Information Systems Frontiers
https://doi.org/10.1007/s10796-025-10691-2

Critical Thinking in AI-Assisted Deontologically-Governed Professional Decision-Making: When and How Explainability, Reliability, and Transparency Matter

Marion Korosec-Serfaty[1] · Pierre-Majorique Léger[2] · Xavier Parent-Rocheleau[3] · Sylvain Sénécal[4]

Received: 10 July 2025 / Accepted: 22 December 2025

Abstract

Critical thinking is a central safeguard for responsibility and accountability in deontologically-governed professions. Artificial intelligence (AI)-assisted decision-making is increasingly being integrated within these professional workflows. However, AI introduces autonomy, learning, and inscrutability, disrupting critical, reflective, decision-making processes. Given these challenges, this research systematically unpacks how embedding explainability, reliability, and transparency into AI fosters critical thinking. Employing a multi-method approach combining cognitive neuroscience, behavioral and self-report measures, we conducted three experiments with practicing professionals tasked with realistic AI-assisted scenarios. Experiment 1 assessed varying levels of AI-generated reconstructive causal explanations under consistent reliability, Experiment 2 introduced variable reliability. Experiment 3 added transparency through model confidence scores. Results reveal that minimal reconstructive explanations enhance analytical reasoning under reliable conditions, while epistemic uncertainty drives critical engagement when reliability varies. Transparency offers limited restoration of explainability benefits. These findings suggest AI reliability primarily drives critical thinking, informing AI design that preserves professional responsibility and accountability.

Keywords Reconstructive causal explanations · Epistemic uncertainty · Model confidence scores · Critical thinking · Deontologically-governed professions · AI-assisted decision-making

1 Introduction

> An expressway, where nothing grows, cannot be a fallow field. Just as we can grow deaf only because we hear, just as we can grow old only because we were young; so, we can grow thought-poor or even thought-less only because man at the core of his being has the capacity to think. (Heidegger, 1959/1966, p.28).

Even in moments of thoughtlessness, we do not relinquish our capacity to think (Heidegger, 1959/1966). Thinking is a defining trait of human nature (Holyoak & Morrison, 2005), and it is precisely this capacity – to reason and reflect – that must be nurtured and actively exercised in navigating the complexities of artificial intelligence (Mauti, 2024).

Artificial intelligence (AI) marks the latest chapter in the grand scheme of technological advancements (Kim et al., 2021). Created to automate tasks, enhance efficiency, and expand human potential, AI aims to liberate us from repetitive drudgery (Kim et al., 2021). However, the agentic nature of AI stands in tension with the reflective engagement demanded by critical thinking over mere ease and efficiency (Mauti, 2024). This tension is particularly pronounced in professions governed by deontological codes,

✉ Marion Korosec-Serfaty
korosec-serfaty.marion@uqam.ca

[1] Department of Analytics, Operations and Information Technology, School of Management, Université du Québec à Montréal (ESG UQAM), Montréal, Canada

[2] Department of Information Technologies, HEC Montréal, Montréal, Canada

[3] Department of Human Resources Management, HEC Montréal, Montréal, Canada

[4] Department of Marketing, HEC Montréal, Montréal, Canada

Published online: 23 February 2026

which presume professionals' active critical engagement and sound judgment to uphold responsibility and accountability and protect the interests of those they serve (Okasha, 1999; Westerholm, 2009). Such obligations hold professionals to higher standards than those required by law (Verrax, 2016), elevating critical thinking from a valuable skill to an essential safeguard for deontological decision-making.

Earlier generations of rule-based and information systems (IS), rooted in the von Neumann architecture (Goldstine, 1993), supported these professional standards by providing transparent and predictable outputs through structured, reliable information frameworks (Phillips-Wren, 2013) that preserved space for critical thinking and human judgment. In stark contrast, AI introduces autonomy, learning, and inscrutability, disrupting the critical, reflective, deontological decision-making processes central to these professions (Smith, 2021).

Simultaneously, across deontologically-governed professions, AI's capacity to operate autonomously and evolve through data and experience is accelerating the integration of AI-assisted decision-making within professional workflows. Unlike the transparent, deterministic processing of traditional von Neumann-based systems, this transformation is fundamentally reshaping professional practices through, for example, machine learning-powered recruitment platforms that streamline candidate evaluation in human resources management (Pedrami & Vaezi, 2025; Yanamala, 2023), neural network-based imaging that enhances clinical diagnostic accuracy in medicine (Aravazhi et al., 2025), generative design algorithms that optimize structural performance in engineering (Bunian et al., 2024), to automated risk assessment tools that detect financial anomalies in accounting and audit (Ajayi-Nifise et al., 2023).

However, these advancements challenge the *society-profession nexus*, wherein professional autonomy is inseparable from accountability and responsibility as a condition for societal trust (Frankel, 1989). Responsibility involves fulfilling the duties and acting according to the standards required for a given role, while bearing credit or blame for the outcomes of such actions (Bivins, 2006). Yet, AI's autonomy disrupts this principle by transferring decision-making from human oversight to machine-driven insights, creating ambiguity about who retains responsibility (Berente et al., 2021). Accountability, by contrast, entails the obligation to answer for one's actions or decisions and requires providing clear justification for the processes and outcomes of those actions to statutory bodies and those affected by them (Bivins, 2006; Smith, 2021). However, AI's inscrutability, resulting from the opacity of its underlying algorithms, often produces probabilistic outputs that professionals cannot fully explain or defend (Asatiani et al., 2021; Berente et al., 2021). Consequently, diminishing opportunities for critical, reflective engagement, increasing the risk of overreliance on AI-generated outputs (Ellenrieder et al., 2023; Jussupow et al., 2021), and potentially severing decisions from the deontological standards these professions are bound to uphold.

Within this context, a growing body of research points to the importance of creating AI with explainability, transparency, and reliability embedded as system-level properties (henceforth, *properties*) to support critical thinking in professional decision-making. Explainability may mitigate AI's inscrutability by clarifying how and why decisions are made in terms understandable to users (Verma et al., 2020), thereby enabling the critical reflection necessary for professionals to identify errors, question outputs, and adjust decisions accordingly (Ellenrieder et al., 2023). Transparency may complement explainability by making AI mechanisms visible and accessible (Doran et al., 2017), fostering the deliberative analysis that supports trust in the system's integrity while reducing the risk of overreliance (Fecho & Zöll, 2023). Reliability, understood as consistent and accurate AI performance across varying conditions (Bansal et al., 2021), may provide professionals with the predictable foundation necessary for maintaining confidence in AI-supported decisions, even in dynamic contexts (Shneiderman, 2020). We thus posit that these properties constitute the deliberately embedded levers to counter AI's disruptive nature and preserve the critical thinking that anchors accountability and responsibility in the society-profession nexus, thereby calling for a closer examination of how each, considered in isolation or in concert, influences the reflective judgment required in AI-assisted decision-making within deontologically-governed practice. Our aim is to clarify the epistemic conditions under which these AI properties foster or hinder the critical-thinking processes that underlie such judgment.

Given these considerations, we ask: *To what extent do the AI properties of explainability, reliability and transparency foster professionals' critical thinking in AI-assisted decision-making processes within deontologically-governed professional contexts?*

To address this question, we draw upon cognitive neuroscience, explanation-based reasoning, and heuristic-systematic processing theories and adopt an exploratory, three-experiment, multi-method approach involving practicing, experienced human resources (HR) professionals that incrementally unpacks how these AI properties influence critical thinking and sound judgment across the phases of AI-assisted professional decision-making.

We start our investigation with AI explainability, given its potential pivotal role in rendering otherwise uninterpretable AI reasoning accessible and actionable (Doshi-Velez & Kim, 2017). Focusing on causal reconstructive explanations that aim to rebuild AI's reasoning through post-hoc, user-centered narratives, Experiment 1 tests the theorization that their level of completeness influences critical thinking, with minimal reconstructive explanations - which provide partial rebuilding narratives requiring users to draw on internal schemas - eliciting greater critical engagement, relative to fully reconstructive explanations and no AI support. Drawing upon prior NeuroIS research, we triangulate electroencephalography (EEG) and electrodermal activity (EDA) with behavioral and self-report measures to establish and assess the neurophysiological basis of these effects under reliable AI. We find that minimal reconstructive explanations foster greater analytical reasoning and experienced decision-confidence. With these neural correlates established, Experiment 2 investigates whether these observed benefits are contingent upon varying AI reliability using behavioral and self-report measures and reveals that reliability overrides these effects. In Experiment 3, we introduce AI transparency, via model confidence scores, to assess whether it can restore the benefits of minimal reconstructive explanations for critical thinking when AI reliability varies. Our findings demonstrate that transparency emerges as a situational heuristic cue rather than restoring these effects, with reliability remaining the primary driver of critical, cognitive engagement and sound judgment. Together, these results reveal that AI properties foster critical thinking conditionally, rather than uniformly.

This research contributes to the fields of IS and human-centered AI by establishing the conditional nature of AI properties in supporting critical thinking and providing evidence-based guidance for designing AI systems that preserve critical thinking in deontologically-governed contexts.

2 Related Literature and Theoretical Background

2.1 Human Thinking

Thinking, arguably one of the most intricate and complex phenomena shaping human experience (Ernst & von Müller, 2005), has long been regarded by philosophical traditions as central to human existence, illustrating its multifaceted nature: from Descartes' *"Cogito, ergo sum" ("I think, therefore I am")*, which situates thought as the foundation of self-awareness and existence, to Kant's account of the mind organizing sensory data into coherent knowledge through universal categories, to Heidegger's view of thinking as an ontological act that reveals and engages with the essence of being, to name but a few (Descartes, 1637/1998; Heidegger, 1954/1968; Kant, 1781/1998).

Beyond its philosophical dimensions, thinking serves as the mechanism through which the coherence and interrelatedness of reality unfold, reflecting its dynamic and transformative nature (Ernst & von Müller, 2005, p.vi). From a cognitive neuroscientific perspective, this manuscript conceptualizes *thinking* as "encompassing all operations by which humans link mental content to gain new insights or perspectives" (Ernst & von Müller, 2005, p.v).

Within this framework, thinking is understood as a core conscious and non-conscious capacity that entails *reasoning*, or the process of drawing inferences from given information (Kraft et al., 2009); *problem-solving*, understood as a goal-oriented process to achieve a solution (Simon, 1960); and *decision-making*, which involves gathering information, evaluating options, and selecting alternatives to meet objectives within a given context (Holyoak & Morrison, 2005; Kahneman & Tversky, 1979; Simon, 1960) and is commonly conceptualized as a dual-phase, interconnected process encompassing a deliberative stage dedicated to information processing and an executive stage concerned with application and action (Rubinstein, 2007).

2.2 AI-Assisted Decision-Making in Deontologically-Governed Professions

Deontology (from the Greek *deonthos* - due, as it should, as it ought) originally referred to religious-moral obligations before evolving to designate moral theory as a whole (Crudu, 2023). Over time, it acquired a more specific meaning, denoting the proper conduct, actions, and responsibilities of professionals (Crudu, 2023) and establishing strict normative standards between morality, ethics, and law (Frankel, 1989). Decision-making in deontologically-governed professions is thus guided by principles, obligations, and duties outlined in formal codes of conduct, irrespective of outcomes or personal consequences (Frankel, 1989; Rest, 1986).

These codes offer a normative framework that clarifies the moral dimensions of practices, breaches of which may result in formal sanctions (Frankel, 1989). Statutory bodies are tasked with creating, maintaining, and enforcing such codes to uphold professional integrity, responsibility, and accountability. Governed by legal and regulatory frameworks, these bodies ensure that codes reflect societal expectations and the evolving demands of professional practices. For example, although tailored to their respective professions, the codes issued by the American Medical Association and the Québec Order of Certified Human Resources Professionals share core principles of responsibility and accountability. Responsibility is reflected in duties such as prioritizing the welfare of those served, maintaining independent judgment, and grounding decisions in thorough and reliable evaluations; accountability, in turn, is expressed through obligations such as safeguarding confidentiality, ensuring equitable treatment, and providing accurate and transparent explanations to patients or clients (Riddick, 2003).

While such codes have traditionally ensured informed, deontologically aligned decision-making (Seitz & O'neill, 1996; Singhapakdi & Vitell, 1991), AI-assisted decision-making introduces an independent and influential third party into these processes to analyze data, predict outcomes, and generate recommendations (Danry et al., 2023; Li et al., 2023). Designed to optimize decision accuracy and operational efficiency, these systems combine human expertise with AI capacities within established workflows to allow professionals to retain formal authority over decisions (Duan et al., 2019). However, AI's distinct nature poses significant challenges to the principles of responsibility and accountability enshrined in deontological codes, as it introduces learning, autonomy, and inscrutability.

The inscrutability of AI undermines accountability by constraining professionals' ability to provide accurate and transparent explanations for AI-informed decisions. For example, AI tools used in performance evaluation or clinical diagnostics often rely on complex analytical models that may obscure the reasoning behind their outputs, reducing professionals' confidence in decisions and their ability to justify outcomes to affected individuals (Moazemi et al., 2023; Tambe et al., 2019).

Moreover, AI's autonomous learning processes can compromise accountability, particularly the obligation to ensure equitable treatment, by reproducing and amplifying inherent biases in training data (Kadiresan et al., 2022). For instance, AI-assisted diagnostic tools used in hospital emergency departments recommended more advanced tests for wealthier patients while advising lower-income patients to forego further testing despite identical clinical presentations (Omar et al., 2025). Similarly, the Optum algorithm, designed to identify high-risk patients for additional care, assigned lower scores to underserved populations due to historical healthcare disparities (Obermeyer et al., 2019). Amazon's AI-driven recruiting recommender system likewise disadvantaged female candidates for technical roles, perpetuating systemic inequities (Logg, 2019). AI's capacity to unpredictably recalibrate its decision-making processes further challenges professionals' responsibility to ground decisions in reliable, thorough evaluations and to exercise independent judgment. For example, AI systems used to analyze video interviews and recommend candidates may unpredictably adjust how they interpret nonverbal cues (e.g., facial expressions or tone of voice) as they learn from new datasets (Biradar et al., 2024). AI clinical decision-support systems may similarly recalibrate treatment protocols without clinicians being aware of the changes (Moazemi et al., 2023). Such unpredictable adjustments, compounded by AI's inscrutability and potential for biases, may threaten the reliability of AI-generated evaluations and professionals' agency in the decision-making process.

The complexity of these cumulative challenges increases the risk of overreliance on AI outputs, whereby professionals may disengage from discernment and thoughtful evaluation (Benbya et al., 2021; Lebovitz et al., 2022). Thus, potentially eroding the exercise of responsibility and accountability, which intrinsically rely on deliberate, reflective, and critical thought (Arendt, 1971; Mauti, 2024).

2.3 Critical Thinking

We draw upon philosophy to conceptualize critical thinking as a self-directed mode of thought conducive to informed judgment and decision-making in conditions of epistemic uncertainty (Cohen & Freeman, 1996; Lipman, 1987; Siegel, 1989). Critical thinking is (1) *self-correcting* as it involves the ability to assess one's thoughts, identify errors and biases, and address them systematically to refine and improve reasoning; (2) *criteria-driven* as it relies on principles of clarity, accuracy, or relevance to guide judgment; and (3) *context-sensitive* as it recognizes that meaning, assumptions and reasoning may change based on the nuances of a particular situation (Lipman, 1987). Through these properties, critical thinking operates dynamically across decision-making phases, supporting the systematic evaluation and integration of information in the deliberative phase, and guiding its application in the executive phase (Facione, 1990).

Critical thinking is thus inherently adaptive, responding to the evolving complexities of context and content (Edwards, 2007). When applied, this reflective mode of thought fosters calibrated decision-confidence through thoughtful evaluation (Facione, 2015), supports agency with deliberate and informed action (Pisani & Haw, 2023), and grounds trust and intentional, conscious reliance in reasoned, evidence-based judgment (Kleinig, 2016), all of which reflect the sustained exercise of professional accountability and responsibility in deontologically-governed professions.

2.4 Explanations and the Heuristic-Systematic Processing Theory

Research suggests that critical thinking can be effectively stimulated through explanations such as those that encourage reflection on reasoning, guide systematic analysis of evidence, and foster understanding of causal relationships and their implications for outcomes (Lombrozo, 2011; Lombrozo & Carey, 2006).

Explanations function as the *process* and the *product* of reasoning, serving as a dynamic mechanism through which individuals generate, refine, and evaluate understanding, and act as the structured outcome that organizes and communicates information (Lombrozo, 2011). In this sense, explanations drive the discovery of new information, confirm its plausibility or utility, and support the evaluation of its trustworthiness (Lombrozo, 2011). By facilitating the efficient and effective development of cognitive frameworks, explanations are expected to have measurable effects on behavior (Lombrozo, 2011).

To further understand how explanations facilitate critical thinking, we draw upon the heuristic-systematic processing theory, which distinguishes between two modes of information processing: heuristic and systematic (Chaiken & Ledgerwood, 2012). Heuristic processing is relatively automatic, requires minimal cognitive effort, and operates through intuitive judgment and mental shortcuts, often prioritizing efficiency over accuracy and depth (Chaiken & Ledgerwood, 2012). While effective in familiar contexts, heuristic processing may be prone to biases in novel or complex situations where more deliberate reasoning is required (Li et al., 2021). In contrast, systematic processing entails deliberate, analytical evaluation, emphasizing effortful reasoning and the careful integration of evidence, and a thorough analysis of factual consequences (Chaiken & Ledgerwood, 2012; Vance et al., 2015), whereby successful critical thinking exemplifies systematic processing (Bonnefon, 2018). In this regard, explanations may generally facilitate the transition between heuristic and systematic processing by refining intuitive judgments, addressing inconsistencies in reasoning, evaluating the strength of evidence, and exploring alternative perspectives and outcomes (Lombrozo, 2011; Williams & Lombrozo, 2010).

Taken together, when navigating the complexities of AI-assisted decision-making in deontologically-governed professions, explanations of AI reasoning may enable professionals to critically evaluate AI-generated outputs, mitigate the risk of overreliance, and support deliberate, reflective and critical thought, even in the face of epistemic uncertainty, thereby safeguarding the responsibility and accountability central to these roles.

3 Research Overview

We develop and test our hypotheses in three experiments (Table 1). In Experiment 1, we theorize how explainability, in the form of causal reconstructive explanations with varying levels of completeness, influences critical thinking under reliable AI conditions and use neurophysiological, behavioral, and self-report measures to examine these effects. In Experiment 2, we introduce varying AI reliability to test whether the effects of explainability are contingent upon consistent system performance. In Experiment 3, we examine whether AI transparency, operationalized through model confidence scores, can modify the influence of explainability when reliability becomes uncertain.

4 Experiment 1: Explainability and Critical Thinking

4.1 Experiment 1: Research Hypotheses

4.1.1 Causal Reconstructive Explanations in AI Explainability

AI explainability seeks to leverage cognitive frameworks associated with representation, modeling, language, understanding, and learning to reach effective human-AI collaborative decision-making by clarifying the inscrutable nature of AI through explanatory mechanisms (Hoffman et al., 2018). These mechanisms differ in strategies, content, and forms. Explanation content pertains to the specific elements that elucidate a classification, model, or decision, which can

Table 1 Experiments overview

	Experiment 1	Experiment 2	Experiment 3
Objectives	• Examine the effect of the levels of completeness of causal reconstructive explanations on critical thinking and behavioral outcomes during decision-making. • Establish neural markers of critical thinking processes.	• Examine whether the effect of causal reconstructive explanations' completeness on critical thinking is contingent on AI reliability during decision-making.	• Examine whether AI transparency (confidence scores) can restore the effect of the levels of completeness of causal reconstructive explanations under varying AI reliability during decision-making.
Design	Within-subject repeated measures	Mixed design	Mixed design
ID	Explainability (causal reconstructive explanations)	Explainability, reliability	Explainability, reliability, and transparency
DV	Neural correlates of critical thinking, reading times (predictive encoding), decision times (cognitive fluency), selective reliance, and decision-confidence.	Attentional-semantic processing, analytical reasoning, appropriate reliance, calibrated decision-confidence, and calibrated trust in AI.	Attentional-semantic processing, analytical reasoning, appropriate reliance, calibrated decision-confidence, and calibrated trust in AI.
Measures	EEG, EDA, behavioral, and self-reports	Behavioral and self-reports	Behavioral and self-reports

EEG :Electroencephalography, *EDA:* Electrodermal activity, *ID:* Independent variables, *DV:* Dependent variables take various forms, such as textual descriptions, statistical graphs, decision trees, feature histograms, color gradients, feature matrices, or rule sets (Danry et al., 2023; Gregor & Benbasat, 1999). Strategies for conveying this information include analogy-based, counterfactual, dialogue-based, and causal explanations (Danry et al., 2023).

Prior work suggests that causal explanations stimulate critical thinking by encouraging reflection, evidence-based reasoning, and multi-step causal analysis (Lombrozo, 2011; Williams & Lombrozo, 2010), a process that mirrors how professionals traditionally explain and justify their decisions (Morrison et al., 2023). However, in AI explainability, the general effectiveness of causal explanations is a matter of debate: such explanations may, on the one hand, foster deliberate reasoning and mitigate overreliance on AI (Kaur et al., 2020; Morrison et al., 2023), while, on the other hand, encouraging overreliance through the bypassing of verification (Danry et al., 2023).

We propose that these contradictory findings may be explained by examining different approaches to causal explanations. One approach to causal explanations in AI explainability involves reconstructive explanations, which aim to rebuild AI's reasoning by aligning it with the system's original contextual decision-making elements and presenting the rationale in a simplified, user-centered narrative (Vilone & Longo, 2021; Wick et al., 1995; Wick & Thompson, 1992). These reconstructive explanations operate as post-hoc rationales that articulate decision-relevant factors in ways that are understandable and usable for the intended user's judgement and, as such, do not provide access to the underlying computational mechanisms of the AI (Doshi-Velez & Kim, 2017). We argue that reconstructive causal explanations (henceforth, *reconstructive explanations*) may facilitate critical thinking in AI-assisted decision-making. These forms of explanations may enable professionals to assess whether the AI's reasoning aligns with the principles, codes, and standards governing their field. Wherein these reconstructive explanations provide a structured framework – or *schema* – that professionals can use as a starting point for critical evaluation.

