Consider an AI-based diagnostic support system used in an emergency department to flag radiological examinations that may require urgent review. A destructive error would occur if the system missed a critical finding and the workflow allowed the case to proceed without human review. A hidden-error configuration would occur if the system produced an apparently precise score without communicating that the case differs substantially from the population on which the model was validated [38]. A productive-fallibility configuration would alter the surrounding interaction: an out-of-distribution indicator, provenance cue, calibrated uncertainty signal, or mismatch warning would make the limitation visible at the point of use, giving the clinician a reason to compare the recommendation with imaging findings, symptoms, and history and, where warranted, activate a second-opinion or escalation procedure. The error does not spontaneously become productive; the socio-technical configuration provides mechanisms that transform otherwise concealed fallibility into an interpretable and actionable signal.

The difference lies not only in the model output, but in the socio-technical configuration surrounding it. Healthcare AI can support clinical work only if human actors retain meaningful capacity to interpret and contextualize system outputs and integrate them into situated judgement [39, 40]. In our framework,

Table 3
Illustrative domains in which productive fallibility can support socio-technical resilience.

Domain	Risk of hidden error	Productive socio-technical function
Healthcare AI	Overconfident decision-support outputs treated as authoritative recommendations.	Reactivates clinical reasoning, contextual verification, and accountable professional judgement.
Education and AI tutors	Fluent but unsupported or misleading explanations accepted as knowledge.	Prompts source verification, metacognition, argument reconstruction, and learner agency.
Organizational decision-making	Algorithmic rankings or classifications absorbed as procedurally neutral outputs.	Exposes bias, hidden assumptions, accountability gaps, and governance limitations.

this requirement extends to the possibility of questioning and contesting recommendations when their applicability to the case is uncertain. The relevant design problem is therefore not simply to maximize apparent precision, but to ensure that probabilistic outputs do not become overconfident substitutes for situated clinical judgement. Productive fallibility contributes to accountable care when it makes uncertainty, mismatch, and disagreement available for professional reasoning.

6.2. Education and AI tutors: error as an occasion for reflection

Educational AI systems and generative tutors increasingly provide explanations, examples, feedback, and writing support. Their fluency creates a pedagogical risk: students may confuse linguistic coherence with epistemic reliability. A hidden error occurs when a plausible answer is accepted because it sounds correct, even though it is incomplete, unsupported, or conceptually misleading. In such cases, the system may weaken learning by encouraging shortcut-based interaction and outsourcing of interpretive effort.

The same fallibility can become productive when it is pedagogically framed. A flawed explanation may invite students to identify missing assumptions, verify sources, reconstruct an argument, compare alternative accounts, or explain why a response is inadequate. The educational value does not lie in the error itself, but in the reflective activity that the error makes possible. This aligns with views of learning as active self-regulation, co-regulation, and shared regulation rather than passive reception of correct answers [41, 42]. For AI tutors, the design question is therefore whether the system merely delivers fluent responses or structures epistemic effort, verification, and learner agency.

6.3. Organizational decision-making: error as resistance to opaque automation

AI systems are increasingly used to support screening, prioritization, risk assessment, resource allocation, and managerial judgement. In these settings, outputs can acquire institutional authority because they are embedded in dashboards, workflows, performance systems, or administrative procedures. Hidden error occurs when a ranking or classification is absorbed into routine practice as if it were neutral, objective, and self-justifying.

A productive error event may appear as an anomalous recommendation, an implausible ranking, or a visibly unfair classification. If the organization provides channels for contestation, review, and explanation, the anomaly can trigger examination of data quality, model assumptions, thresholds, accountability structures, and governance procedures. If those channels are absent, the same error may be individualized: a human operator is blamed for not detecting a problem that the organization made practically difficult to detect or challenge. This is precisely the condition in which nominal oversight can turn the operator into a "moral crumple zone," concentrating responsibility on a person who lacked meaningful control over the broader configuration [30, 31]. Productive fallibility therefore requires organizations to treat errors not as isolated human or technical failures, but as opportunities to examine how responsibility is distributed across people, models, interfaces, and institutions.

Across these domains, productive fallibility performs a common function: it prevents AI from becoming an unquestioned layer of mediation. It sustains professional judgement in healthcare, reflective learning in education, and contestability in organizations. These effects do not arise automatically from error. They require interfaces, workflows, policies, and cultures that make fallibility visible, interpretable, and actionable.

7. Implications for Design

If productive fallibility depends on socio-technical conditions, then design should not focus exclusively on reducing error rates. It should also shape how errors, uncertainty, disagreement, and breakdowns are represented and handled. The following principles translate the framework into design implications.

7.1. Design for inspectable error

AI systems should be designed so that fallibility is inspectable at the point of use. Users need practical cues concerning uncertainty, confidence, provenance, scope conditions, alternative interpretations, known limitations, and the relation between the present case and past cases through which the system has learned to operate [43, 38]. In high-stakes domains, this may include confidence intervals, calibration information, out-of-distribution warnings, links to supporting evidence, or explanations that clarify why an output may not apply to the present case [44, 45]. The goal is not to overwhelm users with technical detail, but to make the output sufficiently contestable and situated to support judgement.

One concrete way to implement this principle is to replace, or at least complement, categorical AI advice with example-based cues. In medical decision making, *pro-hoc explanations* have been proposed as a form of frictional AI in which users are shown analogous cases for each possible outcome, rather than a single post-hoc explanation of the machine's recommendation [46]. A particularly relevant variant consists in presenting cases similar to the one at hand for which the system issued the same class prediction but was later found to be wrong. Evidence from radiological fracture detection suggests that making previous machine errors visible in this way may support case-specific trust calibration, help users interpret advice in light of the present case, reduce some diagnostic errors, and be perceived as useful, while still requiring clinicians to exercise their own interpretative skills [46].

Inspectable error is particularly important for generative AI systems, whose linguistic fluency can conceal factual inaccuracy, unsupported inference, or contextual mismatch. Design should therefore resist the aesthetics of effortless certainty when certainty is not warranted. A system that produces fluent but ungrounded answers may appear more reliable than it is.

7.2. Design for human repair

Detection is insufficient if users cannot act. AI systems should support human repair through correction, annotation, override, escalation, comparison, and documentation of disagreement. In professional settings, repairability may require second-opinion workflows, audit trails, incident reporting, and clear procedures for recurrent failures. These mechanisms prevent human oversight from becoming symbolic. A user who is formally responsible but lacks the tools, authority, or time needed to repair an AI-mediated process is not exercising meaningful control.

Repairability also implies reversibility where possible. When AI-supported decisions can be reviewed and corrected before irreversible consequences occur, the overall system becomes more resilient. Conversely, when outputs are automatically translated into action without review, residual error becomes harder to detect and contain.

7.3. Design for calibrated reliance

AI systems should be designed for calibrated reliance rather than maximum trust. This requires moving beyond the treatment of accuracy as an intrinsic property of the machine and considering

trustworthiness as a situated relation between the system, the user, the case, and the decision context [43]. Reliance is productive only when it is proportionate to the system's actual capabilities, limitations, and context of use [20]. Users should understand not only what the system can do, but also what it cannot do, where performance may degrade, and when human verification remains necessary.

Visible and interpretable errors can help users develop realistic mental models of the system. Hidden errors produce misplaced confidence; unexplained errors produce abandonment or indiscriminate distrust. The design objective is therefore neither blind reliance nor generalized scepticism, but situated and revisable reliance.

7.4. Design for distributed responsibility

Distributed responsibility should not mean diffuse responsibility. Human–AI systems should make explicit which actors are responsible for model development, deployment, monitoring, case-level interpretation, override, escalation, and organizational remediation. Interfaces should distinguish advisory from determinative outputs and make visible when human review is mandatory. Meaningful oversight therefore requires not merely assigning a human to the workflow, but aligning responsibility with actual information, authority, time, competence, and capacity to intervene [28, 47, 17].

This principle is essential because accountability cannot be reduced to individual vigilance. If an organization deploys an AI system without adequate training, documentation, monitoring, or escalation channels, errors cannot fairly be attributed only to frontline users. Productive fallibility requires an institutional environment in which breakdowns can be reported and analysed without merely assigning blame. This organizational emphasis is consistent with recent STPIS work conceptualizing AI governance as a reflexive, multi-layered, and adaptive process embedded in organizational context and supported by continuous monitoring and contextual awareness [48].

7.5. Design for human reflection

A further design implication concerns use cases in which the AI system is not primarily trained to predict the correct answer to the task, but to estimate the probability that the human decision on the present case is wrong. In this configuration, the machine does not function as a direct substitute for human judgement. Rather, it operates as a metacognitive support layer that models the risk of human error by combining information about the case, the decision-maker, and the epistemic conditions under which the decision is being made. This orientation is consistent with the view that the quality of AI-supported decision making cannot be inferred from model accuracy alone, because decision quality emerges from the situated relation between model outputs, human interpretation, case characteristics, and patterns of reliance [43]. Such systems are especially relevant when the central risk is not only that the AI may be wrong, but that the human may be overconfident, insufficiently attentive, or unaware that a case falls outside the range of their reliable competence.