We further posit that the level of completeness of AI reconstructive explanations may influence the level of critical engagement during decision-making processes, whereby explanations featuring minimal reconstructive relationships may better align with professionals' reasoning processes, or *schema*, compared to those involving multiple independent causes for different effects (Korman & Khemlani, 2020; Lombrozo, 2007). Such an assumption aligns with the information processing psychology concept of *schematic anticipation* (Bartlett, 1932; De Groot, 1965; Hark, 2003) and neuroscience predictive encoding theories (Millidge et al., 2021), wherein when setting a concrete goal (i.e., establishing a diagnostic or recruiting a candidate according to pre-set criteria), schematic anticipation of the consequences and stages towards reaching this goal is always implied and acts as the starting point for further reasoning processes (Van Den Berg et al., 2020).

Reconstructive AI explanations may thus increase the ease and speed of human reasoning, determined by the level of completeness of the reconstructive explanation, which provides the missing information necessary to close the gap between the anticipated and final solution. In this context, we further posit that fully reconstructive AI explanations that offer detailed, pre-made schema may substitute for professionals' internal schema, and thus reduce critical thinking. In contrast, minimal reconstructive AI explanations may act as skeletal frameworks, requiring professionals to draw on their internal schema to complete the reasoning process, increasing their propensity to think critically.

4.1.2 Neural Correlates of Critical Thinking (H1a-d)

Critical thinking involves deliberate, analytical evaluation, effortful reasoning, and careful integration of evidence for judgment formation and informed decision-making (Lipman, 1987). Drawing from cognitive neuroscience, critical thinking is understood as emerging from the interaction of higher-order neural processes of attention, information and semantic processing, and analytical reasoning, each supported by the activation and inhibition of interconnected brain regions - or *neural correlates* - expressed through neural oscillations (i.e., rhythmic patterns of electrical activity across groups of neurons that facilitate communication and coordinate cognitive processes) (Klimesch, 2018; Newman, 2019; Sarter et al., 1996) and operating within recurrent loops (Knudsen, 2007). These neural correlates, observable in specific cortices of the human brain through electroencephalography (EEG) as variations in brainwave frequency and amplitude across alpha, beta, gamma, delta, and theta bands, indicate cognitive engagement, which aids in the inference of critical thought, and serve as reliable indicators of such engagement as critical thinking unfolds dynamically across the deliberative and executive phases of decision-making in interdependent ways (Müller-Putz et al., 2015).

Deliberative Phase: Attention, Information, and Semantic Processing In the deliberative phase, critical thinking initially relies on the *sustained* and *selective* allocation of attentional resources to relevant evidence and potential inconsistencies. Sustained attention, understood as the self-directed maintenance of cognitive focus under non-arousing conditions (Clayton et al., 2015), is associated with increased activation in occipital (visual processing) and parietal (attention allocation) brain regions (Coslett &

Schwartz, 2018; Kawasaki & Yamaguchi, 2012), supported by increases in theta (facilitating cognitive control) and delta-band oscillatory activity (facilitating sustained focus) during tasks requiring detailed evaluation of evidence (Cavanagh & Frank, 2014; Harmony, 2013).

Beyond sustained attention, critical thinking requires selectively attending to relevant information while inhibiting distractions, particularly when evaluating conflicting evidence. Selective attention has been associated with increased temporoparietal theta (supporting mismatch detection and reasoning when discrepancies in expected information must be identified) activity (Hiraishi et al., 2021; Jeunet et al., 2018); temporoparietal delta (associated with internal focus and the attention required for discrepancy detection during judgment-building) activity (Harmony, 2013); and temporoparietal beta (facilitating conflict resolution and integration of competing inputs during judgement-building) activity (Park et al., 2018).

Building upon this foundation of attentional processes, critical thinking next relies on *information and semantic processing* to interpret, organize, and understand incoming data. Information processing involves categorizing raw sensory inputs, such as phonological or visual stimuli, into meaningful and coherent representations (Cacioppo et al., 2017). Semantic processing links these representations with stored knowledge and contextual frameworks, facilitating understanding and application (Binder et al., 2009). These interconnected processes engage the frontal (cognitive control) (Miller & Cohen, 2001), centroparietal (information organization and semantic integration) (Petrides, 2013), and parietal (semantic processing) brain regions (Coslett & Schwartz, 2018), and are coupled with increases in theta (supporting cognitive control and integration) and delta (supporting sustained focus) activity, which together facilitate the integration of information into coherent judgment (Coslett & Schwartz, 2018; Petrides, 2013).

Executive Phase: Analytical Reasoning As the deliberative phase progresses toward decision execution, critical thinking culminates in analytical reasoning, where choice-relevant information is synthesized and evaluated to form coherent, informed judgment (Alexander, 2014). Analytical reasoning engages the dorsolateral prefrontal (working memory and manipulation of information) (Barbey et al., 2013), frontal (cognitive reasoning and inhibitory control) (Chayer & Freedman, 2001), parietal (decision-making under uncertainty) (Naaz et al., 2021), centroparietal (analytical realignment and integration) (Kayser et al., 2012), and anterior cingulate (conflict monitoring and decision confidence) cortical regions (Botvinick et al., 2004). Neural activity supporting analytical reasoning is characterized by increases in frontal beta (inhibition) (Rojas et al., 2020) and gamma (facilitating the rapid correlation of information) (Başar-Eroglu et al., 1996); centroparietal theta (reflecting realignment for analytical purposes) (Kayser et al., 2012); frontocentral theta and delta (associated with sustained cognitive engagement and conflict monitoring) (Cavanagh & Frank, 2014; Harmony, 2013); parietal beta (supporting inhibition of irrelevant outputs and cognitive focus) (Naaz et al., 2021); and parietal-occipital delta (integration of visual and conceptual information during judgment consolidation) activity (Harmony, 2013).

Critical thinking reflects the coordinated activation of sustained attention, information processing, semantic analysis, and analytical reasoning across the deliberative and executive decision-making phases. Accordingly, we hypothesize that, compared to decision processes without AI support, the level of completeness in AI reconstructive explanations will influence the neural correlates of critical thinking as follows:

H1a. No AI support will elicit greater deliberative-phase neural activity (attention, information, and semantic processing) than either minimal or fully reconstructive explanations.

H1b. Minimal reconstructive explanations will elicit greater deliberative-phase neural activity (attention, information, and semantic processing) than fully reconstructive explanations.

H1c. No AI support will elicit greater executive-phase neural activity (analytical reasoning) than either minimal or fully reconstructive explanations.

H1d. Minimal reconstructive explanations will elicit greater executive-phase neural activity (analytical reasoning) than fully reconstructive explanations.

4.1.3 Neurophysiological Correlates and Self-Reported Decision-Confidence (H2a-c)

Building on our theorization of critical thinking across decision-making phases, we now consider how varying levels of reconstructive AI explanations contribute, through this process, to the formation of decision-confidence, as a distinct affective-cognitive outcome.

Decision-confidence refers to the subjective sense of certainty regarding the accuracy and validity of a given decision, independent of its actual correctness (Yeung & Summerfield, 2012). In this context, critical thinking grounds decision-confidence in analytical rigor rather than heuristic-driven overconfidence by supporting systematic evaluation, outcome anticipation, and reflective appraisal of evidence (Chaiken & Maheswaran, 1994; Facione, 2015).

In AI-assisted decision-making, the format and structure of explanations are generally expected to influence decision-confidence (Alufaisan et al., 2021; Gregor & Benbasat, 1999). Furthermore, empirical research shows that detailed or complex explanations may introduce cognitive overload, impair judgment quality, or foster a false sense of certainty (Poursabzi-Sangdeh et al., 2021). These findings suggest that the level of completeness of AI explanations, and thus their capacity to stimulate critical thinking, may support the process through which decision-confidence is actively formed.

Decision-confidence is expected to emerge during the executive phase of decision-making, after information has been integrated and a judgment formed, and is reflected in distinct neural and physiological correlates. At the neural level, this process involves increased beta activity in parietal and centroparietal regions (supporting metacognitive certainty, judgment consolidation, and confidence in decision outcomes), reflecting the integration of information and resolution of decisional conflict (Naaz et al., 2021; Wang et al., 2016; Wilhelm et al., 2021). Increased frontocentral beta and gamma activity has been associated with executive attention, internal monitoring and confident task engagement, respectively (Wang et al., 2016). Furthermore, decreased temporoparietal beta activity relates to reduced inhibitory demands and less reliance on internal evaluation under uncertainty, and thus, with greater confidence in one's judgement (Park et al., 2018).

Physiologically, decision-confidence may be inferred through electrodermal activity (EDA), which captures fluctuations in skin conductance that reflect sympathetic nervous system activation during cognitive and emotional processing (Boucsein, 1999; Riedl, 2013; Tranel & Damasio, 1994). Among EDA indices, skin conductance responses (SCRs) are well-established biomarkers of confidence dynamics. Higher SCR amplitudes during decision execution are typically associated with increased arousal, uncertainty, or anxiety, whereas lower amplitudes indicate reduced anticipatory stress and greater subjective confidence in the decision made (Bechara et al., 1997; Dawson et al., 2011).

Accordingly, we propose that minimal reconstructive explanations lead to greater decision-confidence than either fully or reconstructive explanations or no AI support, and more specifically:

> H2a. Minimal reconstructive explanations will elicit greater executive-phase neural correlates of decision-confidence than either fully reconstructive explanations or no AI support.

> H2b. Minimal reconstructive explanations will elicit greater executive-phase physiological correlates of decision-confidence than either fully reconstructive explanations or no AI support.

> H2c. Post-decision, minimal reconstructive explanations will result in greater self-reported decision-confidence than either fully reconstructive explanations or no AI support.

4.1.4 Behavioral Indicators of Schematic Anticipation and Selective Reliance (H3a-c)

We further posit that the varying levels of completeness in reconstructive AI explanations, through their influence on critical thinking via schematic anticipation processes, are reflected in behavioral patterns across the deliberative and executive phases of decision-making.

During the deliberative phase, *reading time* provides a temporal indicator of *predictive encoding*, whereby professionals assess the alignment between AI explanations and their internal schema (Fincher-Kiefer, 1996). Minimal reconstructive explanations may be processed more rapidly, as they allow internal schema to be more readily drawn upon without extensive parsing of AI-generated reasoning. In contrast, fully reconstructive explanations may prompt longer reading time due to the need to process and reconcile detailed, externally provided reasoning.

In the executive phase, *decision time* reflects the cognitive fluency with which AI reasoning is integrated with internally held standards (Buçinca et al., 2021; Cao et al., 2023; Schwarz et al., 2021). Minimal reconstructive explanations may facilitate faster decisions by providing a skeletal narrative that aligns with anticipated schema, supporting efficient predictive encoding, and confident execution. Conversely, fully reconstructive explanations may introduce additional interpretive steps requiring integration, thus extending decision time.

Selective reliance provides a behavioral indicator of critical engagement with AI outputs under conditions of consistent reliability, capturing the extent to which AI-generated reasoning is critically evaluated and acted upon, contrasting with complacency (Parasuraman et al., 1993). In such settings, selective reliance patterns reflect differences in critical appraisal rather than variation in performance accuracy. Fully reconstructive explanations may increase reliance by offering a terminal narrative path that feels complete and discourages further questioning. Conversely, minimal reconstructive explanations may encourage more critical appraisal and, consequently, more selective uptake.

> H3a. Minimal reconstructive explanations will elicit reduced predictive encoding activity (shorter reading time) during the deliberative phase than fully reconstructive explanations, but longer than no AI support.
>
> H3b. Minimal reconstructive explanations will elicit greater cognitive fluency (shorter decision times) during the executive phase than fully reconstructive explanations, but less than no AI support.
>
> H3c. Minimal reconstructive explanations will result in greater selective reliance on AI than fully reconstructive explanations.

4.2 Experiment 1: Methodology

4.2.1 Participants

We recruited 23 active HR professionals, registered members of a recognized professional order (aged 22–56; $M=36.5$; $SD=11.18$; 70% female; with an average of 13.3 years of experience). This sample size aligns with prior EEG and EDA studies employing a within-subject repeated measures design (Boucsein, 1999; Boucsein & Thum, 1997; Riedl et al., 2020). Participants met eligibility criteria for EEG and EDA measurements. The approximately 90-minute study included consent, sensor setup and removal, calibration, scenario reading, task completion, and post-task sociodemographic questionnaires. Participants received CAD150 compensation. The experiment was conducted at a North American university behavioral laboratory and was approved by the Research Ethics Board (certificate #2023–5041). All participants provided written informed consent.

4.2.2 Experimental Task and Design

Experiment 1 employed a scenario-based, one-factor within-subject Wizard-of-Oz design[1] (Riek, 2012) simulating AI-assisted decision-making. Participants were tasked with pre-selecting Web Analyst candidates from pairs of resumes for a fictitious academic institution, based on three recruitment criteria: education, experience, and programming competencies.

The task was implemented across 30 counterbalanced trials, divided equally into three conditions: (1) no AI support (NAS), in which participants evaluated resumes independently, without AI recommendations; (2) minimal reconstructive AI explanations (MRE), where AI provided a recommendation, and a single composite score derived from the three criteria; and (3) fully reconstructive AI explanations (FRE), where AI provided a recommendation, individual scores for each criterion, and highlighted relevant resume content (see Figs. 1 and 2).

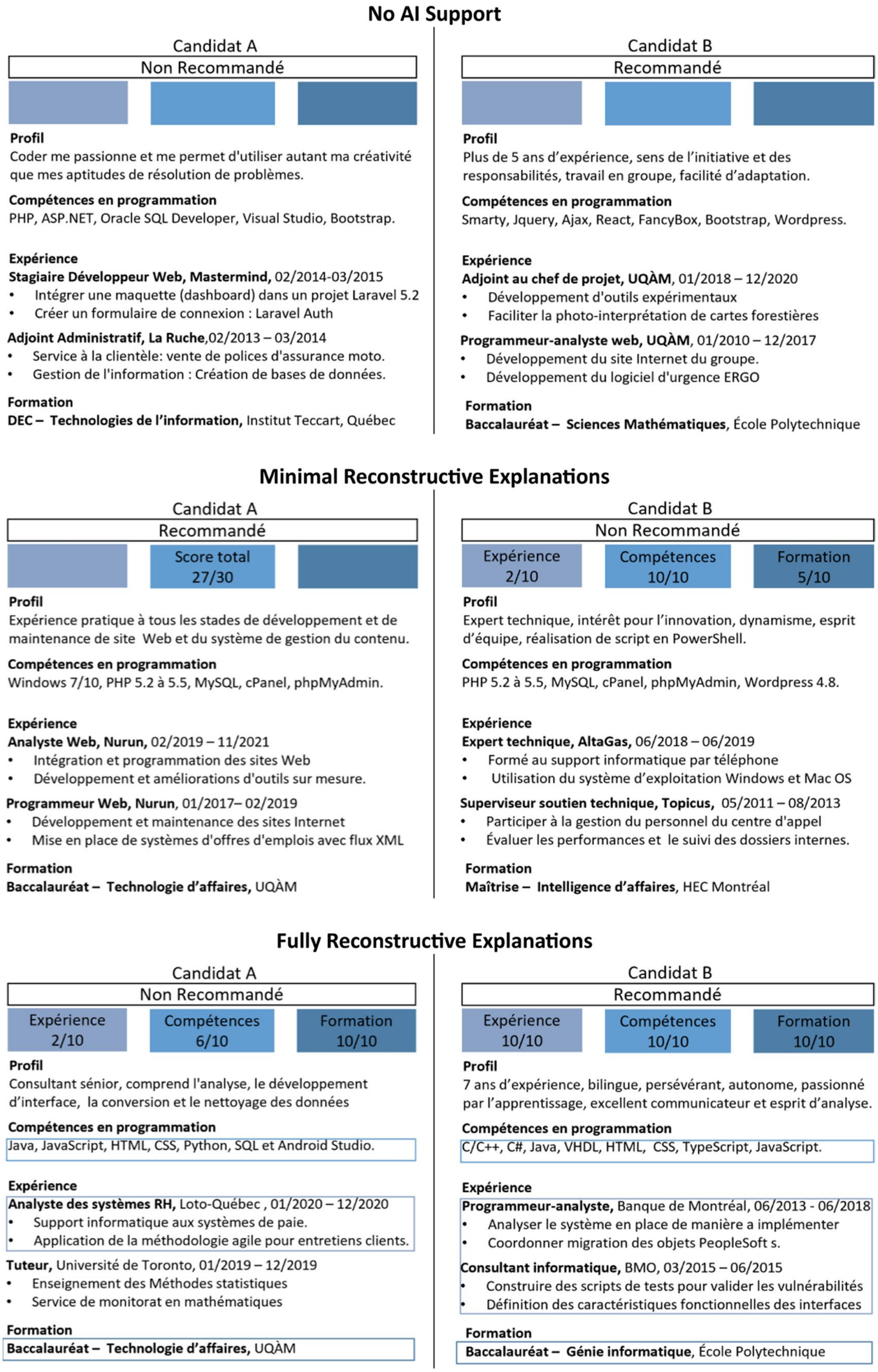
Prior to the task, participants were introduced to the institution, the job description, and the recruitment criteria. They were also informed that, in certain trials, an AI would provide candidate recommendations. Each trial comprised a deliberative phase, during which two unique resumes were displayed side-by-side for self-paced assessment, followed by an executive phase, where a separate screen prompted candidate selection. A 10-second fixation cross was presented between trials. The task was self-paced, though participants were informed that resume reading typically took about one minute.

4.2.3 Stimuli Design

Stimuli consisted of 60 unique resumes created by anonymizing and recombining content from real Web Analyst resumes obtained from the university's HR department. Resumes were presented in a standardized two-column layout (one column per candidate) to control layout-induced neural variability (Wardle et al., 2016). Variations were limited to textual content (i.e., profile, education, experience, and programming skills) and the level of reconstructive AI explainability. Visual cues for explainability (i.e., composite and criterion-level scores, highlighted text) were embedded directly into the resume layout in the MRE and FRE conditions (see Fig. 1). Neutral colors were used throughout to minimize affective bias (Yoto et al., 2007).

AI recommendations were generated using a Preference-Dependent Measure (Aksoy et al., 2011), based on aggregate scores across the three recruitment criteria. Scoring rules allocated up to 10 points per criterion: experience (2/5/10 points for <2, 2–5, or >5 years), education (3/5/10 points depending on degree relevance), and programming (2 points per language, max 10). The candidate with the highest score was always recommended, ensuring consistent reliability and isolating the effects of reconstructive explanation on critical thinking and related outcomes, while minimizing confounds from mistrust or trust recovery. Recommendations were counterbalanced across trials and conditions.

[1] The Wizard-of-Oz methodology (Riek, 2012) is a widely used technique in human-computer interaction research (Steinfeld et al., 2009) for simulating interactions and examining users' responses to hypothetical systems, such as fluent, real-world AI (Hinds et al., 2004). In this paradigm, participants interact with what appears to be an autonomous system, while its functions are covertly operated by a human "wizard." Following Steinfeld et al.'s (2009) classification, this study employed a classical Wizard-of-Oz design, in which the technological functionality is assumed, and the analytic focus remains exclusively on participants' behavior and responses.



No AI Support

Candidat A

Non Recommandé

Profil
Coder me passionne et me permet d'utiliser autant ma créativité que mes aptitudes de résolution de problèmes.

Compétences en programmation
PHP, ASP.NET, Oracle SQL Developer, Visual Studio, Bootstrap.

Expérience
Stagiaire Développeur Web, Mastermind, 02/2014-03/2015
- Intégrer une maquette (dashboard) dans un projet Laravel 5.2
- Créer un formulaire de connexion : Laravel Auth

Adjoint Administratif, La Ruche,02/2013 – 03/2014
- Service à la clientèle: vente de polices d'assurance moto.
- Gestion de l'information : Création de bases de données.

Formation
DEC – Technologies de l'information, Institut Teccart, Québec

Candidat B

Recommandé

Profil
Plus de 5 ans d'expérience, sens de l'initiative et des responsabilités, travail en groupe, facilité d'adaptation.

Compétences en programmation
Smarty, Jquery, Ajax, React, FancyBox, Bootstrap, Wordpress.

Expérience
Adjoint au chef de projet, UQÀM, 01/2018 – 12/2020
- Développement d'outils expérimentaux
- Faciliter la photo-interprétation de cartes forestières

Programmeur-analyste web, UQÀM, 01/2010 – 12/2017
- Développement du site Internet du groupe.
- Développement du logiciel d'urgence ERGO

Formation
Baccalauréat – Sciences Mathématiques, École Polytechnique

Minimal Reconstructive Explanations

Candidat A

Recommandé

Score total
27/30

Profil
Expérience pratique à tous les stades de développement et de maintenance de site Web et du système de gestion du contenu.

Compétences en programmation
Windows 7/10, PHP 5.2 à 5.5, MySQL, cPanel, phpMyAdmin.

Expérience
Analyste Web, Nurun, 02/2019 – 11/2021
- Intégration et programmation des sites Web
- Développement et améliorations d'outils sur mesure.

Programmeur Web, Nurun, 01/2017– 02/2019
- Développement et maintenance des sites Internet
- Mise en place de systèmes d'offres d'emplois avec flux XML

Formation
Baccalauréat – Technologie d'affaires, UQÀM

Candidat B

Non Recommandé

Expérience	Compétences	Formation
2/10	10/10	5/10

Profil
Expert technique, intérêt pour l'innovation, dynamisme, esprit d'équipe, réalisation de script en PowerShell.

Compétences en programmation
PHP 5.2 à 5.5, MySQL, cPanel, phpMyAdmin, Wordpress 4.8.

Expérience
Expert technique, AltaGas, 06/2018 – 06/2019
- Formé au support informatique par téléphone
- Utilisation du système d'exploitation Windows et Mac OS

Superviseur soutien technique, Topicus, 05/2011 – 08/2013
- Participer à la gestion du personnel du centre d'appel
- Évaluer les performances et le suivi des dossiers internes.