The design of this class of systems should avoid treating human error as an individual defect inferred from a single decision. A robust estimate should integrate multiple signals: the intrinsic features of the case; the user's previous accuracy on comparable test or validation cases, as a proxy for task-specific skill; the user's confidence in the current decision; the complexity or ambiguity perceived by the user; the degree of disagreement among raters during ground-truth construction, when available; and the similarity of the current case to cases that were previously found difficult, ambiguous, or error-prone. In this sense, the relevant output is not simply a prediction that "the human is wrong". It is a situated estimate of epistemic risk, grounded in the relation between a particular user, a particular case, and a particular decision context.

This design orientation extends productive fallibility in two ways. First, it makes human fallibility itself inspectable. The system does not merely expose the possible fallibility of the machine; it also helps detect situations in which the human decision may require reconsideration, second reading, additional evidence, or escalation. Second, it supports calibrated reliance in both directions: users should not over-rely on the AI, but they should also avoid over-relying on their own judgement when the available

evidence suggests that the case is unusually difficult or that their confidence is poorly aligned with past performance. The interface should therefore present human-error risk as a reason for reflective verification, not as a verdict that displaces responsibility from the user to the system.

As discussed above, a related design strategy is to present counterfactual analogous cases that make the possible fallibility of the user's initial judgement salient. Pro-hoc explanations operationalize this strategy by showing similar cases in which the system supported the user's tentative judgement but was later found to be wrong, together with cases associated with the alternative outcome. They thereby introduce a lightweight form of cognitive forcing without replacing the clinician's final decision [46, 34]. Prior evidence indicates that such interventions can reduce inappropriate reliance in some conditions but also impose additional effort; cognitive forcing should therefore be treated as a context-dependent design mechanism rather than an inherently beneficial intervention [34].

Several safeguards are necessary. First, a second-order calibration problem arises: a model estimating the probability that a human judgement is wrong is itself fallible. Its estimates therefore require independent validation, calibration assessment, domain-specific monitoring, and mechanisms through which alerts can themselves be inspected and contested. The meta-model should allocate attention to vulnerable cases rather than acquire a new form of epistemic authority. Second, self-reported confidence is a noisy and potentially reactive signal: once users understand how it affects system behaviour, reporting may change strategically or unintentionally. Confidence should therefore be combined with other evidence and evaluated for calibration and measurement reactivity. Third, alerts should be thresholded carefully to avoid alarm fatigue and differentiated by consequence, from self-review to second opinion or formal escalation [49]. Finally, user-specific performance histories and confidence profiles create governance risks. They may support situated assistance but also enable worker monitoring, ranking, or managerial surveillance. Purpose limitation, data minimization, access controls, and separation between decision support and personnel evaluation are therefore important safeguards. In the European context, this concern is especially salient because Annex III, point 4 of the AI Act identifies specified employment and worker-management uses as high-risk, while Article 14 establishes human-oversight requirements for high-risk AI systems [32].

Designed in this way, human-error estimation can become a powerful form of productive fallibility. It turns possible human mistakes into visible and actionable occasions for verification, while preserving the human as the responsible agent who interprets the alert, considers the case context, and decides whether further action is warranted. The objective is not to replace human judgement with a second-order algorithmic judgement about the human. It is to create a socio-technical mechanism that makes vulnerable moments in human reasoning more recognizable, discussable, and repairable. In this respect, human-error estimation should be presented less as an oracle about the correctness of the user and more as a perceptual and semiotic aid that helps users inspect the present case in the light of prior cases, uncertainty, and situated evidence [43].

7.6. Design for learning loops

Finally, AI systems should transform errors into socio-technical learning. Recurrent failures may reveal inadequate training data, poor interface design, unclear task allocation, unrealistic expectations, miscalibrated thresholds, or misalignment between system behaviour and institutional goals. Learning loops should therefore include technical, socio-organizational, and human-learning dimensions. Recent STPIS work on GenAI integration similarly emphasizes socio-technical reflection aimed at balancing technological benefits with wellbeing, autonomy, ethical practice, learning, and the maintenance of human skills [50]. Human learning should not be assumed to follow automatically from repeated exposure to AI-supported decisions. Internalization may support expertise when interaction requires active comparison, explanation, verification, and repair; the same process may instead produce cognitive convergence, biased heuristic transfer, or deskilling when fluent outputs are repeatedly accepted without reflective engagement [51]. The learning loop is therefore bidirectional: it can become a productive loop of reflection and capability development or a degenerative loop of dependence, bias transmission, and progressive loss of independent judgement.