Formation
Maîtrise – Intelligence d'affaires, HEC Montréal

Fully Reconstructive Explanations

Candidat A

Non Recommandé

Expérience	Compétences	Formation
2/10	6/10	10/10

Profil
Consultant sénior, comprend l'analyse, le développement d'interface, la conversion et le nettoyage des données

Compétences en programmation
Java, JavaScript, HTML, CSS, Python, SQL et Android Studio.

Expérience
Analyste des systèmes RH, Loto-Québec , 01/2020 – 12/2020
- Support informatique aux systèmes de paie.
- Application de la méthodologie agile pour entretiens clients.

Tuteur, Université de Toronto, 01/2019 – 12/2019
- Enseignement des Méthodes statistiques
- Service de monitorat en mathématiques

Formation
Baccalauréat – Technologie d'affaires, UQÀM

Candidat B

Recommandé

Expérience	Compétences	Formation
10/10	10/10	10/10

Profil
7 ans d'expérience, bilingue, persévérant, autonome, passionné par l'apprentissage, excellent communicateur et esprit d'analyse.

Compétences en programmation
C/C++, C#, Java, VHDL, HTML, CSS, TypeScript, JavaScript.

Expérience
Programmeur-analyste, Banque de Montréal, 06/2013 - 06/2018
- Analyser le système en place de manière à implémenter
- Coordonner migration des objets PeopleSoft s.

Consultant informatique, BMO, 03/2015 – 06/2015
- Construire des scripts de tests pour valider les vulnérabilités
- Définition des caractéristiques fonctionnelles des interfaces

Formation
Baccalauréat – Génie informatique, École Polytechnique

Fig. 1 Experiment 1: Stimuli as presented to participants. Font sizes reduced for publication layout

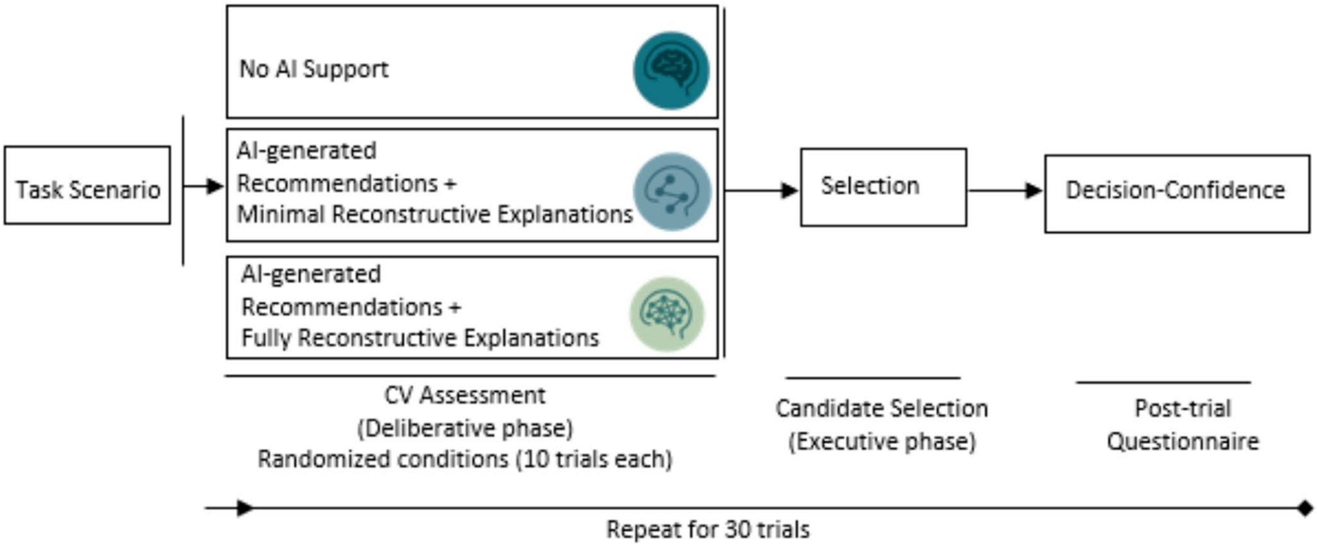

No AI Support
AI-generated Recommendations + Minimal Reconstructive Explanations
AI-generated Recommendations + Fully Reconstructive Explanations
Task Scenario
Selection
Decision-Confidence
CV Assessment (Deliberative phase) Randomized conditions (10 trials each)
Candidate Selection (Executive phase)
Post-trial Questionnaire
Repeat for 30 trials

Fig. 2 Experiment 1: Experimental design

4.2.4 Measures

Electroencephalography (EEG) activity was continuously recorded to extract time–frequency amplitudes in the beta, gamma, delta, and theta bands during the deliberative (resume assessment) and executive (candidate selection) phases. EDA was recorded concurrently to derive SCR. Reading time was measured as the interval from resume display onset to the spacebar press, and decision time as the interval from candidate selection screen onset to the participant's final decision. The number of AI-recommended candidates selected was used as a binary behavioral measure of selective reliance (Candrian & Scherer, 2022). Self-reported decision-confidence was measured after each trial using a single item from Inbar et al. (2011) on a 5-point scale (see Appendix, Table 6). Random attention checks were inserted every seven trials to ensure response validity (Shamon & Berning, 2020).

4.2.5 Procedure

After providing informed consent, participants were fitted with the EEG cap according to the international 10–20 system, and EDA sensors were applied to the palm of the non-dominant hand. They were seated at a desk equipped with a screen, keyboard, and mouse. Following a 90-second fixation cross, participants were presented with the task scenario and asked to verbally recall the three recruitment criteria to ensure comprehension. They then completed three practice trials, one for each experimental condition. After the task, participants completed a sociodemographic questionnaire, and all sensors were removed.

4.2.6 Material and Apparatus

The experimental task was developed and administered using E-Prime 3 software (Psychology Software Tools, Pittsburgh, PA), which sent time markers for stimuli presentation Observer XT (Noldus, Wageningen, The Netherlands). EEG data were acquired using a 32-channel EASYCAP headset (EASYCAP GmbH, Wörthsee, Germany) with electrodes positioned according to the 10–20 system. Recordings were captured at a 250 Hz sampling rate via Brain Vision acquisition software (Brain Vision, Morrisville, USA). Mastoid electrodes served as ground electrodes during recording. EDA data were recorded at 256 Hz using a Biopac MP-150 system running via AcqKnowledge 4.4 software (Biopac Systems Inc., Santa Barbara, CA). Post-hoc synchronization of EDA data was performed using the Cobalt Photobooth software 305 (Courtemanche et al., 2022; Léger et al., 2019). EEG data preprocessing was conducted in MATLAB (version 2020b, MathWorks, Natick, MA, USA) using the EEGLab toolbox (Delorme & Makeig, 2004). Statistical analyses were performed in SPSS (version 25, IBM, Armonk, NY, USA).

4.2.7 Data Preprocessing and Analysis

Electroencephalography (EEG) data were preprocessed following established guidelines for EEG research (Müller-Putz et al., 2015). Noisy EEG channels were removed, and independent component analysis (ICA) was applied to eliminate physiological artifacts and periodic noise. Signals were band-pass filtered between 3 and 40 Hz. Based on observed variability in task timing across conditions, post-stimulus windows of 21 and 3 s were defined for the deliberative and executive phases, respectively. Each window was segmented into three-second intervals. EEG data were epoched from − 1 to 4 s relative to each interval onset. Trials with substantial movement artifacts were excluded after visual inspection. EEG signals were decomposed into delta (2–4 Hz), theta (5–7 Hz), beta (15–29 Hz), and gamma (30–59 Hz) bands (Rempe et al., 2023) using the Hilbert transform. Time-frequency amplitude envelopes for each band were extracted and averaged across epochs within each participant, task phase, and condition. Final mean amplitude values were calculated by averaging over a 0–3 s interval in each window of interest for each participant. These values were used in subsequent statistical analyses. Four participants were excluded due to excessive EEG noise.

Electrodermal activity (EDA) data were preprocessed following standard procedures (Boucsein, 2012). Frequency decomposition was applied to EDA data to derive SCR. Raw signals were down-sampled from 256 Hz to 100 Hz to produce one-second averages. Data were segmented and averaged by participant, trial, and condition. To maintain consistency across conditions and account for the brief duration of the executive period, a fixed 3-second post-stimulus window (from decision screen onset) was used to extract SCR values (Figner & Murphy, 2011). Two participants were excluded due to excessive signal artifacts.

4.2.8 Data Analysis

Statistical significance was set at $p<0.05$. Normality was assessed using Kolmogorov-Smirnov and Shapiro-Wilk tests ($p>0.05$); non-parametric tests were applied where appropriate (Rojas et al., 2020).

Electroencephalography amplitudes were clustered into nine cortical regions of interest corresponding to major brain areas (Bin et al., 2019): frontal (Fp1, FP2, F3, F4, F8, Fz), frontocentral (FC1, FC2, FC5, FC6), temporal (T7, T8), temporoparietal (TP9, TP10), central (C3, C4, Cz), centroparietal (CP1, CP2, CP5, CP6), parietal (P3, P4, P7, P8, Pz), occipital (O1, O2, Oz), and parieto-occipital (PO9, PO10). For each frequency band and task phases (deliberative and executive), within-subject, between-condition comparisons were conducted using the Wilcoxon signed-rank test (Islam et al., 2020; Jebelli et al., 2018; Kurihara et al., 2022).

Skin conductance response (SCR) data were z-transformed per participant for each task window and analyzed using Wilcoxon signed-rank tests to assess within- and between-condition differences (Braithwaite & Watson, 2015; Mühl et al., 2020).

Paired sample t-tests were used to assess main and specific effects of the experimental conditions on reading and decision times. The number of AI-recommended candidates selected vs. not selected was calculated per condition, and differences between conditions were assessed using Fisher's exact test for categorical data (Lury & Fisher, 1972).

Self-reported decision-confidence was averaged per participant per condition and analyzed with paired sample t-tests.

4.3 Experiment 1: Results

4.3.1 Neural Correlates of Critical Thinking (H1a-d)

Results from the between-condition analysis of the EEG data during the deliberative phase support H1a (Fig. 3). Related to sustained attention, delta-band amplitudes were significantly higher in NAS against MRE (occipital: $Z = -2.65$, $p=0.008$; parietal: $Z=-2.61$, $p=0.009$) and FRE (occipital: $Z = -2.65$, $p=0.008$; parietal: $Z = -2.65$, $p=0.008$). Similarly, theta-band amplitudes were significantly higher in NAS compared to MRE (parietal: $Z = -2.25$, $p=0.02$) and FRE (occipital: $Z = -2.85$, $p=0.004$; parietal: $Z = -2.25$, $p=0.02$).

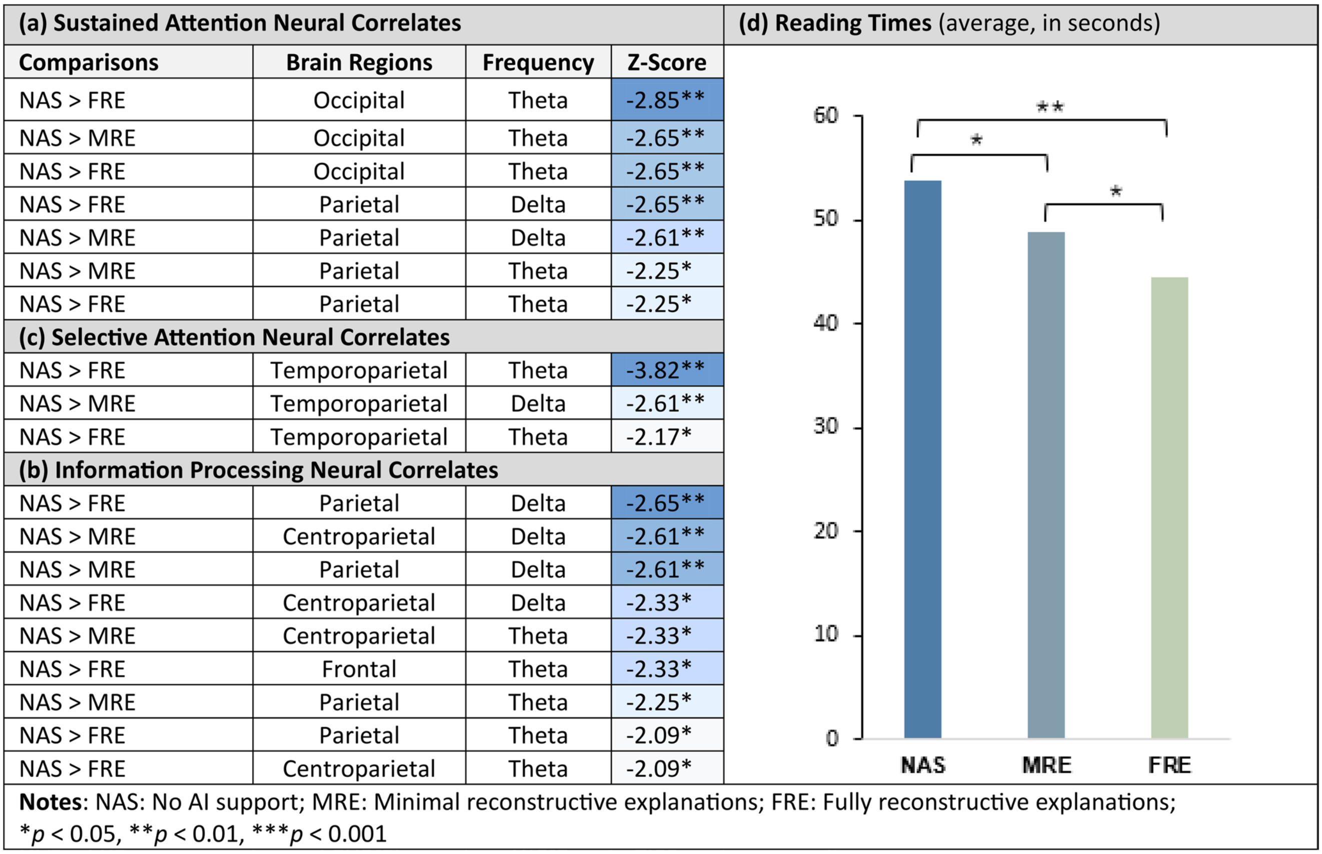
For selective attention, delta-band amplitudes in the temporoparietal region were significantly higher in NAS compared to MRE ($Z=-2.61$, $p=0.009$). When comparing NAS to FER, both delta-band ($Z=-2.17$, $p=0.03$) and theta-band ($Z=-3.82$, $p<0.001$) amplitudes were significantly higher in the temporoparietal region.

Regarding information and semantic processing, delta-band amplitudes were significantly higher in NAS compared to MER (centroparietal: $Z=-2.61$, $p=0.009$; parietal: $Z = -2.61$, $p=0.009$) and FER (centroparietal: $Z = -2.33$, $p=0.02$; parietal: $Z= -2.65$, $p=0.008$). Theta-band amplitudes were also significantly higher in NAS compared to MER (centroparietal: $Z=-2.33$, $p=0.02$; parietal: $Z=-2.25$, $p=0.02$) and FER (centroparietal: $Z=-2.09$, $p=0.03$; frontal: $Z=-2.33$, $p=0.02$; parietal: $Z=-2.09$, $p=0.03$).

With respect to H1b, no significant differences were observed between MER and FER across theta- or delta-band activity in the brain regions associated with sustained and selective attention, or information and systematic processing. Thus, H1b was not supported.

Turning to H1c and analytical reasoning during the executive phase (Fig. 4), beta-band amplitudes in the parietal region ($Z=-2.87$, $p=0.004$), gamma-band amplitudes in the frontocentral region ($Z = -2.05$, $p=0.04$) were significantly higher in NAS against MER. Furthermore, delta amplitudes were significantly higher in the parietal-occipital region in NAS against MR ($Z=-2.15$, $p=0.03$) and FRE ($Z = -2.41$, $p=0.01$). These results support H1c.

Regarding H1d, beta-band amplitudes in the centroparietal region were significantly higher in MER than FER (arding H1d$Z=-2.11$, $p=0.03$), supporting H1d.



(a) Sustained Attention Neural Correlates			
Comparisons	**Brain Regions**	**Frequency**	**Z-Score**
NAS > FRE	Occipital	Theta	-2.85**
NAS > MRE	Occipital	Theta	-2.65**
NAS > FRE	Occipital	Theta	-2.65**
NAS > FRE	Parietal	Delta	-2.65**
NAS > MRE	Parietal	Delta	-2.61**
NAS > MRE	Parietal	Theta	-2.25*
NAS > FRE	Parietal	Theta	-2.25*
(c) Selective Attention Neural Correlates			
NAS > FRE	Temporoparietal	Theta	-3.82**
NAS > MRE	Temporoparietal	Delta	-2.61**
NAS > FRE	Temporoparietal	Theta	-2.17*
(b) Information Processing Neural Correlates			
NAS > FRE	Parietal	Delta	-2.65**
NAS > MRE	Centroparietal	Delta	-2.61**
NAS > MRE	Parietal	Delta	-2.61**
NAS > FRE	Centroparietal	Delta	-2.33*
NAS > MRE	Centroparietal	Theta	-2.33*
NAS > FRE	Frontal	Theta	-2.33*
NAS > MRE	Parietal	Theta	-2.25*
NAS > FRE	Parietal	Theta	-2.09*
NAS > FRE	Centroparietal	Theta	-2.09*

(d) Reading Times (average, in seconds)

Notes: NAS: No AI support; MRE: Minimal reconstructive explanations; FRE: Fully reconstructive explanations; *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

Fig. 3 Experiment 1: Neural correlates of critical thinking and reading times during the deliberative phase

4.3.2 Neurophysiological Correlates of, and Perceived Decision-Confidence (H2a-c)

Turning to H2a, and decision-confidence (Fig. 4), results from the EEG data during the executive phase showed that, compared to FRE, MRE was associated with significantly higher beta amplitudes in centroparietal ($Z = -2.11$, $p=0.03$) and frontocentral ($Z = -2.05$, $p=0.04$), and significantly lower temporoparietal-beta amplitudes ($Z = -3.01, p=0.03$). Moreover, temporoparietal beta amplitudes were significantly lower in MRE against NAS (Z = −2.24, $p=0.04$). However, contrary to our hypothesis, parietal beta amplitudes were significantly higher in MRE than NAS (Z = −2.87, $p=0.004$), therefore partially supporting H2a.

For H2b, analysis of the EDA data showed that SCR amplitudes were, on average, significantly lower in MRE than in NAS ($Z=-4.41$, $p<0.001$), and significantly lower in NAS than in FRE ($Z=-3.19$, $p=0.001$), thereby providing partial support for H2b.

For H2c, self-reported decision-confidence measured post-decision was significantly higher in FER than in MRE ($t(224)=4.87$, $p<0.001$, $d=0.32$), and significantly higher in MRE than in NAS ($t(224)=3.10$, $p=0.002$, $d=0.29$), thereby partially supporting H2c.

4.3.3 Behavioral Indicators of Schematic Anticipation and Selective Reliance (H3a-c)

For H3a, the between-condition analysis of the behavioral data, showed differences in reading times, which were significantly longer in NAS than in MER ($t(224)=3.10$, $p=0.02$, $d=0.21$), in NAS than in FER ($t(224)=5.23$, $p<0.001$, $d=0.35$), and in MRE than in FER ($t(224)=2.42$, $p=0.002$, $d=0.21$) (Fig. 3).

H3b was not supported. Decision times showed a non-significant curvilinear trend, with the shortest average decision time observed in NAS and the longest in MER.

H3c was also not supported. Although the average number of AI recommendations followed was higher in FER than in MER, the difference was not statistically significant.

4.4 Experiment 1: Discussion

Using a NeuroIS approach, Experiment 1 investigated how consistently reliable AI, paired with varying levels of reconstructive causal explainability completeness, influences the cognitive processes and neural dynamics underlying critical thinking, and how these, in turn, contribute to decision-confidence and selective reliance in AI-assisted

(a) Analytical Reasoning Neural Correlates

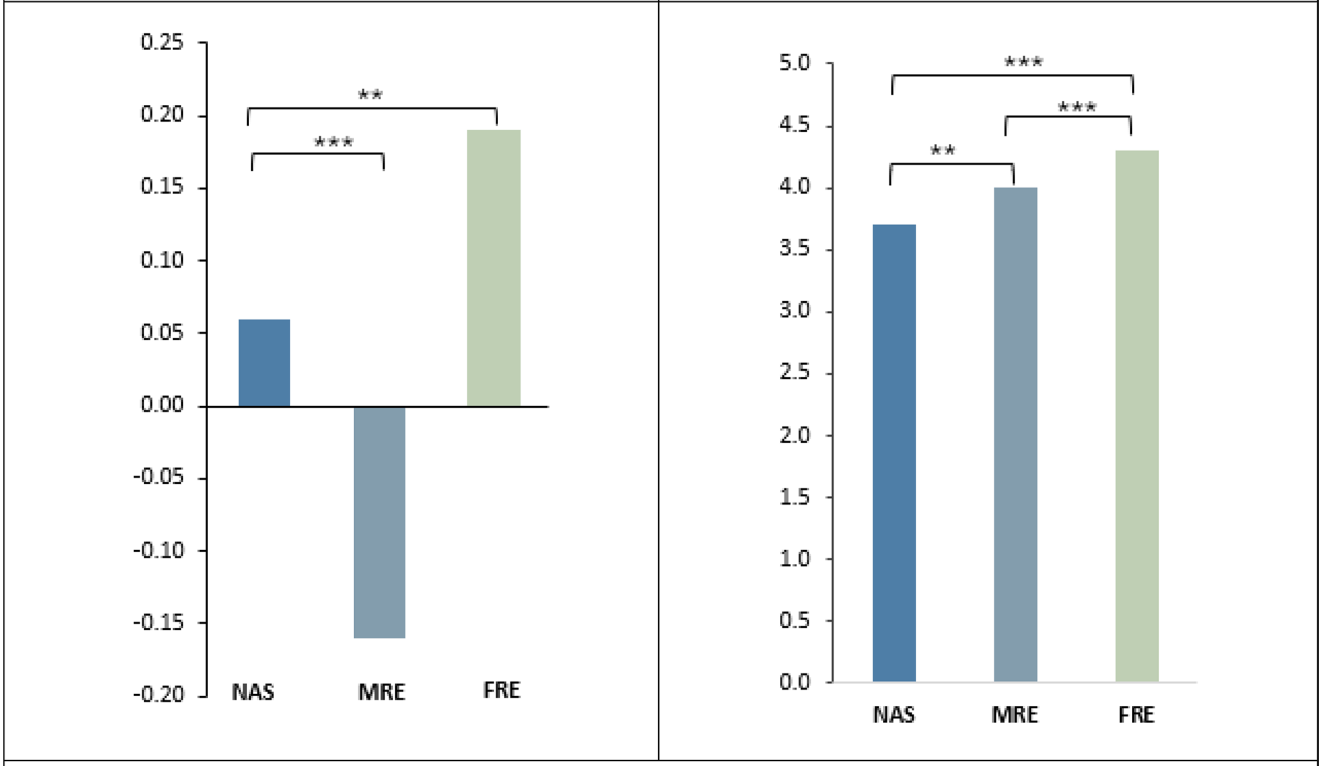
Comparisons	Brain Regions	Frequency	Z-Score
NAS > MRE	Parietal	Beta	-2.87**
NAS > FRE	Parietal-occipital	Delta	-2.41**
NAS > MRE	Parietal-occipital	Delta	-2.15*
MRE > FRE	Centroparietal	Delta	-2.11*
NAS > MRE	Frontocentral	Gamma	-2.05*

(b) Decision-Confidence Neural Correlates

Comparisons	Brain Regions	Frequency	Z-Score
FRE > MRE	Temporoparietal	Beta	-3.01*
NAS > MRE	Parietal	Beta	-2.87**
NAS > MRE	Temporoparietal	Beta	-2.24*
MRE > FRE	Centroparietal	Beta	-2.11*
MRE > FRE	Frontocentral	Beta	-2.05*

(d) Decision-Confidence Physiological Correlates
(Mean SCR per condition)

(c) Self-reported Decision-Confidence
(Mean per condition, 5-point Likert scale)



Notes: NAS: No AI support; MRE: Minimal reconstructive explanations; FRE: Fully reconstructive explanations; SCR: Skin conductance responses; *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

Fig. 4 Experiment 1:Neural correlates of analytical reasoning and decision-confidence in the executive phase

decision-making within deontologically-bound contexts. Our findings enrich knowledge with nuanced insights into these processes.