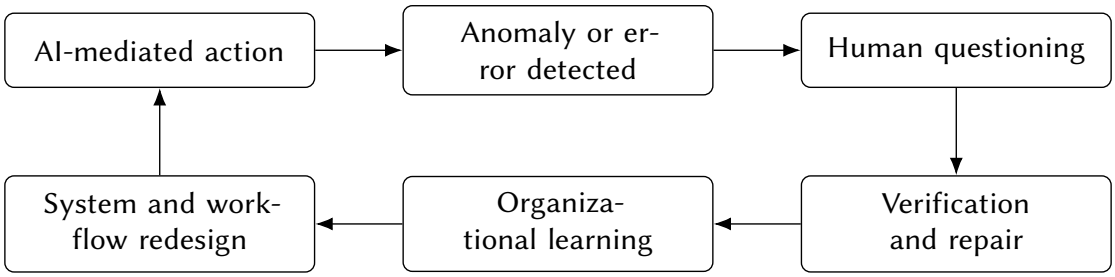


Figure 2: Productive fallibility as a socio-technical learning loop. Visible anomalies and errors can activate questioning, verification, repair, organizational learning, and redesign.

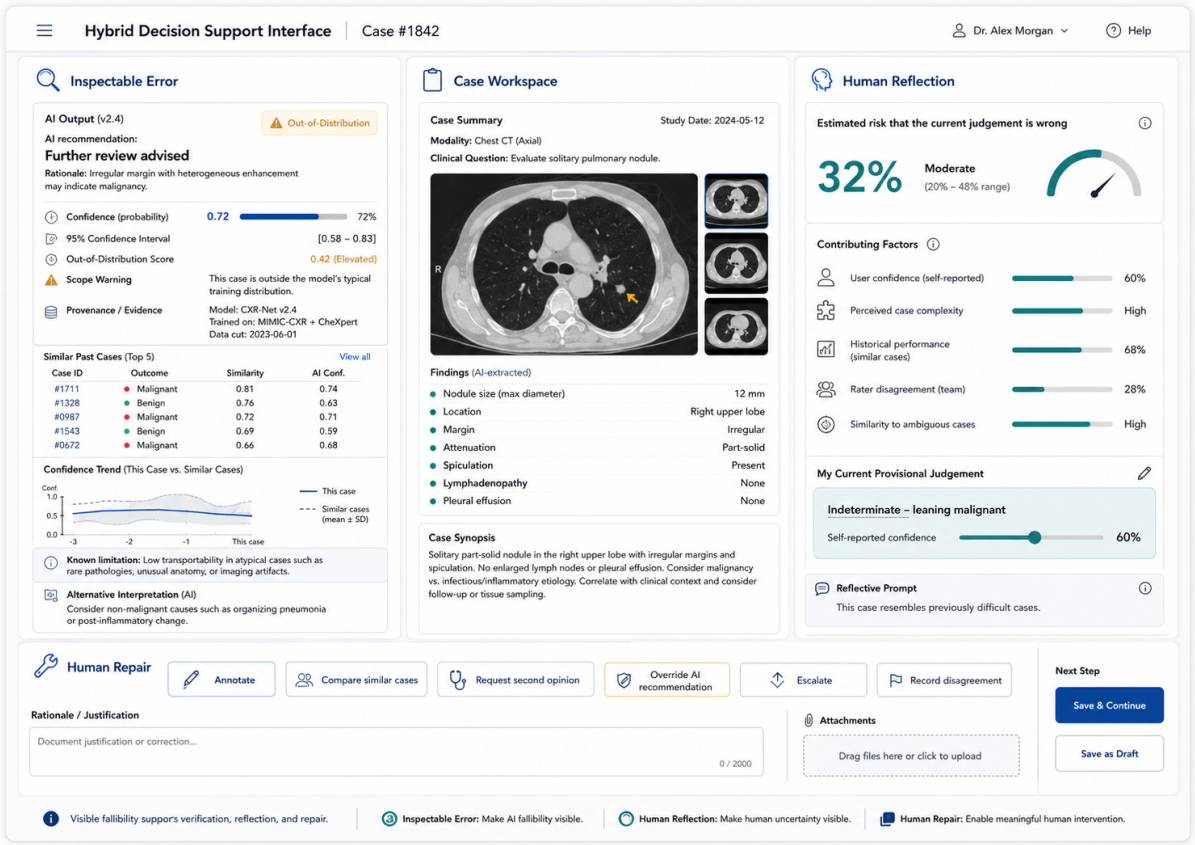


Figure 3: Illustrative hybrid decision-support interface implementing inspectable fallibility, human reflection, and human repair. Model uncertainty, out-of-distribution status, provenance, and analogous cases are surfaced alongside a contestable estimate of human-error risk; repair actions include annotation, comparison, second opinion, override, escalation, and disagreement recording. The mock-up is illustrative rather than a validated clinical interface.

Figure 2 summarizes productive fallibility as a learning loop. Users should be able to report and interpret errors; organizations should analyse patterns; designers should revise interfaces and workflows; and governance structures should update rules, responsibilities, and safeguards. Continuous monitoring should not be understood merely as post-deployment surveillance of model performance, but as ongoing adjustment of the socio-technical system. It should also include attention to what users learn from repeated exposure to AI outputs, since AI-mediated interaction may progressively shape professional judgement, reliance habits, and the transmission of both expertise and bias [51].

To make these implications more concrete, Figure 3 presents an illustrative mock-up of a hybrid decision-support interface that integrates inspectable error, human reflection, and human repair within a single interaction environment.

8. Discussion: From Error Eradication to Governable Fallibility

The argument developed in this paper challenges a dominant assumption in AI design and evaluation: that the primary objective is to eliminate visible error as completely as possible. This assumption is understandable, but incomplete. In human–AI systems, error is not only a technical deviation from correctness. It is also an event that can reorganize attention, redistribute responsibility, expose assumptions, and activate repair.

The distinction between destructive, hidden, and productive configurations reframes the normative goal of human-centered AI. The aim should not be an abstract pursuit of frictionless automation, but the design of systems in which destructive errors are minimized, hidden errors are surfaced, visible-but-non-actionable fallibility is made meaningfully actionable, and productive error events can support oversight and learning. This perspective does not lower reliability standards. Rather, it broadens the meaning of reliability to include detectability, contestability, repairability, and recoverability.

This has implications for evaluation. Standard metrics such as accuracy, precision, recall, calibration, and error rate remain necessary, but they are not sufficient. They do not indicate whether users can recognize when the system is wrong, understand uncertainty, contest outputs, intervene before harm occurs, or ensure that recurrent breakdowns inform redesign. A socio-technical evaluation of AI systems should therefore include measures of reliance behaviour, error detectability, quality of human intervention, cognitive workload, escalation effectiveness, documentation practices, and organizational learning.

The framework also complicates the aspiration toward seamless AI. Seamlessness can be valuable for routine, low-risk, reversible tasks. However, in consequential, ambiguous, or value-laden decisions, some forms of friction are necessary to preserve attention, accountability, and interpretive agency. The relevant design question is not whether interaction should be smooth or effortful in general, but where friction should be removed, where it should be preserved, and how it should be made meaningful rather than merely burdensome. Operationalizing these principles also creates implementation and governance requirements. Inspectable and repairable AI may require reliable provenance, traceable records of system outputs and human interventions, appropriate retention and access policies, and documentation of overrides, escalation, and recurrent failure patterns. These mechanisms must be proportionate to domain risk and aligned with applicable data-protection, safety, accountability, and AI-governance requirements; for high-risk systems, Articles 12–14 of the EU AI Act provide relevant requirements on record-keeping, transparency and information to deployers, and human oversight [32].

8.1. Limitations and future research

This paper is conceptual. It does not provide empirical evidence that productive fallibility improves performance, learning, safety, or accountability in specific domains. The vignettes are illustrative rather than evidential. Future work should operationalize the framework and test its claims through controlled experiments, field studies, and longitudinal evaluations.

One complementary direction is to investigate how system- or provider-level evidence about content reliability can be integrated with situated measures of productive fallibility. Indicators such as citation correctness, source quality, provenance, evidential support, and appropriately designed user evaluations could provide useful evidence about the reliability of AI outputs and support calibrated reliance and deployment decisions. Such aggregate indicators, however, should complement rather than substitute for interaction- and configuration-level evaluation: a system that performs reliably on average may still produce consequential errors in particular cases, and aggregate reliability does not by itself determine whether residual fallibility is detectable, interpretable, actionable, contestable, or repairable in use. Future empirical work should therefore examine how these levels of evidence can be combined without turning reliability scores themselves into proxies for warranted trust.

Several research directions follow. First, productive fallibility requires metrics for *detectability*: whether users can notice that an output may be wrong, uncertain, or out of scope. Signal detection theory could help distinguish sensitivity to genuinely problematic outputs from indiscriminate warning

responses, separating detection performance from response bias [52]. Second, it requires metrics for *contestability*: whether users can challenge outputs and whether such challenges are procedurally consequential. Third, it requires metrics for *repairability* and *recoverability*: whether users and organizations can correct errors before harm occurs and learn from recurrent breakdowns. Fourth, it requires empirical investigation of trade-offs. Friction can support vigilance, but it can also increase workload, slow action, and produce alert fatigue. The challenge is to identify the conditions under which friction is productive rather than merely obstructive. Human-error estimation further requires research on second-order calibration, strategic or reactive confidence reporting, privacy and workplace surveillance, and whether metacognitive alerts improve joint decision quality rather than merely shifting authority to another algorithm.