During the deliberative decision-making phase (resumes assessment), NAS elicited the greatest neural activation across brain regions and band frequencies associated with sustained attention, selective attention, and information and semantic reasoning. This pattern supports the hypothesis that, without access to external reasoning, greater demands are placed on the neural processes supporting critical thinking, requiring greater engagement of attentional, integrative, and evaluative cognitive processes (H1a).

Interestingly, no significant differences were observed between minimal and fully reconstructive explanations in these neural markers (H1b, not supported), suggesting that any AI-generated reconstructive explanation may partially substitute for internal evaluative processes during initial evidence appraisal.

However, despite this convergence in neural activity, the observed significant decrease in reading time across conditions (H3a) supports the hypothesis that such minimal reconstructive explanations may facilitate more efficient predictive encoding processes, allowing internal schemas to be more readily drawn upon and reducing the need for extensive critical analysis or interpretation of AI-generated reasoning. Consequently, diminishing the extent to which critical thinking was required during the deliberative phase.

Conversely, during the executive phase (i.e., candidate selection), explanation completeness distinctly influenced analytical reasoning. Minimal reconstructive explanations were associated with greater activation of neural correlates associated with independent, criteria-driven judgement compared to fully reconstructive explanations (H1d), while the absence of AI support required the maximum cognitive effort and application of these analytical resources (H1c). Furthermore, while it was posited that minimal reconstructive explanations would facilitate cognitive fluency, and thereby reduce decision times (H3b, not supported), the behavioral data instead point to a non-significant trend toward longer decision times. Rather than indicating inefficiency, this result may reflect a deeper engagement in analytical reasoning associated with systematic processing (Evans, 2008), consistent with the demands of critical thinking.

With respect to executive-phase decision-confidence, minimal reconstructive explanations were associated with greater experienced confidence in judgement accuracy compared to either NAS or fully reconstructive explanations (H2a). This was reflected in a concurrent increase and decrease in neural markers: increased neural activity associated with certainty, judgment consolidation, and confident engagement, and decreased cognitive effort spent to manage doubt and uncertainty in the decision process. Physiological data further reflect these differences, with arousal following a curvilinear pattern across conditions, reaching its lowest with minimal reconstructive explanations, suggesting lower anticipatory stress and greater decision-confidence (H2b, partially supported). In contrast, self-reported decision-confidence increased most with fully reconstructive explanations (H2c, partially supported). This dissociation suggests that minimal reconstructive explanations may better support the reasoning processes that ground experienced confidence in analytical rigor, rather than uncritical overconfidence, potentiated by fully reconstructive explanations.

No significant differences were found in selective reliance between explanation conditions (H3c, not supported). This finding suggests that, despite observed differences in critical thinking and decision-confidence, explanations alone may not suffice to influence the selective uptake of consistently reliable AI recommendations.

Overall, these findings highlight the nuanced ways in which the level of completeness of reconstructive explanations influences critical thinking and decision-confidence when AI aligns with deontological standards. However, these effects did not extend to observable changes in selective reliance, suggesting that explainability alone may be insufficient to ensure selective uptake of AI recommendations.

These findings thus raise the question: *To what extent does AI reliability further support or constrain the effects of explainability on critical thinking?* Experiment 2 extends this inquiry by examining the combined roles of varying levels of reconstructive explanation completeness and reliability.

5 Experiment 2: Explainability, Reliability, and Critical Thinking

When AI outputs are consistently reliable, minimal reconstructive explanations may foster critical thinking by supporting greater analytical reasoning and experienced decision-confidence in the executive phase of decision-making compared to fully reconstructive explanations. However, their effectiveness may change under conditions of epistemic uncertainty, where the reliability of AI outputs is ambiguous and cannot be readily determined.

Under such conditions, we posit that minimal reconstructive explanations may act as epistemic prompts, stimulating the activation of domain-relevant reasoning schemas to resolve uncertainty and fill inferential gaps, and thus fostering more effortful deliberation. By contrast, fully reconstructive explanations, by virtue of their completeness and coherence, may increase persuasive appeal and attenuate critical scrutiny (Danry et al., 2025), thereby serving a rhetorical rather than epistemic function, appearing meaningful while remaining indifferent to truth (Frankfurt, 2005).

Experiment 2 tests this premise by systematically examining whether the value of explanation completeness is contingent on variations in AI reliability. Specifically, we compare contexts of epistemic uncertainty with those of manifest unreliability, where errors are prominent and readily identifiable.

5.1 Experiment 2: Research Hypotheses

5.1.1 AI Reliability and Critical Thinking Under Epistemic Uncertainty (H4-H5)

Reliability is the property of AI to consistently operate in a predictable, robust, and accurate manner under varying conditions, while adapting to new data and environments (Bansal et al., 2021). Maintaining such reliability is especially challenging as AI is probabilistic in nature, continuously evolves and may, unpredictably, alter how it interprets information or generates recommendations, often without users' awareness (Schuetz & Venkatesh, 2020). This adaptability introduces epistemic uncertainty, that is, ambiguity or lack of certainty regarding the accuracy of AI outputs (Hudon et al., 2021; Jiang et al., 2022).

Epistemic uncertainty is often recognized as a catalyst for cognitive engagement, stimulating inquiry and reflective thought (Dewey, 1933; Muis et al., 2021). This effect is particularly evident in the way relative or dynamic changes in perceived certainty trigger changes in thinking processes from fast, intuitive heuristics to slower, more systematic and effortful reasoning (Christensen & Ball, 2018).

Building on these insights, we posit that during the deliberative phase of AI-assisted decision-making, epistemic uncertainty compels greater sustained and selective attention, more extensive information processing, and increased semantic integration to identify subtle inconsistencies and resolve ambiguity. Thereby resulting in prolonged engagement with the evidence and an intensified search for coherence between new information and existing knowledge schemas. Conversely, under manifest unreliability, cognitive effort is minimized as errors are apparent and require less deliberation.

> H4a. Attentional-semantic processing will be greater under epistemic uncertainty than under manifest unreliability, regardless of the level of explanation completeness.
>
> H4b*[2].Attentional-semantic processing will vary depending on the level of explanation completeness, regardless of epistemic uncertainty or manifest unreliability.
>
> H4c*. The influence of epistemic uncertainty versus manifest unreliability on attentional-semantic processing will vary depending on the level of explanation completeness.

Furthermore, we expect that during the executive phase, residual uncertainty from the deliberative phase will elevate the need for analytical reasoning. Here, epistemic uncertainty amplifies the processes required for judgment consolidation, including the deliberate weighing of alternatives, the rigorous evaluation of competing explanations, and the inhibition of premature or intuitive responses. Accordingly, we posit:

> H5a. Analytical reasoning will be greater under epistemic uncertainty than under manifest unreliability, regardless of the level of explanation completeness.
>
> H5b*. Analytical reasoning will vary depending on the level of explanation completeness, regardless of epistemic uncertainty or manifest unreliability.
>
> H5c*. The influence of epistemic uncertainty versus manifest unreliability on analytical reasoning will vary depending on the level of explanation completeness.

5.1.2 Appropriate Reliance (H6)

Reliance centers on epistemic certainty in the dependability of an entity and plays an indefeasible role in guiding reasoning and action (Kleinig, 2016). As a condition for human agency, the possibility of acting together presupposes the capacity to appropriately rely on others' intentions and actions (Alonso, 2009). Within this framework, critical thinking provides the reflective cognitive mechanisms for determining when appropriate reliance is warranted, through reasoning and evidence-based evaluation (Kleinig, 2016).

In AI-assisted decision-making, under conditions of epistemic uncertainty, reliance is the process of engaging or disengaging with AI (Lee & See, 2004), reflecting the judgment involved in determining when to defer to the system and when to retain control. In this context, achieving appropriate reliance reflects the normative behavior of endorsing correct AI-generated recommendations, while rejecting incorrect ones, and acting upon this discrimination (Schemmer et al., 2023). Rather than passive compliance, appropriate reliance reflects the culmination of critical thinking across the deliberative and executive phases, integrating attentional-semantic appraisal, analytical reasoning, and the inhibition of uncritical acceptance.

> H6a. Appropriate reliance will be greater under epistemic uncertainty than under manifest unreliability, regardless of the level of explanation completeness.
>
> H6b*. Appropriate reliance will vary depending on the level of explanation completeness, regardless of epistemic uncertainty or manifest unreliability.
>
> H6c*. The influence of epistemic uncertainty versus manifest unreliability on appropriate reliance will vary depending on the level of explanation completeness.

5.1.3 Reflective Outcomes: Decision-Confidence and Calibrated Trust (H7-H8)

Critical thinking extends beyond decision execution to influence the reflective appraisal of one's judgement (decision-confidence) and the system that informed it (trust). Whereas post-decision-confidence captures perceived accuracy in one's judgment, trust reflects the belief that an agent will help achieve one's goal amid conditions of uncertainty and vulnerability (Lee & See, 2004, p.51).

Both decision-confidence and trust require calibration to reflect the level of warrantedness provided by available evidence and the context in which they are placed, rather than

[2] Asterisked hypotheses are exploratory, investigating the main effect of explanation completeness and its interaction with reliability.

being blind or excessive (Cohen et al., 2022; Gefen et al., 2008; Kleinig, 2016). Critical thinking supports this calibration by fostering systematic reasoning, evidence evaluation, and anticipation of potential outcomes (Marin & Copeland, 2024), rather than the reflexive heuristic judgment, leading to a more accurate assessment of one's judgment and the trustworthiness of AI-generated reasoning.

> H7a. Decision-confidence will be better calibrated under epistemic uncertainty than under manifest unreliability, regardless of the level of explanation completeness.

> H7b*. Calibrated decision-confidence will vary depending on the level of completeness of explanation, regardless of epistemic uncertainty or manifest unreliability.

> H7c*. The influence of epistemic uncertainty versus manifest unreliability on calibrated decision-confidence will vary depending on the level of explanation completeness.

> H8a. Trust will be better calibrated under epistemic uncertainty than under manifest unreliability, regardless of the level of explanation completeness.

> H8b*. Calibrated trust will vary depending on the level of explanation completeness, regardless of epistemic uncertainty or manifest unreliability.

> H8c*. The influence of epistemic uncertainty versus manifest unreliability on calibrated trust will depend on the level of explanation completeness.

5.2 Experiment 2: Methodology

5.2.1 Participants

We recruited 66 active HR professionals, all registered members of a recognized professional order (aged 22–58; $M=36$ years, $SD=8.1$; 73% female; with an average of 12 years of experience). Experiment 2 lasted approximately 30 min, including consent, setup, scenario, and task completion. All participants also completed Experiment 3 during the same session with a counterbalanced order (see Experiment 3: Methodology). Participants received CAD100 compensation for the complete session. Both experiments were conducted in the same laboratory under the same ethics approval certificate as Experiment 1. All participants provided written informed consent.

5.2.2 Experimental Task and Design

Participants completed a scenario-based Wizard-of-Oz recruitment task similar to Experiment 1, with two changes: (1) all candidate assessments (deliberative phase) were AI-assisted, and (2) participants reviewed individual resumes rather than pairs and decided whether to select each candidate or not on the subsequent screen.

The experiment used a 2 × 2 mixed design. Explainability (minimal vs. fully reconstructive) was the between-subjects manipulation. Reliability (reliable vs. unreliable) was manipulated within subjects. Participants were randomly assigned to explanation conditions and completed 20 counterbalanced trials (see Table 2). Explainability conditions were operationalized as in Experiment 1. Reliability was manipulated by varying the accuracy of AI-generated recommendations and scores (reliable vs. unreliable) (see Fig. 5).

To operationalize epistemic uncertainty and simulate real-world scenarios where AI outputs appear generally reliable, but individual output accuracy remains uncertain, participants encountered mostly reliable trials (15 out of 20), interspersed with a smaller number of unreliable trials (5 out of 20). Manifest unreliability was introduced at the trial level through outputs containing clear, identifiable errors. The task procedure was identical to Experiment 1.

5.2.3 Stimuli Design

Stimuli design followed Experiment 1, with, in this case, 20 unique resumes in a standardized one-column format. As in Experiment 1, reliable AI recommendations and scores were generated using the Preference-Dependent Measure (Aksoy et al., 2011). In the unreliable condition, recommendations and scores were intentionally inaccurate, but explanations

Table 2 Experiment 2: Experimental conditions

Explainability	Reliability	
	Reliable recommendations	Unreliable recommendations
Minimal reconstructive explanations	15 trials	5 trials
Fully reconstructive explanations	15 trials	5 trials

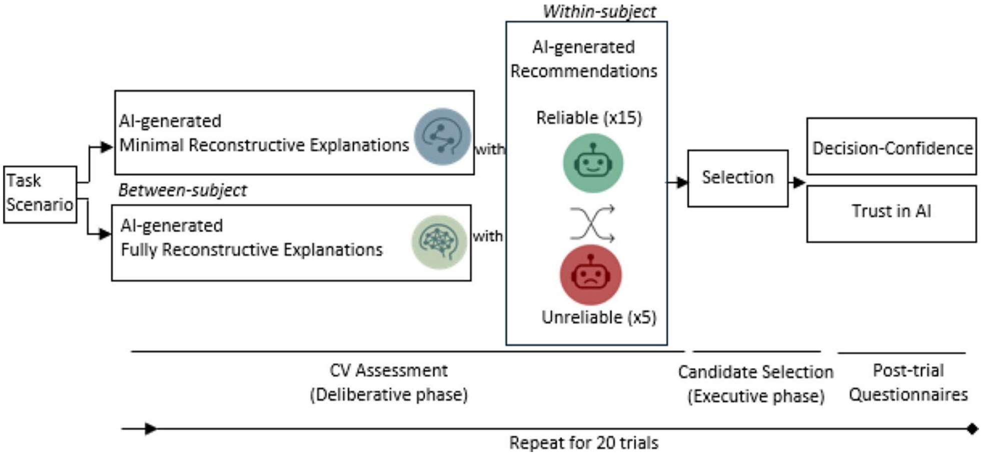

Within-subject
AI-generated Recommendations
Reliable (x15)
Unreliable (x5)
Task Scenario
AI-generated Minimal Reconstructive Explanations
with
Between-subject
AI-generated Fully Reconstructive Explanations
with
Selection
Decision-Confidence
Trust in AI
CV Assessment (Deliberative phase)
Candidate Selection (Executive phase)
Post-trial Questionnaires
Repeat for 20 trials

Fig. 5 Experiment 2: Experimental design

remained factually correct and thus did not justify the inaccurate outputs. All recommendations were counterbalanced across conditions.

5.2.4 Measures

Appropriate reliance was measured as the proportion of trials in which participants' final decisions aligned with accurate AI recommendations or diverged from inaccurate ones (Wang & Yin, 2021). After each trial, self-reported decision-confidence was measured using a single item from Inbar et al. (2011) and trust in the AI recommendation was assessed with a single item from Chen et al. (2020) (see Appendix, Table 6). For each participant, calibrated decision-confidence and trust were indexed as the average absolute difference between confidence/trust ratings and appropriate reliance scores across trials and conditions, such that lower values reflected better-calibrated confidence and trust (Merritt et al., 2015).

Reading time and decision time were measured as in Experiment 1. To capture attentional-semantic processing, we computed an index by multiplying each participant's average reading time by their appropriate reliance scores across trials and conditions, with higher values reflecting increased cognitive engagement and more accurate evidence appraisal. To capture analytical reasoning, we computed an index by multiplying each participant's average decision time by their appropriate reliance scores, such that higher values reflected increased analytical reasoning in support of accurate judgment formation.

All self-report items were adapted to the task context, translated into French using independent double translation and back-translation, and pretested with a monolingual sample (see Appendix, Table 6). Three random attention checks were inserted throughout the 20 trials to ensure response validity (Shamon & Berning, 2020).

5.2.5 Procedure

After providing informed consent, participants reviewed the task scenario and were asked to verbally recall the three recruitment criteria to ensure comprehension. Upon task completion, participants completed a sociodemographic questionnaire. All experimental stimuli and questionnaires were administered using Qualtrics (Provo, UT, USA).

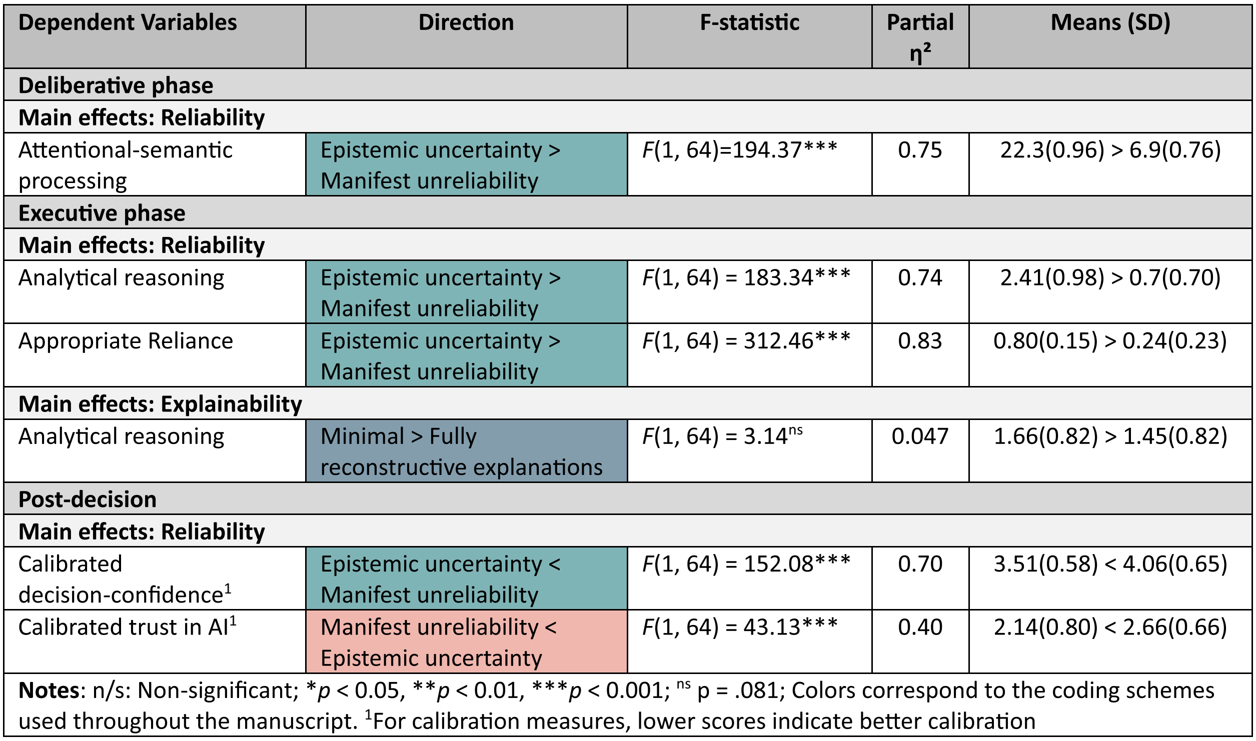
5.3 Experiment 2: Results

A 2 (Explainability: between-subjects; minimal vs. fully reconstructive) x 2 (Reliability: within-subjects; reliable vs. unreliable) mixed-design repeated measures ANOVA (Bonferroni corrected) was conducted to test main and interaction effects on all dependent variables (see Table 3). Analyses used SPSS (v28) with significance levels set at $p<0.05$.

No statistically significant interaction effects were observed between explanation completeness and AI reliability. No main effects of explanation completeness were found on attentional-semantic processing (H4b-c), analytical reasoning (H5c), appropriate reliance (H6b-d), calibrated decision-confidence (H7b-c), or trust calibration (H8b-c). A marginally non-significant main effect of explanation completeness was observed for analytical reasoning (H5b), with higher scores under minimal reconstructive explanations than fully reconstructive explanations ($F(1, 64)=3.14$, $p=0.081$, partial $\eta^2 = 0.047$).