Finally, the framework should be adapted to domain-specific risk profiles. Productive fallibility in education cannot be governed in the same way as productive fallibility in emergency medicine or public administration. The visibility of error, the acceptability of delay, the reversibility of decisions, the authority of the user, and the consequences of failure vary substantially across domains. A mature research agenda should therefore connect the general framework proposed here with situated studies of practice.

9. Conclusion

This paper has argued that error in human–AI systems should not be understood only as a technical defect to be eliminated. Error can also be a socio-technical phenomenon that shapes how humans interpret, question, trust, repair, and govern AI-mediated processes. The central contribution is the distinction between destructive, hidden, and productive socio-technical configurations of error. Destructive errors should be prevented and mitigated; hidden errors should be surfaced and made contestable; visible-but-non-actionable fallibility should be coupled to meaningful response mechanisms; and productive error events can support verification, repair, learning, and accountability.

The broader claim is that reliability in AI systems is not merely a property of model performance. It is a socio-technical achievement sustained by visibility, contestability, repairability, calibrated reliance, and organizational learning. Human-centered AI should therefore not aim only to make automation more seamless. It should also design the conditions under which residual fallibility can be detected, interpreted, contested, and responsibly addressed. In human–AI systems, the absence of visible error may not indicate reliability; it may indicate the disappearance of opportunities for critique.

Declaration on Generative AI

During the preparation of this work, the authors used OpenAI ChatGPT for language editing, structural revision, and argumentative refinement. The tool was used to improve clarity, readability, and organization of author-generated content, and to support critical review and refinement of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, E. Vayena, AI4People—an ethical framework for a good ai society: Opportunities, risks, principles, and recommendations, Minds and Machines 28 (2018) 689–707. doi:10.1007/s11023-018-9482-5.

[2] B. Shneiderman, Human-Centered AI, Oxford University Press, 2022.

[3] S. Passi, S. Barocas, Problem formulation and fairness, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, ACM, 2019, pp. 39–48. doi:10.1145/3287560.3287567.

[4] S. Russell, Human Compatible: Artificial Intelligence and the Problem of Control, Penguin Publishing Group, 2019.
[5] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, D. Mané, Concrete problems in ai safety, 2016. doi:10.48550/arXiv.1606.06565. arXiv:1606.06565.
[6] U. Agudo, K. G. Liberal, M. Arrese, et al., The impact of ai errors in a human-in-the-loop process, Cognitive Research: Principles and Implications 9 (2024) 1. doi:10.1186/s41235-023-00529-3.
[7] R. Parasuraman, V. Riley, Humans and automation: Use, misuse, disuse, abuse, Human Factors 39 (1997) 230–253. doi:10.1518/001872097778543886.
[8] K. L. Mosier, L. J. Skitka, S. Heers, M. Burdick, Automation bias: Decision making and performance in high-tech cockpits, The International Journal of Aviation Psychology 8 (1998) 47–63. doi:10.1207/s15327108ijap0801_3.
[9] L. J. Skitka, K. L. Mosier, M. Burdick, Does automation bias decision-making?, International Journal of Human-Computer Studies 51 (1999) 991–1006. URL: https://doi.org/10.1006/ijhc.1999.0252. doi:10.1006/ijhc.1999.0252.
[10] L. J. Skitka, K. L. Mosier, M. Burdick, Accountability and automation bias, International Journal of Human-Computer Studies 52 (2000) 701–717. URL: https://doi.org/10.1006/ijhc.1999.0349. doi:10.1006/ijhc.1999.0349.
[11] E. Mumford, Socio-technical design: An unfulfilled promise or a future opportunity?, in: R. Baskerville, J. Stage, J. I. DeGross (Eds.), Organizational and Social Perspectives on Information Technology, volume 41 of *IFIP Advances in Information and Communication Technology*, Springer, Boston, MA, 2000, pp. 33–46. doi:10.1007/978-0-387-35505-4_3.
[12] G. Baxter, I. Sommerville, Socio-technical systems: From design methods to systems engineering, Interacting with Computers 23 (2011) 4–17. doi:10.1016/j.intcom.2010.07.003.
[13] P. Checkland, S. Holwell, Information, Systems and Information Systems: Making Sense of the Field, John Wiley & Sons, Chichester, 1998.
[14] S. Alter, Work system theory: Overview of core concepts, extensions, and challenges for the future, Journal of the Association for Information Systems 14 (2013) 72–121. doi:10.17705/1jais.00323.
[15] S. Alter, Applications of AI in sociotechnical work systems: Fundamental change or just a type of partial automation?, in: Proceedings of the 10th International Conference on Socio-Technical Perspectives in Information Systems (STPIS 2024), volume 3857 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 12–21. URL: https://ceur-ws.org/Vol-3857/paper1.pdf.
[16] L. Gkinko, A. Elbanna, AI chatbots sociotechnical research: An overview and future directions, in: Proceedings of the 8th International Workshop on Socio-Technical Perspective in Information Systems Development (STPIS 2022), volume 3239 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 173–183. URL: https://ceur-ws.org/Vol-3239/paper17.pdf.
[17] A. Baird, L. M. Maruping, The next generation of research on is use: A theoretical framework of delegation to and from agentic is artifacts, MIS Quarterly 45 (2021) 315–341. doi:10.25300/MISQ/2021/15882.
[18] N. Berente, B. Gu, J. Recker, R. Santhanam, Managing artificial intelligence, MIS Quarterly 45 (2021) 1433–1450. doi:10.25300/MISQ/2021/16274.
[19] L. Bainbridge, Ironies of automation, Automatica 19 (1983) 775–779. doi:10.1016/0005-1098(83)90046-8.
[20] J. D. Lee, K. A. See, Trust in automation: Designing for appropriate reliance, Human Factors 46 (2004) 50–80. doi:10.1518/hfes.46.1.50_30392.
[21] J. Reason, Human Error, Cambridge University Press, Cambridge, 1990. URL: https://www.cambridge.org/highereducation/books/human-error/281486994DE4704203A514F7B7D826C0.
[22] E. Hollnagel, Safety-I and Safety-II: The Past and Future of Safety Management, Ashgate, Farnham, 2014. URL: https://erikhollnagel.com/books/safety-i-and-safety-ii.html.
[23] M. Chalmers, A. Galani, Seamful interweaving: Heterogeneity in the theory and design of interactive systems, in: Proceedings of the 5th Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques, ACM, 2004, pp. 243–252. doi:10.1145/1013115.1013149.

[24] S. Inman, D. Ribes, "Beautiful Seams": Strategic Revelations and Concealments, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA, 2019, pp. 1–14. doi:10.1145/3290605.3300508.
[25] C. Natali, Frictional ai: Designing desirable inefficiencies in decision support systems for knowledge work, in: Proceedings of the 22nd European Conference on Computer-Supported Cooperative Work: Doctoral Colloquium Contributions, European Society for Socially Embedded Technologies, 2024. doi:10.48340/ecscw2024_dc07.
[26] E. A. Poe, Maelzel's chess-player, Southern Literary Messenger, 1836.
[27] A. M. Turing, Computing machinery and intelligence, Mind 59 (1950) 433–460. doi:10.1093/mind/LIX.236.433.
[28] F. Santoni de Sio, J. van den Hoven, Meaningful human control over autonomous systems: A philosophical account, Frontiers in Robotics and AI 5 (2018) 15. doi:10.3389/frobt.2018.00015.
[29] B. Green, The flaws of policies requiring human oversight of government algorithms, Computer Law & Security Review 45 (2022) 105681. doi:10.1016/j.clsr.2022.105681.
[30] M. C. Elish, Moral crumple zones: Cautionary tales in human–robot interaction, Engaging Science, Technology, and Society 5 (2019) 40–60. doi:10.17351/ests2019.260.
[31] M. C. Elish, Moral crumple zones: Cautionary tales in human–robot interaction, in: R. Calo, A. M. Froomkin, K. Thomasen (Eds.), Robot Law: Volume II, Edward Elgar Publishing, Cheltenham, UK, 2025, pp. 83–105. doi:10.4337/9781800887305.00010.
[32] European Parliament and Council of the European Union, Regulation (eu) 2024/1689 of 13 june 2024 laying down harmonised rules on artificial intelligence (artificial intelligence act), Official Journal of the European Union, L 2024/1689, 12 July 2024, 2024. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj.
[33] A. La Rosa, A. Beretta, Frictional ai in joint cognitive systems: Towards a human-centered approach at higher levels, in: Proceedings of the Workshops at the Fourth International Conference on Hybrid Human-Artificial Intelligence (HHAI 2025), volume 4074 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025, pp. 90–97. URL: https://ceur-ws.org/Vol-4074/short3-1.pdf.
[34] Z. Buçinca, M. B. Malaya, K. Z. Gajos, To trust or to think: Cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making, Proceedings of the ACM on Human-Computer Interaction 5 (2021) 1–21. doi:10.1145/3449287.
[35] M. Kapur, Productive failure, Cognition and Instruction 26 (2008) 379–424. doi:10.1080/07370000802212669.
[36] C. D. Wickens, J. G. Hollands, S. Banbury, R. Parasuraman, Engineering Psychology and Human Performance, 4 ed., Psychology Press, 2015.
[37] B. Shneiderman, Human-centered artificial intelligence: Reliable, safe & trustworthy, International Journal of Human–Computer Interaction 36 (2020) 495–504. URL: https://doi.org/10.1080/10447318.2020.1741118. doi:10.1080/10447318.2020.1741118.
[38] L. Marconi, F. Cabitza, Show and tell: A critical review on robustness and uncertainty for a more responsible medical ai, International Journal of Medical Informatics 202 (2025) 105970. doi:10.1016/j.ijmedinf.2025.105970.
[39] A. Bussone, S. Stumpf, D. O'Sullivan, The role of explanations on trust and reliance in clinical decision support systems, in: 2015 International Conference on Healthcare Informatics, 2015, pp. 160–169. doi:10.1109/ICHI.2015.26.
[40] A. S. Islind, H. V. Hult, Balancing overreliance and mistrust in data-driven decision making: A critical view on the role of quantified self in diabetes management, in: Proceedings of the 8th International Workshop on Socio-Technical Perspective in Information Systems Development (STPIS 2022), volume 3239 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 74–84. URL: https://ceur-ws.org/Vol-3239/paper8.pdf.
[41] E. Kasneci, K. Sessler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, S. Krusche, G. Kutyniok, T. Michaeli, C. Nerdel, J. Pfeffer, O. Poquet, M. Sailer, A. Schmidt, T. Seidel, M. Stadler, J. Weller, J. Kuhn, G. Kasneci, Chatgpt for good? on opportunities and challenges of large language models for education, Learning and Individual