Conversely, we observed strong main effects of AI reliability on all dependent variables. Attentional-semantic processing, analytical reasoning, and appropriate reliance were significantly higher under epistemic uncertainty than under manifest unreliability, supporting H4a, H5a, and H6a ($F(1, 64)=194.37$, $p<0.001$, partial $\eta^2 = 0.75$; $F(1, 64)=183.34$, $p<0.001$, partial $\eta^2 = 0.74$; $F(1, 64)=312.46$, $p<0.001$, partial $\eta^2 = 0.83$), respectively. Decision-confidence was more calibrated under epistemic uncertainty ($F(1, 64)=152.08$,

Table 3 Experiment 2: Effects of AI reliability and explainability on critical thinking across decision-making phases



Dependent Variables	Direction	F-statistic	Partial η²	Means (SD)
Deliberative phase				
Main effects: Reliability				
Attentional-semantic processing	Epistemic uncertainty > Manifest unreliability	*F*(1, 64)=194.37***	0.75	22.3(0.96) > 6.9(0.76)
Executive phase				
Main effects: Reliability				
Analytical reasoning	Epistemic uncertainty > Manifest unreliability	*F*(1, 64) = 183.34***	0.74	2.41(0.98) > 0.7(0.70)
Appropriate Reliance	Epistemic uncertainty > Manifest unreliability	*F*(1, 64) = 312.46***	0.83	0.80(0.15) > 0.24(0.23)
Main effects: Explainability				
Analytical reasoning	Minimal > Fully reconstructive explanations	$F(1, 64) = 3.14^{ns}$	0.047	1.66(0.82) > 1.45(0.82)
Post-decision				
Main effects: Reliability				
Calibrated decision-confidence[1]	Epistemic uncertainty < Manifest unreliability	*F*(1, 64) = 152.08***	0.70	3.51(0.58) < 4.06(0.65)
Calibrated trust in AI[1]	Manifest unreliability < Epistemic uncertainty	*F*(1, 64) = 43.13***	0.40	2.14(0.80) < 2.66(0.66)

Notes: n/s: Non-significant; *p < 0.05, **p < 0.01, ***p < 0.001; ^{ns} p = .081; Colors correspond to the coding schemes used throughout the manuscript. [1]For calibration measures, lower scores indicate better calibration

n/s: Non-significant; *p<0.05, **p<0.01, ***p<0.001; ^{ns}p=0.081; Colors correspond to the coding schemes used throughout the manuscript. [1]For calibration measures, lower scores indicate better calibration

p<0.001, partial η² = 0.70), consistent with H7a. However, contrary to H8a, trust was better calibrated under manifest unreliability (F(1, 64)=43.13, p<0.001, partial η^2 = 0.403).

5.4 Experiment 2: Discussion

Experiment 2 investigated whether the influence of reconstructive explanation completeness on stimulating critical thinking is contingent upon the reliability of AI outputs.

Our findings suggest that AI reliability was the dominant driver of the cognitive processes underlying critical thinking and their reflective outcomes across all phases of decision-making, whereas the anticipated effects of explanations were not statistically significant.

Specifically, a key distinction emerged in the contextual variation of AI reliability: compared to manifest unreliability, epistemic uncertainty was associated with greater attentional-semantic processing, analytical reasoning, and appropriate reliance. These findings support the hypothesis that, when the reliability of AI-generated outputs cannot be readily determined, more sustained attention and semantic integration are required to resolve ambiguity and evaluate evidentiary coherence during the deliberative phase of decision-making. This carries forward into judgment formation, which prompts more systematic evaluation and accurate weighing of alternatives during the executive phase in support of sound decision-making. The increased systematic evaluation and accuracy of decision-making were further reflected in better-calibrated decision-confidence, suggesting that the critical thinking processes activated across both phases support more accurate reflective appraisal of one's judgment.

Conversely, trust calibration diverged from expectations, with more calibrated trust toward AI recommendations emerging under manifest unreliability than under epistemic uncertainty (contrary to H8a), potentially suggesting disengagement from the system in the face of obvious errors, rather than the more effortful, reflective calibration process hypothesized to occur under conditions of epistemic uncertainty.

Furthermore, although explanation completeness did not exert significant main or interaction effects, a marginal non-significant trend toward greater analytical reasoning under minimal reconstructive explanations (H5b; p=0.081) echoes findings from Experiment 1, whereby such explanations were associated with increased activation of neural correlates supporting independent judgment. This may suggest that minimal explanations, while not significantly influencing deliberative phase processes, may potentially continue to facilitate greater reflective engagement during the executive phase under epistemic uncertainty.

Overall, these findings suggest a reorientation of critical thinking, from the reasoning underlying AI-generated outputs to system-level appraisal, in response to variations in AI reliability. Thus, under epistemic uncertainty, explanations may no longer guide content-level engagement; instead, uncertainty itself serves as a cognitive prompt, eliciting increased scrutiny during evidence appraisal and a more deliberate evaluation of system reliability during the formation and execution of sound judgment, though minimal reconstructive explanations may still modestly support analytical reasoning, irrespective of reliability. In such contexts, minimal reconstructive explanations alone may be insufficient to sustain critical thinking, suggesting that AI transparency may help by reducing uncertainty and restoring the benefits of explainability.

This raises a further question: *To what extent does AI transparency sustain the effects of explainability on critical thinking when AI reliability is uncertain?* Experiment 3 extends this inquiry by examining the combined roles of varying levels of reconstructive explanation, reliability, and transparency.

6 Experiment 3: Explainability, Reliability, Transparency, and Critical Thinking

Experiment 2 established that variations in AI reliability redirect critical thinking away from the reasoning underlying AI-generated outputs toward system-level appraisal, thereby constraining the influence of the level of completeness of reconstructive explanations (minimal vs. fully) on critical thinking.

Experiment 3 extends this investigation by introducing AI transparency to test whether the provision of confidence scores can sustain or restore the value of reconstructive explanation completeness under conditions of epistemic uncertainty. To test this premise, Experiment 3 manipulated three factors: explainability (between-subjects: minimal vs. fully reconstructive), reliability (within-subjects: reliable vs. unreliable), and transparency (within-subjects: confidence scores vs. no confidence scores).

6.1 Experiment 3: Research Hypotheses

Transparency involves making AI's processes and internal mechanisms accessible and visible to the intended users (Doran et al., 2017). In this study, transparency is operationalized as the provision of probabilistic recommendation confidence scores, offering explicit information about the system's internal certainty for each output (Le et al., 2023). We posit that, in doing so, transparency may function as a metacognitive cue, promoting deeper engagement with AI-generated reasoning (explanations) across cognitive domains. By signaling internal coherence certainty, confidence scores may facilitate attentional-semantic processing (Eva & Regehr, 2005), analytical reasoning (Eva & Regehr, 2005) and improve decision-confidence and trust calibration (Kizilcec, 2016). Alternatively, transparency may further redirect critical thinking away from content-level reasoning toward system-level appraisals, diminishing the influence of reconstructive explanations altogether.

By systematically manipulating explanation completeness, reliability, and transparency, Experiment 3 directly tests these competing accounts and seeks to delineate the boundary conditions under which explainability meaningfully supports critical thinking in AI-assisted decision-making.

Building on the findings of Experiment 2, we posit:

> H9a-e. The influence of explanation completeness (minimal versus fully reconstructive) on (a) attentional-semantic processing, (b) analytical reasoning, (c) appropriate reliance, (d) calibrated decision-confidence, and (e) calibrated trust in AI will vary depending on whether a confidence score is present or absent.

> H10a-e. The influence of epistemic uncertainty versus manifest uncertainty on (a) attentional-semantic processing, (b) analytical reasoning, (c) appropriate reliance, (d) calibrated decision-confidence, and (e) calibrated trust in AI will vary depending on whether a confidence score is present or absent.

> H11a-e*[3]**.** The influence of explanation completeness on (a) attentional-semantic processing, (b) analytical reasoning, (c) appropriate reliance, (d) calibrated decision-confidence, and (e) calibrated trust in AI will vary depending on the combined effects of epistemic uncertainty or manifest unreliability and the presence or absence of a confidence score.

> H12a-e: (a) Attentional-semantic processing, (b) analytical reasoning, (c) appropriate reliance, (d) calibrated decision-confidence, and (e) calibrated trust in AI will be greater in the presence of a confidence score than its absence, regardless of explanation completeness or reliability.

> H13a-e: (a) Attentional-semantic processing, (b) analytical reasoning, (c) appropriate reliance, and (d) calibrated trust in AI scores will be greater while (e) calibrated decision-confidence scores will be lower under epistemic uncertainty than manifest uncertainty, regardless of explanation completeness or transparency.

[3] Asterisked hypotheses are exploratory, investigating the main and interaction effects of explanation completeness, reliability and transparency.

6.2 Experiment 3: Methodology

6.2.1 Participants

The same 66 h professionals from Experiment 2 completed Experiment 3 during the same session, after a 15-minute break between the two. Experiment order was counterbalanced across participants with half (N=33) completing Experiment 2 first and half (N=33) completing Experiment 3 first. The approximately 30-minute study included consent, setup, scenario reading, and task completion. The session concluded with sociodemographic questionnaires.

6.2.2 Experimental Task and Design

Following Experiment 2, participants completed a scenario-based Wizard-of-Oz (Riek, 2012) AI-assisted recruitment task, this time pre-selecting Web Project Manager candidates, according to three revised criteria: project management skillsets, experience, and education.

The experiment used a 2 × 2 × 2 repeated-measures mixed design. Explainability (minimal vs. fully reconstructive) was the between-subjects manipulation. Reliability (reliable vs. unreliable) and transparency (confidence scores vs. no confidence scores) were manipulated within subjects. Participants were randomly assigned to one of the explainability conditions and completed 22 counterbalanced trials (see Table 4). Explainability and reliability conditions were operationalized as in previous experiments. Transparency was manipulated by including or excluding model confidence scores. The task procedure was identical to Experiment 2.

6.2.3 Stimuli Design

Stimuli design was otherwise identical to Experiment 2, with 22 unique resumes. In transparent conditions, the model confidence score was always set above 50%, since values below chance would signal that the AI's recommendation was likely incorrect and would not provide a meaningful reference point for participants (see Fig. 6). All recommendations were counterbalanced across conditions.

6.2.4 Measures

All behavioral and self-report measures were identical to those used in Experiment 2. Three random attention checks were inserted throughout the 22 trials to ensure response validity (Shamon & Berning, 2020).

6.3 Experiment 3: Results

A 2 (Explanation: between-subjects; minimal vs. fully reconstructive) x 2 (Reliability: within-subjects; reliable vs. unreliable) x 2 (Transparency: within-subjects; confidence scores vs. no-confidence scores) mixed-design repeated measures ANOVA (Bonferroni corrected) was conducted to test main and interaction effects on all dependent variables. Analyses used SPSS (v28) with a significance level set a*t* $p<0.05$.

No statistically significant interaction effects were observed between explanation completeness and transparency, nor between explanation completeness, AI reliability, and transparency, on any of the dependent variables, providing no support for H9a-e or H11a-e.

Significant interaction effects between AI reliability and transparency were observed for appropriate reliance and calibrated decision-confidence. Appropriate reliance was higher under epistemic uncertainty than manifest unreliability, with this difference being greater when confidence scores were present (F(1, 64)=17.27, $p<0.001$, partial η^2 = 0.21), supporting H10c. Conversely, decision-confidence was better calibrated under epistemic uncertainty than manifest unreliability, and this difference was greater when confidence scores were absent (F(1, 64)=9.38, p=0.003, partial η^2 = 0.13), supporting H10d. Furthermore, analytical reasoning was marginally significantly greater under epistemic uncertainty than manifest unreliability, with this difference being greater when confidence scores were absent (F(1, 64)=3.54, p=0.064, partial η^2 = 0.052) (H10b, not supported). No significant interaction effects were observed for attentional-semantic processing (H10a) or calibrated trust (H10e).

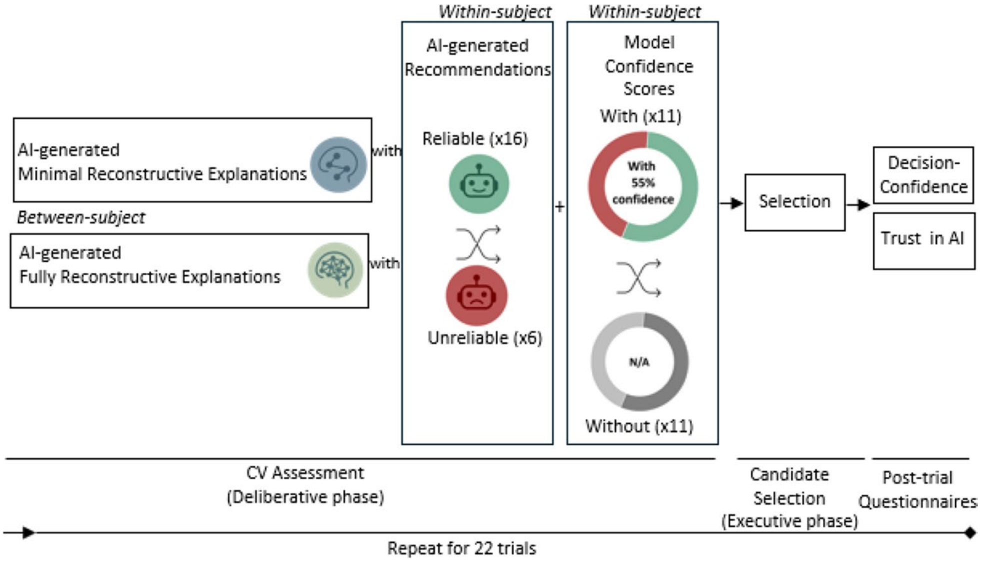
A marginally non-significant main effect of transparency was observed for decision-confidence, with slightly

Table 4 Experiment 3: Experimental conditions

Expainabilityl	Reliability	Transparency	
		With model confidence score	Without model confidence score
MRE	Reliable recommendation	8 trials	8 trials
	Unreliable recommendation	3 trials	3 trials
FRE	Reliable recommendation	8 trials	8 trials
	Unreliable recommendation	3 trials	3 trials

MRE: Minimal reconstructive explanations; *FRE:* Fully reconstructive explanations

Fig. 6 Experiment 3: Experimental Design


Within-subject
AI-generated Recommendations
Reliable (x16)
Unreliable (x6)
Within-subject
Model Confidence Scores
With (x11)
With 55% confidence
N/A
Without (x11)
AI-generated Minimal Reconstructive Explanations
with
Between-subject
AI-generated Fully Reconstructive Explanations
with
+
Selection
Decision-Confidence
Trust in AI
CV Assessment (Deliberative phase)
Candidate Selection (Executive phase)
Post-trial Questionnaires
Repeat for 22 trials

better-calibrated confidence when confidence scores were present (F(1, 64) = 3.52, p = 0.065, partial η^2 = 0.05) (H12d, not supported). No significant main effects of transparency were found for attentional-semantic processing (H12a), analytical reasoning (H12b), appropriate reliance (H12c), or calibrated trust (H12e).

Strong main effects of AI reliability were observed across all dependent variables. Attentional-semantic processing, analytical reasoning, appropriate reliance and calibrated trust scores were all significantly higher under epistemic uncertainty than under manifest unreliability, supporting H13a-d (F(1, 64) = 234.73, p < 0.001, partial η^2 = 0.79; F(1, 64) = 233.95, p < 0.001, partial η^2 = 0.79; F(1, 64) = 1055.61, p < 0.001, partial η^2 = 0.94; F(1, 64) = 39.48, p < 0.001, partial η^2 = 0.38, respectively). Furthermore, and supporting H13e, calibrated decision-confidence scores were lower under epistemic uncertainty (F(1, 64) = 249.92, p < 0.001, partial η^2 = 0.80) (Table 5).

6.4 Experiment 3: Discussion

Experiment 3 explored whether transparency might restore the influence of explanation completeness on critical thinking and related outcomes, particularly in the face of varying AI reliability.

Contrary to our initial theorization, our findings suggest that the addition of model confidence scores (i.e., transparency) did not restore or increase the influence of reconstructive explanations across attentional-semantic processing, analytical reasoning, appropriate reliance, calibrated decision-confidence, and trust in AI.

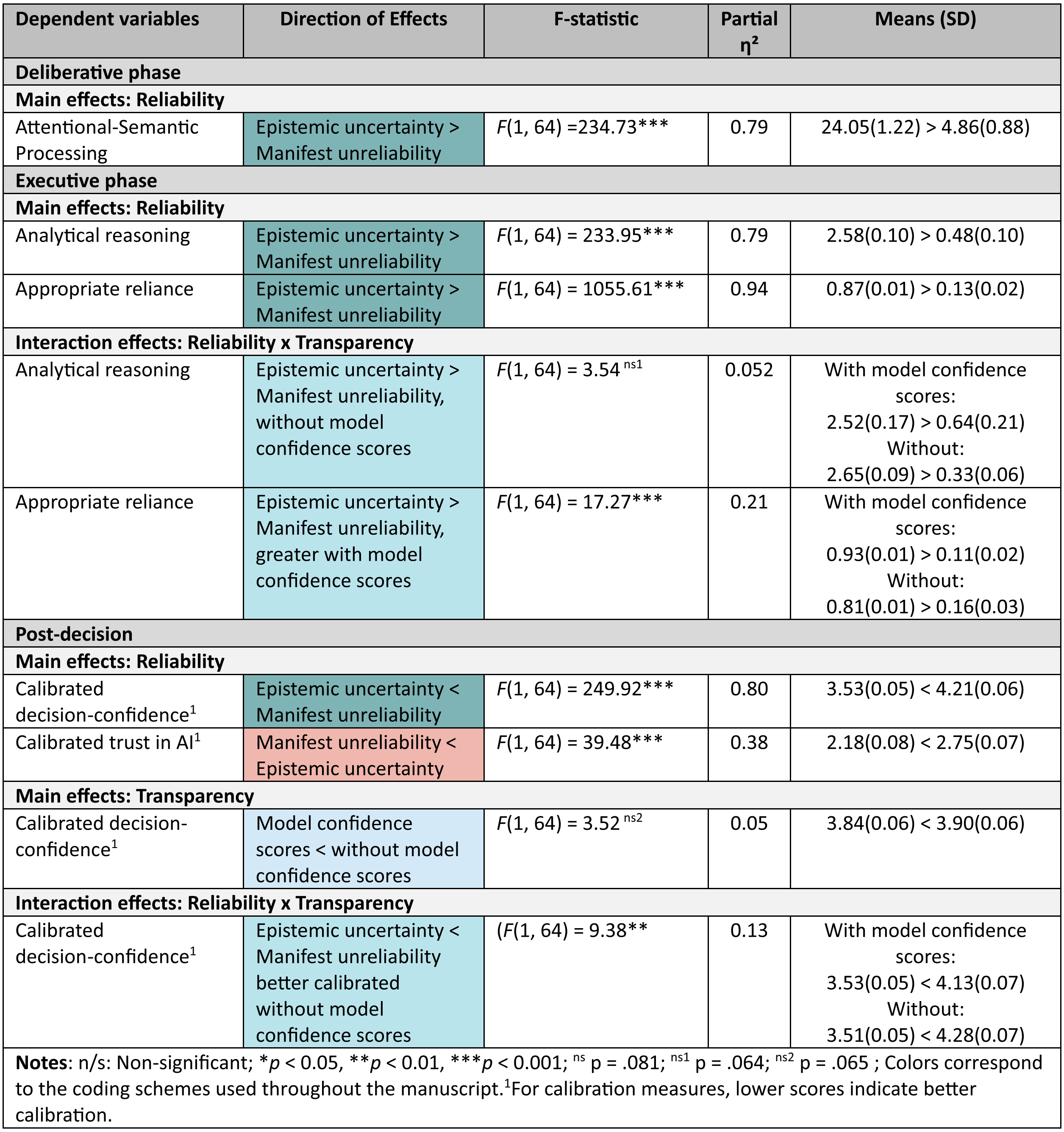
Instead, and consistent with Experiment 2, AI reliability remained the dominant driver of critical thinking, shaping its cognitive and behavioral outcomes. Epistemic uncertainty continued to be associated with greater attentional-semantic processing, analytical reasoning, appropriate reliance, and better-calibrated decision-confidence. Moreover, manifest unreliability was associated, as anticipated, with more calibrated trust in AI, further supporting the observed effects in Experiment 2.

Furthermore, whereas transparency did not exert any significant main effects, a marginal non-significant trend was observed toward better-calibrated decision-confidence (H12d; p = 0.065, partial η^2 = 0.05), suggesting that transparency may improve judgment appraisal.

Building on these main effects, we identified important nuances in how transparency and reliability interact. Under epistemic uncertainty, showing model confidence scores led to a modest trend toward reduced analytical reasoning (H10b; p = 0.064, partial η^2 = 0.052), but greater appropriate reliance on AI recommendations (H10c, supported). This may suggest that making AI's internal confidence explicit serves as a heuristic cue that facilitates appropriate reliance rather than prompting more critical engagement with uncertain recommendations. Supporting this pattern, model confidence scores reduced the calibration benefits seen under epistemic uncertainty, suggesting that transparency may interfere with the systematic self-evaluation processes that enhance calibration under uncertainty.

In sum, findings from Experiment 3 do not support the assumption that transparency can restore the influence of minimal reconstructive explanations on critical thinking when AI reliability varies. Instead, epistemic uncertainty remained the dominant driver of critical thinking, with transparency showing only selective, context-dependent effects. Thus, in this context, transparency appears to serve as a situational heuristic cue, rather than a consistent enabler of the positive effect of minimal reconstructive explanations on critical thinking.