Differences 103 (2023) 102274. URL: https://doi.org/10.1016/j.lindif.2023.102274. doi:10.1016/j.lindif.2023.102274.

[42] A. F. Hadwin, S. Järvelä, M. Miller, Self-regulation, co-regulation, and shared regulation in collaborative learning environments, Educational Psychologist 53 (2018) 242–264. URL: https://doi.org/10.1080/00461520.2018.1471249. doi:10.1080/00461520.2018.1471249.

[43] F. Cabitza, A. Campagner, E. Datteri, To err is (only) human. reflections on how to move from accuracy to trust for medical ai, in: F. Ceci, A. Prencipe, P. Spagnoletti (Eds.), Exploring Innovation in a Digital World: Cultural and Organizational Challenges, volume 51 of *Lecture Notes in Information Systems and Organisation*, Springer, Cham, 2021, pp. 36–49. doi:10.1007/978-3-030-87842-9_4.

[44] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, J. Teevan, R. Kikin-Gil, E. Horvitz, Guidelines for human-ai interaction, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 1–13. URL: https://doi.org/10.1145/3290605.3300233. doi:10.1145/3290605.3300233.

[45] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the gdpr, Harvard Journal of Law & Technology 31 (2018) 841–887. URL: https://doi.org/10.2139/ssrn.3063289. doi:10.2139/ssrn.3063289.

[46] F. Cabitza, C. Natali, L. Famiglini, A. Campagner, V. Caccavella, E. Gallazzi, Never tell me the odds: Investigating pro-hoc explanations in medical decision making, Artificial Intelligence in Medicine 150 (2024) 102819. doi:10.1016/j.artmed.2024.102819.

[47] K. Alfrink, I. Keller, G. Kortuem, N. Doorn, Contestable ai by design: Towards a framework, Minds and Machines 33 (2023) 613–639. doi:10.1007/s11023-022-09611-z.

[48] L. Jiříčková, Towards fourth-order cybernetics in AI governance: Sociotechnical perspectives on AI risk management in knowledge-based organizations, in: Proceedings of the 11th International Workshop on Socio-Technical Perspectives in Information Systems (STPIS 2025), volume 4134 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025. URL: https://ceur-ws.org/Vol-4134/paper30.pdf.

[49] H. van der Sijs, J. Aarts, A. Vulto, M. Berg, Overriding of drug safety alerts in computerized physician order entry, Journal of the American Medical Informatics Association 13 (2006) 138–147. doi:10.1197/jamia.M1809.

[50] L. H. Fischer, How sociotechnical reflection influence wellbeing and productivity during GenAI integration, in: Proceedings of the 10th International Conference on Socio-Technical Perspectives in Information Systems (STPIS 2024), volume 3857 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 9–11. URL: https://ceur-ws.org/Vol-3857/short1.pdf.

[51] L. Vicente, F. Cabitza, Machine mentoring by machine learning: Setting a research agenda on how human-ai collaboration can trigger knowledge and bias transmission, in: International Conference on Human-Computer Interaction, Springer, 2025, pp. 72–94.

[52] D. M. Green, J. A. Swets, Signal Detection Theory and Psychophysics, John Wiley & Sons, New York, 1966.