Table 5 Experiment 3: Effects of AI reliability & transparency on critical thinking across decision-making phases



Dependent variables	Direction of Effects	F-statistic	Partial η²	Means (SD)
Deliberative phase				
Main effects: Reliability				
Attentional-Semantic Processing	Epistemic uncertainty > Manifest unreliability	*F*(1, 64) =234.73***	0.79	24.05(1.22) > 4.86(0.88)
Executive phase				
Main effects: Reliability				
Analytical reasoning	Epistemic uncertainty > Manifest unreliability	*F*(1, 64) = 233.95***	0.79	2.58(0.10) > 0.48(0.10)
Appropriate reliance	Epistemic uncertainty > Manifest unreliability	*F*(1, 64) = 1055.61***	0.94	0.87(0.01) > 0.13(0.02)
Interaction effects: Reliability x Transparency				

Analytical reasoning	Epistemic uncertainty > Manifest unreliability, without model confidence scores	*F*(1, 64) = 3.54 ^{ns1}	0.052	With model confidence scores:
				2.52(0.17) > 0.64(0.21)
				Without:
				2.65(0.09) > 0.33(0.06)

Appropriate reliance	Epistemic uncertainty > Manifest unreliability, greater with model confidence scores	*F*(1, 64) = 17.27***	0.21	With model confidence scores:
				0.93(0.01) > 0.11(0.02)
				Without:
				0.81(0.01) > 0.16(0.03)

Post-decision				
Main effects: Reliability				
Calibrated decision-confidence[1]	Epistemic uncertainty < Manifest unreliability	*F*(1, 64) = 249.92***	0.80	3.53(0.05) < 4.21(0.06)
Calibrated trust in AI[1]	Manifest unreliability < Epistemic uncertainty	*F*(1, 64) = 39.48***	0.38	2.18(0.08) < 2.75(0.07)
Main effects: Transparency				
Calibrated decision-confidence[1]	Model confidence scores < without model confidence scores	*F*(1, 64) = 3.52 ^{ns2}	0.05	3.84(0.06) < 3.90(0.06)
Interaction effects: Reliability x Transparency				

Calibrated decision-confidence[1]	Epistemic uncertainty < Manifest unreliability better calibrated without model confidence scores	(*F*(1, 64) = 9.38**	0.13	With model confidence scores:
				3.53(0.05) < 4.13(0.07)
				Without:
				3.51(0.05) < 4.28(0.07)

Notes: n/s: Non-significant; *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$; ^{ns} p = .081; ^{ns1} p = .064; ^{ns2} p = .065 ; Colors correspond to the coding schemes used throughout the manuscript.[1]For calibration measures, lower scores indicate better calibration.

n/s: Non-significant; *$p<0.05$, **$p<0.01$, ***$p<0.001$; $^{ns}p=0.081$; $^{ns1}p=0.064$; $^{ns2}p=0.065$; Colors correspond to the coding schemes used throughout the manuscript.[1] For calibration measures, lower scores indicate better calibration.

7 General Discussion

This research was motivated by a central concern in contemporary professional practices: that the distinct nature AI - its autonomy, learning, and inherent inscrutability – may erode the cognitive conditions necessary for professionals to engage in active critical thinking and informed judgment, thereby compromising the responsibility and accountability required by deontological codes to protect the interests of those they serve (Okasha, 1999; Westerholm, 2009).

To better understand these challenges, we investigated how embedding into AI system-level properties of explainability, reliability, and transparency may foster critical thinking in AI-assisted professional contexts. Drawing on cognitive neuroscience, explanation-based reasoning, and heuristic-systematic theories, we conducted three multi-method experiments to clarify how these properties, individually or in combination, influence the neurophysiological, cognitive, and behavioral mechanisms underlying sound judgment. Our findings provide a nuanced understanding of how professionals engage with AI and help delineate the conditions under which these properties can facilitate critical engagement (see Appendix, Table 7 for a comprehensive summary of hypothesis testing results).

7.1 Minimal Reconstructive Explanations Foster Reasoning but only Under Reliable AI

Experiment 1 established a baseline under conditions of consistently reliable AI. In this setting, we examined how two levels of reconstructive causal explanations - explanations that narratively reconstruct the reasoning underlying AI-generated recommendations as post-hoc, non-mechanistic user-level rationales - influence the neural dynamics and behavioral outcomes of critical thinking, compared with no AI support. Both minimal and reconstructive explanations reduced neural markers of sustained, selective attention and semantic information processing integration, suggesting that any form of AI-causal reconstructive narrative may substitute for internal evidence appraisal, compared to making decisions independently. However, minimal reconstructive explanations increased neural correlates of analytical reasoning and physiological markers of experienced decision-confidence. In contrast, fully reconstructive explanations inflated self-reported decision-confidence despite reduced cognitive effort. This asymmetry suggests that minimal reconstructive explanations may support confidence grounded in analytical rigor, whereas fully reconstructive explanations risk fostering unwarranted decision-confidence decoupled from reasoning.

Experiment 2 introduced variable AI reliability to test the boundary conditions of these observed effects. Under these conditions, however, these effects were no longer statistically significant. Instead, epistemic uncertainty emerged as the primary driver of critical thinking. Compared to manifest unreliability, epistemic uncertainty prompted greater attentional and semantic processing, analytical reasoning, more appropriate reliance on AI, and better-calibrated decision-confidence. These findings point to a reorientation from explanation-level engagement to system-level appraisal, with uncertainty serving as a cognitive prompt for greater deliberative evaluative reasoning.

Experiment 3 thus introduced AI transparency, through instance-based AI confidence scores, as a potential mechanism to reinstate the influence of explanation completeness on critical thinking. However, we found minimal support for this conjecture. Epistemic uncertainty, rather than transparency, remained the dominant driver shaping cognitive, behavioral, and reflective outcomes. Compared to conditions without transparency, AI model confidence scores alone did not significantly increase proxies of critical thinking and its related outcomes. Furthermore, under epistemic uncertainty, AI model confidence scores showed mixed effects: reducing analytical reasoning and decision-confidence calibration but improving appropriate reliance, suggesting they serve as heuristic cues that disrupt critical engagement while facilitating reliance decisions under uncertain conditions, eventually failing to restore the earlier benefits of explanation completeness.

7.2 Emerging Metacognitive Process of Critical Thinking in AI-Assisted Professional Decision-Making

Together, these findings point to an emergent metacognitive process wherein critical thinking is not uniformly influenced by AI explainability but rather shaped by the interaction between AI properties and cognitive orientation toward uncertainty, unfolding dynamically across decision-making phases.

In the deliberative phase, engagement with AI outputs involves allocating sustained and selective attention and the evaluation of semantic coherence relative to internal judgment schemas. When epistemic uncertainty is low and AI performance is stable, minimal reconstructive explanations facilitate efficient predictive encoding. By providing skeletal causal frameworks, they enable the activation of internal schemas with less interpretative effort, thus supporting semantic integration while maintaining evaluative control. Fully reconstructive explanations, by contrast, present complete causal reasoning that may preempt the need for semantic alignment, thereby displacing active cognitive effort. Under epistemic uncertainty, however, these observed patterns are reoriented. Uncertainty acts as a metacognitive prompt, increasing attentional and semantic processing regardless of the level of causal reconstructive explanations. Cognitive effort is redirected from content-level reasoning to system-level scrutiny, prompting more deliberate appraisal of AI reliability.

In the executive phase, analytical reasoning and conflict monitoring support judgment consolidation and decision-confidence. When epistemic uncertainty is low, minimal reconstructive explanations again foster greater analytical effort and experienced decision-confidence. Conversely, under epistemic uncertainty, analytical reasoning increases broadly, driven not by causal reconstructive explanations, but by the cognitive demands of resolving ambiguity in AI outputs.

Together, these phases support informed judgment, and with it, appropriate reliance on AI outputs, grounded in reflective evaluation. Crucially, such reliance is not driven by the nature of reconstructive causal explanations but rather arises through active engagement with epistemic uncertainty, whereby transparency may serve as a heuristic cue that bypasses rather than aids in assessing AI reliability.

Reflective appraisal is similarly conditional. Under consistent reliability, fully reconstructive explanations inflate self-reported confidence despite reduced cognitive effort, suggesting an illusion of warranted judgment. Epistemic uncertainty, however, enhances decision-confidence calibration by activating critical thinking processes that support more accurate reflective appraisal of one's judgment. Moreover, trust in AI outputs becomes more discerning under manifest unreliability, as epistemic authority is reoriented from AI outputs toward autonomous evaluative processes.

7.3 Critical Thinking as Epistemic Safeguard in AI-Assisted Professional Contexts

In deontologically-governed professions, responsibility involves acting in accordance with established duties and standards of care, bearing credit or blame for the outcomes of one's actions, while accountability entails justifying and explaining decisions to oversight bodies and those affected (Bivins, 2006; Frankel, 1989). These normative obligations rest on the capacity for deliberate, reflective, and critical thinking (Arendt, 1971; Mauti, 2024), even when assisted by increasingly autonomous and inscrutable technologies.

Within this context, our findings suggest that explainability, reliability, and transparency do not uniformly foster critical thinking. Rather, their influence is conditional on the broader epistemic context in which they are encountered.

Under conditions of consistent AI reliability, minimal reconstructive explanations can support these normative demands. By scaffolding, rather than substituting internal reasoning, such explanations help retain control over judgment and align confidence with analytical effort (Lombrozo, 2011; Wick & Thompson, 1992). This support, however, erodes when reliability becomes variable. In contexts of epistemic uncertainty, the level of completeness of reconstructive causal explanation loses salience, and critical appraisal shifts toward the system itself, prompted by ambiguity (Chaiken & Ledgerwood, 2012).

Transparency, in turn, offers limited reinforcement. Under epistemic uncertainty, model confidence scores disrupt the decision-confidence calibration process and do not restore the benefits of minimal reconstructive explanations, suggesting that transparency may act more as a situational heuristic cue than a standalone support for critical thinking.

We argue that this reorientation, from engaging with explanation content to appraising the system, when otherwise accurate explanations no longer offer a reliable basis for justification due to unreliable outputs, may reflect an effort to preserve control to remain answerable and uphold normative obligations (Pisani & Haw, 2023). Epistemic uncertainty and manifest unreliability may serve as cognitive safeguards, prompting renewed vigilance and a reflective stance, conditions necessary for safeguarding responsibility and accountability (Arendt, 1971; Facione, 1990). In both cases, critical thinking reasserts itself not despite AI fallibility, but because of it.

7.4 Implications for Research

This work contributes to AI explainability research by empirically distinguishing between the implicit and explicit cognitive responses elicited by different levels of causal reconstructive explanations during professional decision-making under varying AI reliability. Under high-reliability conditions, minimal reconstructive explanations support analytical reasoning and decision-confidence, without displacing cognitive evaluative effort. These results suggest that, under favorable epistemic conditions, minimal explanations scaffold coherent interpretation while preserving critical thinking. Under uncertainty, however, the benefits of explainability diminish and engagement shifts to system-level evaluation, with ambiguity itself acting as a metacognitive prompt (Alter et al., 2007). This pattern aligns with emerging perspectives that frame explainability not as a tool for clarity, but as a mechanism for supporting metacognitive control and adaptive reliance (Danry et al., 2025; Morrison et al., 2023).

We further contribute to the growing body of knowledge on XAI in information systems by documenting the circumstances under which the presentation of model confidence scores (i.e., AI transparency) in AI-assisted decision-making shapes appropriate reliance. Prior work has shown that presenting post-hoc calibrated model uncertainty in frequency format can help users better calibrate their reliance on AI compared to probability-based presentation, especially in high-stakes contexts (Cao et al., 2024). Complementing this, our findings with non-calibrated model confidence scores suggest a more nuanced picture: under epistemic uncertainty, such scores may facilitate appropriate reliance decisions, yet appear to function as heuristic cues that prompt users to bypass critical evaluation of choice-relevant information and support less context-sensitive decision-confidence calibration.

Together, these findings add to recent work that challenges the assumption that the most accurate - or more fully explainable and transparent - AI necessarily best supports human decision-making (Bansal et al., 2021; Karran et al., 2024). Specifically, we show that epistemic uncertainty and manifest unreliability serve not only as risk signals but also as cognitive prompts that elicit critical thinking. While model transparency (e.g., instance-based confidence scores) can support appropriate reliance when interacting with uncertainty, it is ultimately the variability in AI performance that more consistently drives critical engagement.

More broadly, this research enhances understanding of how critical thinking emerges from the interaction of higher-order cognitive processes — including sustained attention, semantic integration, and analytical reasoning—coordinated across deliberative and executive decision phases. Building on existing cognitive neuroscience, we develop a novel theoretical account of critical thinking as a dynamically regulated process, supported by distinct neural mechanisms. Our findings suggest that neural correlates, as measured through EEG markers such as changes in brainwave frequency and amplitude, reflect cognitive engagement associated with components of critical thinking and offer reliable, phase-specific indicators of such reflective engagement during AI-assisted decision tasks.

Additionally, this work supports the human-centered explainable AI (HC-XAI) research agenda by examining practicing experienced and certified HR professionals, addressing calls for empirical research that assesses stakeholder-specific needs with explanatory AI systems in applied decision-making contexts (Casalino et al., 2025). Our multi-method approach, combining neurophysiological, behavioral, and self-report measures, strengthens evaluation frameworks for HC-XAI by demonstrating how different explanation types influence cognitive processes across decision-making phases. These findings inform the development of effective explanation mechanisms that consider not only technical accuracy but also the cognitive and ethical requirements of deontologically-governed professional contexts.

7.5 Implications for Design and Practice

Our findings carry important implications for human-centered AI (HCAI) interaction design, whereby designers may wish to consider integrating explainability not only to support comprehension but also to facilitate self-directed, context-sensitive reasoning. In professional settings where decisions must be justified to others, user interfaces should go beyond delivering detailed, ready-made causal explanations. Instead, designers might implement layered, minimal explanations that prompt users to reconstruct the causal reasoning behind AI-generated outputs, encouraging critical evaluation on their own terms.

Furthermore, our results point toward a new set of design principles, in which ambiguity is intentionally surfaced as a means to foster user vigilance and discretionary judgment, rather than a flaw to be eliminated. Rather than over-optimizing to obscure unreliability, design might leverage the inherent uncertainty expressed through visual uncertainty bands or hedging language that signals when extra scrutiny is warranted (Ferson et al., 2015; Kay et al., 2016). However, our findings suggest caution with model confidence scores, which may operate as heuristic cues that reduce, rather than stimulate, critical engagement. User interfaces might also integrate reflective prompts that encourage users to validate AI-generated output and pause before taking action (Karran et al., 2024; Malaguti et al., 2025).

While our findings do not directly address in situ AI literacy interventions, they nevertheless point to opportunities for supporting critical thinking capacities needed to assess the validity of AI-generated reasoning. For example, interfaces could integrate in situ AI literacy interventions, such as logic-based mini-challenges, step-by-step breakdowns of inference chains, or interactive prompts to challenge unsupported conclusions, directly into the workflow (Danry et al., 2025; Panciroli et al., 2023).

7.6 Implications for Early-Career Professional Training

Across our experiments, participants were experienced professionals, averaging approximately twelve years of practice, whose critical thinking with AI likely drew on well-established schematic knowledge and domain-specific expertise (Chi et al., 2014). By contrast, however, early-career professionals, whose cognitive schemata and professional judgment are still developing, may engage with AI differently, raising an important concern about how these individuals will acquire the required domain knowledge and evaluative skills needed to engage critically with AI as it becomes increasingly embedded in professional practice. In line with empirical evidence that structured AI-assisted practice enables novices to develop expert-like patterns of critical engagement and calibrated reliance in high-accountability contexts (Kawakami et al., 2023) our findings point to the value of developing AI-assisted decision-making training environments that deliberately alternate between reliable AI providing minimal reconstructive explanations and variable AI reliability. Such training environments can reinforce analytical reasoning and the associative links that underlie schematic learning, while fostering metacognitive monitoring, and sustaining the cognitive effort required for schematic knowledge development and reflective judgment (Chi et al., 2014; Ericsson, 2004).

8 Limitations and Future Research

Notwithstanding its contributions, we would like to acknowledge that the research presented in this manuscript has several limitations that can be addressed in future research. First, the use of the Wizard-of-Oz methodology in controlled experimental settings may not fully reflect the complexity of real-world interactions with AI. Future research should examine these dynamics in more naturalistic environments, where professionals engage with functioning AI systems embedded within their actual workflows. Second, we conducted multiple experiments to bolster our findings and address threats to generalizability. However, the relatively small sample size in Experiment 1 reduces the positive impact of the observed results. Nonetheless, this sample aligns with prior research and provides adequate power to support the tentative conclusions presented (Boucsein, 1999; Boucsein & Thum, 1997; Riedl et al., 2020). Third, the consecutive administration of Experiments 2 and 3 may have introduced fatigue effects despite counterbalancing and a 15-minute break. Fourth, individual characteristics, such as prior experience with AI, were not captured and might have influenced how professionals engage with AI explanations and reliability cues.

Our findings may not fully generalize across all domains, despite deontologically-governed professions sharing common standards. The underlying cognitive mechanisms of critical thinking may prove transferable, but the specific conditions that trigger epistemic uncertainty or the effectiveness of minimal reconstructive explanations may vary. Different professional contexts involve different types of judgments, decision stakes, and forms of evidence, which may influence how practitioners engage with AI properties. Additionally, as the study was conducted in North America, cultural or institutional differences may also affect how professionals in other contexts engage with AI. Future research should explore how domain- and context-specific factors shape the relationship between explainability, reliability, transparency, and critical thinking.

A related consideration concerns the extent to which our findings engage cognitive mechanisms that may similarly arise when professionals evaluate human-generated recommendations. However, unlike human advisors who operate within shared social norms, transparent forms of expertise, and accountable agency, the distinctive nature of AI is associated with different patterns of confidence, trust, and reliance compared to human advice even when the informational content is matched (Alon-Barkat & Busuioc, 2023; Logg et al., 2019; You et al., 2022). This distinction was also evident in Experiment 1, where the no-AI-support condition elicited measurably different cognitive and neural engagement than AI-assisted conditions. Moreover, the cognitive processes uncovered in this work unfolded in an explicitly AI-assisted decision-making context, with participants evaluating outputs identified as AI-generated, consistent with prior experimental paradigms (Efendić et al., 2024; Logg et al., 2019; Sachin & Schecter, 2024). Future research may extend this work by comparing human-human and human-AI-assisted decision-making under equivalent tasks and conditions.

9 Conclusion

This research examined how AI explainability, reliability, and transparency influence critical thinking in deontological professional contexts. Through three experimental studies involving practicing HR professionals, we demonstrate that minimal reconstructive causal explanations enhance analytical reasoning and decision-confidence, provided that AI reliability remains consistent. When reliability is uncertain, epistemic uncertainty drives critical engagement. We further illuminate an emergent metacognitive process, suggesting that critical thinking is a context-sensitive phenomenon unfolding across decision-making phases, as attention, reasoning, and judgment interact synergistically in response to these embedded AI properties. These insights advance our understanding of HCAI in complex professional decision-making and provide actionable guidance for designing more human-centered AI.

Appendix

Table 6 Measurement items of self-report scales

Variable	Items	References
Trust in AI	How much do you trust the recommendation of the intelligent agent?	(Chen et al., 2020)
Decision-Confidence	How confident are you that you made the right decision?	(Inbar et al., 2011)

Table 7 Summary of hypotheses tests

Hypothesis	Description	Tested in	Supported
H1a	NAS will elicit greater deliberative-phase neural activity (attention, information, and semantic processing) than either MRE or FRE.	Experiment 1	Yes
H1b	MRE will elicit greater deliberative-phase neural activity (attention, information, and semantic processing) than FRE	Experiment 1	No
H1c	NAS will elicit greater executive-phase neural activity (analytical reasoning) than either MRE or FRE	Experiment 1	Yes
H1d	MRE will elicit greater executive-phase neural activity (analytical reasoning) than FRE.	Experiment 1	Yes
H2a	MRE will elicit greater executive-phase neural correlates of decision-confidence than FRE or NAS.	Experiment 1	Partially
H2b	MRE will elicit greater executive-phase physiological correlates of decision-confidence than either FRE or NAS.	Experiment 1	Partially
H2c	Post-decision, MRE will result in greater self-reported decision-confidence than either FRE or NAS.	Experiment 1	Partially
H3a	MRE will elicit reduced predictive encoding activity (shorter reading time) during the deliberative phase than FRE, but longer than NAS.	Experiment 1	Yes
H3b	MRE will elicit greater cognitive fluency (shorter decision times) during the executive phase than FRE, but less than NAS.	Experiment 1	No
H3c	MRE will result in greater selective AI reliance than FRE	Experiment 1	No
H4a	Attentional-semantic processing will be greater under epistemic uncertainty than under manifest unreliability, regardless of the level of explanation completeness.	Experiment 2	Yes
H4b*	Attentional-semantic processing will vary depending on the level of explanation completeness, regardless of epistemic uncertainty or manifest unreliability.	Experiment 2	No
H4c*	The influence of epistemic uncertainty versus manifest unreliability on attentional-semantic processing will vary depending on the level of explanation completeness.	Experiment 2	No
H5a	Analytical reasoning will be greater under epistemic uncertainty than under manifest unreliability, regardless of the level of explanation completeness.	Experiment 2	Yes
H5b*	Analytical reasoning will vary depending on the level of explanation completeness, regardless of epistemic uncertainty or manifest unreliability.	Experiment 2	No
H5c*	The influence of epistemic uncertainty versus manifest unreliability on analytical reasoning will vary depending on the level of explanation completeness.	Experiment 2	No
H6a	Appropriate reliance will be greater under epistemic uncertainty than under manifest unreliability, regardless of the level of explanation completeness.	Experiment 2	Yes
H6b*	Appropriate reliance will vary depending on the level of explanation completeness, regardless of epistemic uncertainty or manifest unreliability.	Experiment 2	No
H6c*	The influence of epistemic uncertainty versus manifest unreliability on appropriate reliance will vary depending on the level of explanation completeness.	Experiment 2	No
H7a	Decision-confidence will be better calibrated under epistemic uncertainty than under manifest unreliability, regardless of the level of explanation completeness.	Experiment 2	Yes
H7b*	Calibrated decision-confidence will vary depending on the level of completeness of explanation, regardless of epistemic uncertainty or manifest unreliability	Experiment 2	No
H7c*	The influence of epistemic uncertainty versus manifest unreliability on calibrated decision-confidence will vary depending on the level of explanation completeness.	Experiment 2	No
H8a	Trust will be better calibrated under epistemic uncertainty than under manifest unreliability, regardless of the level of explanation completeness.	Experiment 2	Inverted
H8b*	Calibrated trust will vary depending on the level of explanation completeness, regardless of epistemic uncertainty or manifest unreliability.	Experiment 2	No
H8c*	The influence of epistemic uncertainty versus manifest unreliability on calibrated trust will depend on the level of explanation completeness.	Experiment 2	No
H9a-e	The influence of explanation completeness (minimal versus fully reconstructive) on (a) attentional-semantic processing, (b) analytical reasoning, (c) appropriate reliance, (d) calibrated decision-confidence, and (e) calibrated trust in AI will vary depending on whether a confidence score is present or absent.	Experiment 3	No
H10a	The influence of epistemic uncertainty versus manifest uncertainty on attentional-semantic processing will vary depending on whether a confidence score is present or absent.	Experiment 3	No
H10b	The influence of epistemic uncertainty versus manifest uncertainty on analytical reasoning will vary depending on whether a confidence score is present or absent.	Experiment 3	No

Table 7 (continued)

Hypothesis	Description	Tested in	Supported
H10c	The influence of epistemic uncertainty versus manifest uncertainty on appropriate reliance will vary depending on whether a confidence score is present or absent.	Experiment 3	Yes
H10d	The influence of epistemic uncertainty versus manifest uncertainty on calibrated decision-confidence will vary depending on whether a confidence score is present or absent.	Experiment 3	Yes
H10e	The influence of epistemic uncertainty versus manifest uncertainty on calibrated trust in AI will vary depending on whether a confidence score is present or absent.	Experiment 3	No
H11a-e	The influence of explanation completeness on (a) attentional-semantic processing, (b) analytical reasoning, (c) appropriate reliance, (d) calibrated decision-confidence, and (e) calibrated trust in AI will vary depending on the combined effects of epistemic uncertainty or manifest unreliability and the presence or absence of a confidence score.	Experiment 3	No
H13a	Attentional-semantic processing will be greater under epistemic uncertainty than manifest uncertainty, regardless of explanation completeness or transparency.	Experiment 3	Yes
H13b	Analytical reasoning will be greater under epistemic uncertainty than manifest uncertainty, regardless of explanation completeness or transparency.	Experiment 3	Yes
H13c	Appropriate reliance will be greater under epistemic uncertainty than manifest uncertainty, regardless of explanation completeness or transparency	Experiment 3	Yes
H13d	Calibrated trust in AI scores will be greater under epistemic uncertainty than manifest uncertainty, regardless of explanation completeness or transparency.	Experiment 3	Yes
H13e	Calibrated decision-confidence scores will be lower under epistemic uncertainty than manifest uncertainty, regardless of explanation completeness or transparency.	Experiment 3	Yes

NAS: No AI support: *MRE:* Minimal reconstructive explanations, *FRE:* Fully reconstructive explanations; Asterisked hypotheses are exploratory, investigating the main effects and interactions of explanation completeness, reliability and transparency

Acknowledgements This research was supported by the 2021 IVADO Strategic Research Funding Program (PRF2021-05). The authors would like to thank Dr. Alexander John Karran for valuable discussions and feedback during the research process.

Authors' Contributions MK-S: ideation, coordination, design and development, experimental procedures, analysis, interpretation and writing. P-ML, XP-R, and SS: ideation, insight, editorial review. All authors contributed to the manuscript and approved the final version.

Funding This research was supported by the 2021 IVADO Strategic Research Funding Program (PRF2021-05).

Data Availability The data supporting this research are available in anonymized form on request to the corresponding author. The data are not publicly available due to privacy or ethical limits.

Declarations

Ethics Approval All experiments in this research were approved by the HEC Montréal Research Ethics Board (REB) (certificate #2023–5041). All participants provided written informed consent and could withdraw participation at any time.

Competing Interests The authors report no potential conflict of interest.

References

Ajayi-Nifise, A. O., Odeyemi, O., Mhlongo, N. Z., Ibeh, C. V., Elufioye, O. A., & Awonuga, K. F. (2023). The future of accounting: Predictions on automation and AI integration. *World Journal of Advanced Research and Reviews, 21*(2), 399–407. https://doi.org/10.30574/wjarr.2024.21.2.0466

Aksoy, L., Cooil, B., & Lurie, N. H. (2011). Decision quality measures in recommendation agents research. *Journal of Interactive Marketing, 25*(2), 110–122. https://doi.org/10.1016/j.intmar.2011.01.001

Alexander, P. A. (2014). Thinking critically and analytically about critical-analytic thinking: An introduction. *Educational Psychology Review, 26*, 469–476. https://doi.org/10.1007/s10648-014-9283-1

Alon-Barkat, S., & Busuioc, M. (2023). Human-AI interactions in public sector decision making: Automation bias and selective adherence to algorithmic advice. *Journal of Public Administration Research and Theory*, *33*(1), 153–169. https://doi.org/10.1093/jopart/muac007

Alonso, F. M. (2009). Shared intention, reliance, and interpersonal obligations. *Ethics, 119*(3), 444–475.

Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, R. N. (2007). Overcoming intuition: Metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology: General, 136*(4), 569–576. https://doi.org/10.1037/0096-3445.136.4.569

Alufaisan, Y., Marusich, L. R., Bakdash, J. Z., Zhou, Y., & Kantarcioglu, M. (2021). Does explainable artificial intelligence improve human decision-making? *Proceedings of the AAAI Conference on Artificial Intelligence*, 6618–6626.

Aravazhi, P. S., Gunasekaran, P., Benjamin, N. Z. Y., Thai, A., Chandrasekar, K. K., Kolanu, N. D., Prajjwal, P., Tekuru, Y., Brito, L. V., & Inban, P. (2025). The integration of artificial intelligence into clinical medicine: Trends, challenges, and future directions. *Disease-a-Month,* , Article 101882. https://doi.org/10.1016/j.disamonth.2025.101882

Arendt, H. (1971). Thinking and moral considerations: A lecture. *Social Research*, *38*(3), 417–446.

Asatiani, A., Malo, P., Nagbøl, P. R., Penttinen, E., Rinta-Kahila, T., & Salovaara, A. (2021). Sociotechnical envelopment of artificial intelligence: An approach to organizational deployment of inscrutable artificial intelligence systems. *Journal of the Association for Information Systems*, *22*(2), 325–352. https://doi.org/10.17705/1jais.00664

Başar-Eroglu, C., Strüber, D., Kruse, P., Başar, E., & Stadler, M. (1996). Frontal gamma-band enhancement during multistable visual perception. *International Journal of Psychophysiology*, *24*(1–2), 113–125. https://doi.org/10.1016/S0167-8760(96)00055-4

Bansal, G., Nushi, B., Kamar, E., Horvitz, E., & Weld, D. S. (2021). Is the most accurate AI the best teammate? Optimizing AI for teamwork. *Proceedings of the AAAI Conference on Artificial Intelligence*, 11405–11414. https://www.aaai.org

Barbey, A. K., Koenigs, M., & Grafman, J. (2013). Dorsolateral prefrontal contributions to human working memory. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior, 49*, 1195–1205. https://doi.org/10.1016/j.cortex.2012.05.022

Bartlett, F. C. (1932). Remembering: a study in experimental and social psychology. Cambridge University Press.

Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (1997). Deciding advantageously before knowing the advantageous strategy. *Science*, *275*(5304), 1293–1295. https://www.science.org

Benbya, H., Pachidi, S., & Jarvenpaa, S. L. (2021). Special issue editorial: Artificial intelligence in organizations: Implications for information systems research. *Journal of the Association for Information Systems (Vol*, *22*(2), 281–303. https://doi.org/10.17705/1jais.00662. Association for Information Systems.

Berente, N., Gu, B., & Recker, J. (2021). Managing artificial intelligence. *MIS Quarterly*, *45*(3), 1433–1450. https://doi.org/10.25300/MISQ/2021/16274

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, *19*, 2767–2796. https://doi.org/10.1093/cercor/bhp055

Bin, N. W., Awang, S. A., Fook, C. Y., Chin, L. C., & Ying, O. Z. (2019). A study of informative EEG channel and brain region for typing activity. *Journal of Physics: Conference Series*, *1372*(1), 1–6. https://doi.org/10.1088/1742-6596/1372/1/012008

Biradar, A., Ainapur, J., Kalyanrao, K., Aishwarya, A., Sudharani, S., Shivaleela, & Monika (2024). The impact of artificial intelligence on modern recruitment practices: A multi-company case study analysis. *International Journal of Business and Management Invention*, *13*(9), 143–150. https://doi.org/10.35629/8028-1309143150

Bivins, T. H. (2006). Responsibility and accountability. In *Ethics in public relations: Responsible advocacy*, pp. 19–38.

Bonnefon, J. F. (2018). The pros and cons of identifying critical thinking with system 2 processing. *Topoi*, *37*, 113–119. https://doi.org/10.1007/s11245-016-9375-2

Botvinick, M. M., Cohen, J. D., & Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: An update. *Trends in Cognitive Sciences, 8*(12), 539–546. https://doi.org/10.1016/j.tics.2004.10.003

Boucsein, W. (1999). Electrodermal activity as an indicator of emotional processes. *Science of Emotion and Sensibility, 2*(1), 1–25.

Boucsein, W. (2012). *Electrodermal activity*. Springer Science & Business Media.

Boucsein, W., & Thum, M. (1997). Design of work/rest schedules for computer work based on psychophysiological recovery measures. *International Journal of Industrial Ergonomics, 20*(1), 51–57. https://doi.org/10.1016/S0169-8141(96)00031-5

Braithwaite, J. J., & Watson, D. G. (2015). Issues surrounding the normalization and standardisation of skin conductance responses (SCRs).Technical research note. Selective attention & awareness laboratory (SAAL). In *Univ. Birmingham, Birmingham, UK, Tech. Rep*

Buçinca, Z., Malaya, M. B., & 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*(5(CSCW1)), 1–21.

Bunian, S., Al-Ebrahim, M. A., & Nour, A. A. (2024). Role and applications of Artificial Intelligence and machine learning in manufacturing engineering: A review. *Engineered Science, 29*, Article 1088. https://doi.org/10.30919/es1088

Cacioppo, J. T., Tassinary, L. G., & Berntson, G. G. (2017). *Handbook of psychophysiology*. Cambridge University Press.

Candrian, C., & Scherer, A. (2022). Rise of the machines: Delegating decisions to autonomous AI. *Computers in Human Behavior, 134*(March), Article 107308. https://doi.org/10.1016/j.chb.2022.107308

Cao, S., Gomez, C., & Huang, C. M. (2023). How time pressure in different phases of decision-making influences human-AI collaboration. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW2), 1–26. https://doi.org/10.1145/3610068

Cao, S., Liu, A., & Huang, C. M. (2024). Designing for appropriate reliance: The roles of AI uncertainty presentation, initial user decision, and user demographics in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction, 8*(CSCW1), 1–32. https://doi.org/10.1145/3637318

Casalino, G., Castellano, G., Kaymak, U., & Zaza, G. (2025). Call for papers: Special issue on explainability in human-centric AI. *Information Systems Frontiers*. https://persone.ict.uniba.it/rubrica/gabriella.casalino

Cavanagh, J. F., & Frank, M. J. (2014). Frontal theta as a mechanism for cognitive control. *Trends in Cognitive Sciences, 18*(8), 414–421. https://doi.org/10.1016/j.tics.2014.04.012

Chaiken, S., & Ledgerwood, A. (2012). A theory of heuristic and systematic information processing. *Handbook of theories of social psychology* , 1, 246–266. Sage.

Chaiken, S., & Maheswaran, D. (1994). Heuristic processing can bias systematic processing: Effects of source credibility, argument ambiguity, and task importance on attitude judgment. *Journal of Personality and Social Psychology*, *66*(3), 460–473.

Chayer, C., & Freedman, M. (2001). Frontal lobe functions. *Current Neurology and Neuroscience Reports*, *1*(6), 547–552. https://doi.org/10.1007/s11910-001-0060-4

Chen, M., Soh, H., Hsu, D., Nikolaidis, S., & Srinivasa, S. (2020). Trust-aware decision making for human-robot collaboration: Model learning and planning. *ACM Transactions on Human-Robot Interaction*, *9*(2), 1–23. https://doi.org/10.1145/3359616

Chi, M. T. H., Claser, R., & Farr, M. J. (2014). *The nature of expertise*. Psychology.

Christensen, B. T., & Ball, L. J. (2018). Fluctuating epistemic uncertainty in a design team as a metacognitive driver for creative cognitive processes. *CoDesign: International Journal of CoCreation in Design and the Arts, 14*(2), 133–152. https://doi.org/10.1080/15710882.2017.1402060

Clayton, M. S., Yeung, N., & Cohen Kadosh, R. (2015). The roles of cortical oscillations in sustained attention. *Trends in Cognitive Sciences, 19*(4), 188–195. https://doi.org/10.1016/j.tics.2015.02.004

Cohen, A. S., Lutzke, L., Otten, C. D., & Árvai, J. (2022). I think, therefore I act: The influence of critical reasoning ability on trust and behavior during the COVID-19 pandemic. *Risk Analysis*, *42*(5), 1073–1085. https://doi.org/10.1111/risa.13833

Cohen, M. S., & Freeman, J. T. (1996). Thinking naturally about uncertainty. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 179–183.

Coslett, H. B., & Schwartz, M. F. (2018). The parietal lobe and language. In *Handbook of Clinical Neurology*, 151, 365–375. Elsevier. https://doi.org/10.1016/B978-0-444-63622-5.00018-8

Courtemanche, F., Léger, P. M., Fredette, M., & Sénécal, S. (2022). *COBALT - Photobooth: Integrated UX Data System* (Patent VAL-0045).

Crudu, C. (2023). Professional ethics and deontology in the practice of social work. *Scientific Annals of the "Alexandru Ioan Cuza" University, Iaşi. New Series SOCIOLOGY AND SOCIAL WORK Section, 16*(2), 17–23. https://doi.org/10.47743/asas-2023-2-739

Danry, V., Pataranutaporn, P., Groh, M., & Epstein, Z. (2025). Deceptive explanations by large language models lead people to change their beliefs about misinformation more often than honest explanations. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems.*, 1–31. https://doi.org/10.1145/3706598.3713408

Danry, V., Pataranutaporn, P., Mao, Y., & Maes, P. (2023). Don't just tell me, ask me: AI systems that intelligently frame explanations as questions improve human logical discernment accuracy over causal ai explanations. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3544548.3580672

Dawson, M. E., Schell, A. M., & Courtney, C. G. (2011). The skin conductance response, anticipation, and decision-making. *Journal of Neuroscience, Psychology, and Economics, 4*(2), 111–116. https://doi.org/10.1037/a0022619

De Groot, A. D. (1965). *Thought and choice in chess*. Mouton.

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Descartes, R. (1998). *Discourse on the method* (D. A. Cress, Trans.). Hackett Publishing (Original work published 1637).

Dewey, J. (1933). How we think: a restatement of the relation of reflective thinking to the educative process. D.C. Heath.

Doran, D., Schulz, S., & Besold, T. R. (2017). What does explainable AI really mean? a new conceptualization of perspectives. *ArXiv Preprint ArXiv*:171000794. http://arxiv.org/abs/1710.00794

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *ArXiv Preprint ArXiv:1702.08608.* http://arxiv.org/abs/1702.08608

Duan, Y., Edwards, J. S., & Dwivedi, Y. K. (2019). Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda. *International Journal of Information Management, 48*, 63–71. https://doi.org/10.1016/j.ijinfomgt.2019.01.021

Edwards, S. L. (2007). Critical thinking: A two-phase framework. *Nurse Education in Practice, 7*(5), 303–314. https://doi.org/10.1016/j.nepr.2006.09.004

Efendić, E., de Van Calsey, P. P. F. M., Bahník, Š, & Vranka, M. A. (2024). Taking algorithmic (vs. human) advice reveals different goals to others. *International Journal of Human-Computer Interaction, 40*(1), 45–54. https://doi.org/10.1080/10447318.2023.2210886

Ellenrieder, S., Kallina, E. M., Pumplun, L., Gawlitza, J. F., Ziegelmayer, S., & Buxmann, P. (2023). Promoting learning through explainable artificial intelligence: An experimental study in radiology. *Proceedings of the International Conference on Information Systems*, 3.

Ericsson, K. A. (2004). Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Academic Medicine, 79*(10), 70–81.

Ernst, P., & von Müller, A. (2005). What is thinking? - Trying to define an equally fascinating and elusive phenomenon. In E. Kraft, G. Balázs, & P. Ernst (Eds.), *Neural correlates of thinking* (pp. v–vii). Springer.

Eva, K. W., & Regehr, G. (2005). Self-assessment in the health professions: A reformulation and research agenda. *Academic Medicine, 10*(80), S46–S54.

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology, 59*(1), 255–278. https://doi.org/10.1146/annurev.psych.59.103006.093629

Facione, P. A. (1990). Critical thinking: a statement of expert consensus for purposes of educational assessment and instruction. Research Findings and Recommendations.

Facione, P. A. (2015). Critical thinking: what it is and why it counts. *Insight Assessment*. https://doi.org/https://doi.org/ISBN13:978-1-891557-07-1.

Fecho, M., & Zöll, A. (2023). The power of trust: designing trustworthy machine learning systems in healthcare. *Proceedings of the International Conference on Information Systems.*, 3. https://aisel.aisnet.org/icis2023

Ferson, S., O'Rawe, J., Antonenko, A., Siegrist, J., Mickley, J., Luhmann, C. C., Sentz, K., & Finkel, A. M. (2015). Natural language of uncertainty: Numeric hedge words. *International Journal of Approximate Reasoning, 57*, 19–39. https://doi.org/10.1016/j.ijar.2014.11.003

Figner, B., & Murphy, R. O. (2011). Using skin conductance in judgment and decision making research. In M. Schulte-Mecklenbeck, A. Kuehberger, & R. Ranyards (Eds.), *A handbook of process tracing methods for decision research* (pp. 163–184). Psychology.

Fincher-Kiefer, R. (1996). Encoding differences between bridging and predictive inferences. *Discourse Processes*, *22*(3), 225–246. https://doi.org/10.1080/01638539609544974

Frankel, M. S. (1989). Professional codes: Why, how, and with what impact? *Journal of Business Ethics, 8*(3), 109–115.

Frankfurt, H. G. (2005). *On bullshit*. Princeton University Press.

Gefen, D., Benbasat, I., & Pavlou, P. A. (2008). A research agenda for trust in online environments. *Journal of Management Information Systems*, *24*(4), 275–286. https://doi.org/10.2753/MIS0742-1222240411

Goldstine, H. H. (1993). *The computer from Pascal to von neumann*. Princeton University Press.

Gregor, S., & Benbasat, I. (1999). Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS Quarterly*, *23*(4), 497–530.

Hark, M. Ter. (2003). Searching for the searchlight theory: From Karl Popper to Otto Selz. *Journal of the History of Ideas, 64*(3), 465–487. https://doi.org/10.1353/jhi.2003.0038

Harmony, T. (2013). The functional significance of delta oscillations in cognitive processing. *Frontiers in Integrative Neuroscience*, *7*(DEC), 1–10. https://doi.org/10.3389/fnint.2013.00083

Heidegger, M. (1966). *Discourse on thinking* (J. Anderson & E. Freund, Trans.). Harper & Row. (Original work published 1959).

Heidegger, M. (1968). *What is called thinking?* (F. D. Wieck & J. G. Gray, Trans.). Harper & Row (Original work published 1954).

Hinds, P. J., Roberts, T. L., & Jones, H. (2004). Whose Job Is It Anyway? A Study of Human- Robot Interaction in a Collaborative Task. *Human–Computer Interaction, 19*(1–2), 151–181. https://doi.org/10.1080/07370024.2004.9667343

Hiraishi, H., Ikeda, T., Saito, D. N., Hasegawa, C., Kitagawa, S., Takahashi, T., Kikuchi, M., & Ouchi, Y. (2021). Regional and temporal differences in brain activity with morally good or bad judgments in men: A magnetoencephalography study. *Frontiers* *in Neuroscience, 15*, Article 396. https://doi.org/10.3389/fnins.2021.596711

Hoffman, R. R., Klein, G., & Mueller, S. T. (2018). Explaining explanation for explainable AI. *Proceedings of the Human Factors and Ergonomics Society*, *1*, 197–201. https://doi.org/10.1177/1541931218621047

Holyoak, K. J., & Morrison, R. G. (2005). Thinking and reasoning: a reader's guide. In K. J. Holyoak, & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning*. Cambridge University Press.

Hudon, A., Demazure, T., Karran, A., Léger, P. M., & Sénécal, S. (2021). Explainable artificial intelligence (XAI): How the visualization of AI predictions affects user cognitive load and confidence. *Information Systems and Neuroscience: NeuroIS Retreat 2021*, *52 LNISO*, 237–246. https://doi.org/10.1007/978-3-030-88900-5_27

Inbar, Y., Botti, S., & Hanko, K. (2011). Decision speed and choice regret: When haste feels like waste. *Journal of Experimental Social Psychology*, *47*(3), 533–540. https://doi.org/10.1016/j.jesp.2011.01.011

Islam, M. R., Barua, S., Ahmed, M. U., Begum, S., Aricò, P., Borghini, G., & Flumeri, G. . Di. (2020). A novel mutual information based feature set for drivers' mental workload evaluation using machine learning. *Brain Sciences, 10*(8), 1–23. https://doi.org/10.3390/brainsci10080551

Jebelli, H., Hwang, S., & Lee, S. (2018). EEG signal-processing framework to obtain high-quality brain waves from an off-the-shelf wearable EEG device. *Journal of Computing in Civil Engineering*. https://doi.org/10.1061/(asce)cp.1943-5487.0000719

Jeunet, C., Albert, L., Argelaguet, F., & Lécuyer, A. (2018). Do you feel in control? Towards novel approaches to characterise, manipulate and measure the sense of agency in virtual environments. *IEEE Transactions on Visualization and Computer Graphics*, *24*(4), 1486–1495. https://doi.org/10.1109/TVCG.2018.2794598

Jiang, J., Kahai, S., & Yang, M. (2022). Who needs explanation and when? Juggling explainable AI and user epistemic uncertainty. *International Journal of Human-Computer Studies, 165*, Article 102839. https://doi.org/10.1016/j.ijhcs.2022.102839

Jussupow, E., Spohrer, K., Heinzl, A., & Gawlitza, J. (2021). Augmenting medical diagnosis decisions? An investigation into physicians' decision-making process with artificial intelligence. *Information Systems Research*, *32*(3), 713–735. https://doi.org/10.1287/ISRE.2020.0980

Kadiresan, A., Baweja, Y., & Ogbanufe, O. (2022). Bias in AI-based decision-making. In M. V. Albert, L. Lin, M. J. Spector, & L. S. Dunn (Eds.), *Bridging human intelligence and artificial intelligence. Educational communications and technology: Issues and innovations*. Springer. https://doi.org/10.1007/978-3-030-84729-6_19

Kahneman, B. Y. D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*(2), 363–391.

Kant, I. (1998). *Critique of pure reason* (P. Guyer & A. W. W. Wood, Trans.). Cambridge University Press (Original work published 1781).

Karran, A. J., Korosec-Serfaty, M., Malaguti, P., Le, D., Tyler, C., Léger, P. M., & Sénécal, S. (2024). When interacting with AI, make me think. *5th International Neuroergonomics Conference*, 343–348.

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Wortman Vaughan, J. (2020). Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3313831.3376219

Kawakami, A., Guerdan, L., Cheng, Y., Lee, M., Carter, S., Arechiga, N., Glazko, K., Zhu, H., & Holstein, K. (2023). Training towards critical use: Learning to situate AI predictions relative to human knowledge. *Proceedings of the ACM Collective Intelligence Conference, 63-78*. https://doi.org/10.1145/3582269.3615595

Kawasaki, M., & Yamaguchi, Y. (2012). Effects of subjective preference of colors on attention-related occipital theta oscillations. *NeuroImage, 59*(1), 808–814. https://doi.org/10.1016/j.neuroimage.2011.07.042

Kay, M., Kola, T., R Hullman, J., & A Munson, S. (2016). When (ish) is my bus? User-centered visualizations of uncertainty in everyday, mobile predictive systems. *Conference on Human Factors in Computing Systems - Proceedings*, *5092–5103*. https://doi.org/10.1145/2858036.2858558

Kayser, C., Ince, R. A. A., & Panzeri, S. (2012). Analysis of slow (theta) oscillations as a potential temporal reference frame for information coding in sensory cortices. *PLoS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1002717

Kim, T. W., Maimone, F., Pattit, K., Sison, A. J., & Teehankee, B. (2021). Master and slave: The dialectic of Human-Artificial Intelligence engagement. *Humanistic Management Journal, 6*(3), 355–371. https://doi.org/10.1007/s41463-021-00118-w

Kizilcec, R. F. (2016). How much information? effects of transparency on trust in an algorithmic interface. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2390–2395. https://doi.org/10.1145/2858036.2858402

Kleinig, J. (2016). Trust and critical thinking. *Educational Philosophy and Theory*. https://doi.org/10.1080/00131857.2016.1144167

Klimesch, W. (2018). The frequency architecture of brain and brain body oscillations: An analysis. *European Journal of Neuroscience, 48*(7), 2431–2453. https://doi.org/10.1111/ejn.14192

Knudsen, E. I. (2007). Fundamental components of attention. *Annual Review of Neuroscience, 30*(1), 57–78. https://doi.org/10.1146/annurev.neuro.30.051606.094256

Korman, J., & Khemlani, S. (2020). Explanatory completeness. *Acta Psychologica, 209*, Article 103139. https://doi.org/10.1016/j.actpsy.2020.103139

Kraft, E., Gulyás, B., & Pöppel, E. (2009). *Neural correlates of thinking*. Springer Berlin Heidelberg.

Kurihara, Y., Takahashi, T., & Osu, R. (2022). The relationship between stability of interpersonal coordination and inter - brain EEG synchronization during anti - phase tapping. *Scientific Reports, 0123456789*, 1–13. https://doi.org/10.1038/s41598-022-10049-7

Lebovitz, S., Lifshitz-Assaf, H., & Levina, N. (2022). To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis. *Organization Science*, *33*(1), 126–148. https://doi.org/10.1287/ORSC.2021.1549

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Le, T., Miller, T., Singh, R., & Sonenberg, L. (2023). Explaining model confidence using counterfactuals. *Proceedings of the AAAI Conference on Artificial Intelligence*, 11856–11864. https://www.aaai.org

Léger, P. M., Courtemanche, F., Fredette, M., & Sénécal, S. (2019). A cloud-based lab management and analytics software for triangulated human-centered research. *Information Systems and Neuroscience: NeuroIS Retreat 2018, 93–99*. https://doi.org/10.1007/978-3-030-01087-4

Lipman, M. (1987). Critical thinking: what can it be? *Analytic Teaching*, *8*(1).

Li, S., Ren, X., Schweizer, K., Brinthaupt, T. M., & Wang, T. (2021). Executive functions as predictors of critical thinking: Behavioral and neural evidence. *Learning and Instruction, 71*, Article 101376. https://doi.org/10.1016/j.learninstruc.2020.101376

Li, Z., Lu, Z., & Yin, M. (2023). Modeling human trust and reliance in AI-assisted decision Making: A Markovian approach. *The*

Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI-23), 6056–6064.

Logg, J. M. (2019). Using algorithms to understand the biases in your organization. *Harvard Business Review*, 1–2. https://hbr.org/2019/08/using-algorithms-to-understand-the-biases-in-your-organization

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes, 151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology*, *55*(3), 232–257. https://doi.org/10.1016/j.cogpsych.2006.09.006

Lombrozo, T. (2011). The instrumental value of explanations. *Philosophy Compass*, *8*(6), 539–551.

Lombrozo, T., & Carey, S. (2006). Functional explanation and the function of explanation. *Cognition*, *99*(2), 167–204. https://doi.org/10.1016/j.cognition.2004.12.009

Lury, D. A., & Fisher, R. A. (1972). Statistical methods for research workers. *The Statistician*, *21*(3), 229. https://doi.org/10.2307/2986695

Malaguti, P., Karran, A. J., Le, D., Mortin, H., Coursaris, C. K., Sénécal, S., & Léger, P. M. (2025). Investigating interaction friction in generative AI: Improving user experience and decision-making. *SIGHCI 2024 Proceedings*, 26. https://aisel.aisnet.org/sighci2024.

Marin, L., & Copeland, S. M. (2024). Self-trust and critical thinking online: A relational account. *Social Epistemology*, *6*(38), 696–708. https://doi.org/10.1080/02691728.2022.2151330

Mauti, G. (2024). Educating for artificial intelligence: critical thinking, responsibility and resistance. *Journal of Inclusive Methodology and Technology in Learning and Teaching*, *1*(4). www.inclusive-teaching.it

Merritt, S. M., Lee, D., Unnerstall, J. L., & Huber, K. (2015). Are well-calibrated users effective users? Associations between calibration of trust and performance on an automation-aided task. *Human Factors*, *57*(1), 34–47. https://doi.org/10.1177/0018720814561675

Mühl, K., Strauch, C., Grabmaier, C., Reithinger, S., Huckauf, A., & Baumann, M. (2020). Get ready for being chauffeured: Passenger's preferences and trust while being driven by human and automation. *Human Factors*, *62*(8), 1322–1338. https://doi.org/10.1177/0018720819872893

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*(1), 167–202. www.annualreviews.org

Millidge, B., Seth, A., & Buckley, C. L. (2021). Predictive coding: a theoretical and experimental review. *ArXiv Preprint ArXiv:2107.12979*. http://arxiv.org/abs/2107.12979

Müller-Putz, G. R., Riedl, R., & Wriessnegger, S. C. (2015). Electroencephalography (EEG) as a research tool in the information systems discipline: Foundations, measurement, and applications. *Communications of the Association for Information Systems*, *37*(1), 911–948. https://doi.org/10.17705/1cais.03746

Moazemi, S., Vahdati, S., Li, J., Kalkhoff, S., Castano, L. J. V., Dewitz, B., Bibo, R., Sabouniaghdam, P., Tootooni, M. S., Bundschuh, R. A., Lichtenberg, A., Aubin, H., & Schmid, F. (2023). Artificial intelligence for clinical decision support for monitoring patients in cardiovascular ICUs: A systematic review. *Frontiers in Medicine*. https://doi.org/10.3389/fmed.2023.1109411

Morrison, K., Shin, D., Holstein, K., & Perer, A. (2023). Evaluating the impact of human explanation strategies on human-AI visual decision-making. *Proceedings of the ACM on Human-Computer Interaction*. https://doi.org/10.1145/3579481

Muis, K. R., Chevrier, M., Denton, C. A., & Losenno, K. M. (2021). Epistemic emotions and epistemic cognition predict critical thinking about socio-scientific issues. *Frontiers in Education*, *6*, 669908. https://doi.org/10.3389/feduc.2021.669908

Naaz, F., Chen, L., Gold, A. I., Samuels, J., Krasnow, J., Wang, Y., Nestadt, P., Kamath, V., Chib, V. S., Nestadt, G., & Bakker, A. (2021). Neural correlates of doubt in decision-making. *Psychiatry Research - Neuroimaging*, *317*(March), 111370. https://doi.org/10.1016/j.pscychresns.2021.111370

Newman, A. (2019). *Research methods for cognitive neuroscience* (1st ed.). SAGE Publications Ltd.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting Racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.2307/26843725

Okasha, A. (1999). Prevention of deontological mistakes: The Role of ethical codes. In G. Christodoulou, D. Lecic-Tosevski, & V. Kontaxakis (Eds.), *Issues in Preventive Psychiatry* (pp. 134–142).

Omar, M., Soffer, S., Agbareia, R., Bragazzi, N. L., Apakama, D. U., Horowitz, C. R., Charney, A. W., Freeman, R., Kummer, B., Glicksberg, B. S., Nadkarni, G. N., & Klang, E. (2025). Sociodemographic biases in medical decision making by large Language models. *Nature Medecine*, 1–9.

Panciroli, C., Allegra, M., Gentile, M., & Rivoltella, P. C. (2023). Towards AI literacy: a proposal of a framework based on the episodes of situated learning. *3rd CINI National Conference on Artificial Intelligence (ITAL-IA 2023)*. http://ceur-ws.org

Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation-induced complacency. *The International Journal of Aviation Psychology*, *3*(1), 1–23. https://doi.org/10.1207/s15327108ijap0301_1

Park, J., Kim, H., Sohn, J. W., Choi, J. R., & Kim, S. P. (2018). EEG beta oscillations in the temporoparietal area related to the accuracy in estimating others' preference. *Frontiers in Human Neuroscience*, *12*(February), 1–11. https://doi.org/10.3389/fnhum.2018.00043

Pedrami, M., & Vaezi, S. K. (2025). Factors influencing artificial intelligence adoption in human resource management: A meta-synthesis and systematic review of multidimensional considerations. *Journal of Work-Applied Management*. https://doi.org/10.1108/JWAM-10-2024-0158

Petrides, M. (2013). Neuroanatomy of Language regions of the human brain. Academic.

Phillips-Wren, G. (2013). Intelligent decision support systems. In *Multicriteria Decision Aid and Artificial Intelligence: Links, Theory and Applications* (pp. 25–44).

Pisani, S., & Haw, M. D. (2023). Learner agency in a chemical engineering curriculum: Perceptions and critical thinking. *Education for Chemical Engineers*, *44*, 200–215. https://doi.org/10.1016/j.ece.2023.06.003

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, W., J., & Wallach, H. (2021). Manipulating and measuring model interpretability. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–52. https://doi.org/10.1145/3411764.3445315

Rempe, M. P., Ott, L. R., Picci, G., Penhale, S. H., Christopher-Hayes, N. J., Lew, B. J., Petro, N. M., Embury, C. M., Schantell, M., Johnson, H. J., Okelberry, H. J., Losh, K. L., Willett, M. P., Losh, R. A., Wang, Y. P., Calhoun, V. D., Stephen, J. M., Heinrichs-Graham, E., Kurz, M. J., & Wilson, T. W. (2023). Spontaneous cortical dynamics from the first years to the golden years. *Proceedings of the National Academy of Sciences, 120*(4), Article e2212776120. https://doi.org/10.1073/pnas.2212776120

Rest, J. R. (1986). Moral development: advances in research and theory. Praeger.

Riddick, F. A. (2003). The code of medical ethics of the American Medical Association. *American Medical Association*. www.ama-assn.org/go/ceja

Riedl, R. (2013). On the biology of technostress: literature review and research agenda. *ACM SIGMIS Database: The DATABASE for Advances in Information Systems*, *44*(1), 18–55. https://doi.org/10.1145/2436239.2436242

Riedl, R., Minas, R. K., Dennis, A. R., & Müller-Putz, G. R. (2020). Consumer-grade EEG instruments: insights on the measurement quality based on a literature review and implications for neurois research. *Information Systems and Neuroscience: NeuroIS Retreat*, *43*, 350–361. https://doi.org/10.1007/978-3-030-60073-0_41

Riek, L. D. (2012). Wizard of Oz studies in HRI: a systematic review and new reporting guidelines. *Journal of Human-Robot Interaction*, *1*(1), 119–136. https://doi.org/10.5898/jhri.1.1.riek

Rojas, R. F., Essam, D., Fidock, J., Barlow, M., Kasmarik, K., Anavatti, S., Garratt, M., & Abbass, H. (2020). Electroencephalographic workload indicators during teleoperation of an unmanned aerial vehicle shepherding a swarm of unmanned ground vehicles in contested environments. *Frontiers in Neuroscience*, *14*(40). https://doi.org/10.3389/fnins.2020.00040

Rubinstein, A. (2007). Instinctive and cognitive reasoning: A study of response times. *The Economic Journal*, *117*(523), 1243–1259. https://doi.org/10.1111/j.1468-0297.2007.02081.x.

Sachin, P. K., & Schecter, A. (2024). Advice utilization in combined human-algorithm decision-making: An analysis of preferences and behaviors. *Journal of the Association for Information Systems, 25*(6), 1439–1465. https://doi.org/10.17705/1jais.00896

Sarter, M., Berntson, G. G., & Cacioppo, J. T. (1996). Brain imaging and cognitive neuroscience: Toward strong inference in attributing function to structure. *American Psychologist, 51*(1), 13–21.

Schemmer, M., Kühl, N., Benz, C., Bartos, A., & Satzger, G. (2023). Appropriate reliance on AI advice: Conceptualization and the effect of explanations. *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 410–422. https://doi.org/10.1145/3581641.3584066

Schuetz, S., & Venkatesh, V. (2020). The rise of human machines: How cognitive computing systems challenge assumptions of user-system interaction. *Journal of the Association for Information Systems, 21*(2), 460–482. https://doi.org/10.17705/1jais.00608

Schwarz, N., Jalbert, M., Noah, T., & Zhang, L. (2021). Metacognitive experiences as information: Processing fluency in consumer judgment and decision making. *Consumer Psychology Review*, *4*(1), 4–25. https://doi.org/10.1002/arcp.1067

Seitz, J., & O'neill, P. (1996). Ethical decision-making and the code of ethics of the Canadian Psychological Association. *Canadian Psychology = Psychologie Canadienne, 1*(37), 23.

Shamon, H., & Berning, C. (2020). Attention check items and instructions in online surveys with incentivized and non-incentivizedquality?Samples: Boon or bane for data. *Survey Research Methods, 14*(1), 55–77. https://doi.org/10.18148/srm/2020.v14i1.7374

Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human-Computer Interaction, 36*(6), 495–504. https://doi.org/10.1080/10447318.2020.1741118

Siegel, H. (1989). *Educating reason: rationality, critical thinking and education*. Routledge Publishing.

Simon, H. A. (1960). The new science of management decision. Harper & Brothers.

Singhapakdi, A., & Vitell, S. J. (1991). Ethical and legal issues in selling and sales management. *Journal of Personal Selling & Sales Management, 11*(4), 1–12.

Smith, H. (2021). Clinical AI: Opacity, accountability, responsibility and liability. *AI and Society, 36*(2), 535–545. https://doi.org/10.1007/s00146-020-01019-6

Steinfeld, A., Jenkins, O. C., & Scassellati, B. (2009, March). The oz of wizard: simulating the human for interaction research. In Proceedings of the 4th ACM/IEEE *International Conference on Human-Robot Interaction*, 101-107. https://dl.acm.org/doi/epdf/10.1145/1514095.1514115

Tambe, P., Cappelli, P., & Yakubovich, V. (2019). Artificial intelligence in human resources management: Challenges and a path forward. *California Management Review, 61*(4), 15–42. https://doi.org/10.1177/0008125619867910

Tranel, D., & Damasio, H. (1994). Neuroanatomical correlates of electrodermal skin conductance responses. *Psychophysiology, 31*(5), 427–438. https://doi.org/10.1111/j.1469-8986.1994.tb01046.x

Vance, A., Lowry, P. B., & Eggett, D. (2015). Increasing accountability through user-interface design artifacts. *MIS Quarterly, 39*(2), 345–366. https://doi.org/10.2307/26628357

Van Den Berg, B., De Bruin, A. B. H., Marsman, J. B. C., Lorist, M. M., Schmidt, H. G., Aleman, A., & Snoek, J. W. (2020). Thinking fast or slow? Functional magnetic resonance imaging reveals stronger connectivity when experienced neurologists diagnose ambiguous cases. *Brain Communications*. https://doi.org/10.1093/braincomms/fcaa023

Verma, T., Lingenfelder, C., & Klakow, D. (2020). Defining explanation in an AI context. *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP.*, 314–322.

Verrax, F. (2016). Beyond professional ethics: GIS, codes of ethics, and emerging challenges. *echnoscience and citizenship: Ethics and governance in the digital society* (pp. 143–161). Springer Science and Business Media B.V. https://doi.org/10.1007/978-3-319-32414-2_10

Vilone, G., & Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, *76*, 89–106. https://doi.org/10.1016/j.inffus.2021.05.009

Wang, X., & Yin, M. (2021). Are explanations helpful? a comparative study of the effects of explanations in AI-Assisted decision-Making. *Proceedings of the 26th International Conference on Intelligent User Interfaces*, 318–328. https://doi.org/10.1145/3397481.3450650

Wang, Y., Zhang, Z., Jing, Y., Valadez, E. A., & Simons, R. F. (2016). How do we trust strangers? The neural correlates of decision making and outcome evaluation of generalized trust. *Social Cognitive and Affective Neuroscience*, *11*(10), 1666–1676. https://doi.org/10.1093/scan/nsw079

Wardle, S. G., Kriegeskorte, N., Grootswagers, T., Khaligh-Razavi, S. M., & Carlson, T. A. (2016). Perceptual similarity of visual patterns predicts dynamic neural activation patterns measured with MEG. *NeuroImage, 132*, 59–70. https://doi.org/10.1016/j.neuroimage.2016.02.019

Westerholm, P. (2009). Codes of ethics in occupational health - Are they important? *Continuing Medical Education*, *11*(27). http://students.kennesaw

Wick, M. R., Dutta, P., Wineinger, T., & Conner, J. (1995). Reconstructive explanation: A case study in integral calculus. *Expert Systems with Applications*, *8*(4), 463–473.

Wick, M. R., & Thompson, W. B. (1992). Reconstructive expert system explanations. *Artificial Intelligence, 54*, 33–70.

Wilhelm, R. A., Threadgill, A. H., & Gable, P. A. (2021). Motor preparation and execution for performance difficulty: Centroparietal beta activation during the effort expenditure for rewards task as a function of motivation. *Brain Sciences*. https://doi.org/10.3390/brainsci11111442

Williams, J. J., & Lombrozo, T. (2010). The role of explanation in discovery and generalization: Evidence from category learning. *Cognitive Science, 34*(5), 776–806. https://doi.org/10.1111/j.1551-6709.2010.01113.x

Yanamala, K. K. R. (2023). Transparency, privacy, and accountability in AI-enhanced HR processes. *Journal of Advanced Computing Systems, 3*(3), 10–18. https://doi.org/10.69987/JACS.2023.30302

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: Confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1310–1321. https://doi.org/10.1098/rstb.2011.0416

Yoto, A., Katsuura, T., Iwanaga, K., & Shimomura, Y. (2007). Effects of object color stimuli on human brain activities in perception and attention referred to EEG alpha band response. *Journal of Physiological Anthropology*, *26*(3), 373–379. https://doi.org/10.2114/jpa2.26.373

You, S., Yang, C. L., & Li, X. (2022). Algorithmic versus human advice: Does presenting prediction performance matter for algorithm appreciation? *Journal of Management Information Systems, 39*(2), 336–365. https://doi.org/10.1080/07421222.2022.2063553

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Marion Korosec-Serfaty is Assistant Professor of Information Technologies in the Department of Analytics, Operations, and Information Technologies at Université du Québec à Montréal's School of Management (ESG UQAM) in Montréal, Québec,Canada. Her research uses neuroscience-informed Information Systems methods (NeuroIS) to examine human–computer and human–AI interaction, with an emphasis on the responsible use of AI.

Pierre-Majorique Léger is Professor of Information Technology, Chair in User Experience, Director of ERPsim Lab, and Co-founder of Tech3Lab at HEC Montréal. His research explores human–computer interaction through the use of neurophysiological methods to study user experience in digital environments. His work has appeared in journals such as MIS Quarterly, Journal of Management Information Systems, and Journal of the Association for Information Systems.

Xavier Parent-Rocheleau is Associate Professor of Human Resource Management at HEC Montreal. He holds a PhD from the Université du Québec à Montréal (Canada). His research explores the digitalization of human resources management, the datafication of work, algorithmic management of the workforce and electronic surveillance of the workplace. His work has been published in influential journals such as Human Resource Management, Nature Reviews Psychology, Human Resource Management Review, Journal of Business Research, and Public Administration Review.

Sylvain Sénécal is Professor of Marketing, RBC Financial Group Chair of E-Commerce, and Co-founder of Tech3Lab at HEC Montréal. His research focuses on online consumer behavior and consumer neuroscience. His work has been published in journals such as Journal of the Academy of Marketing Science, Journal of Retailing, and Journal of the Association for Information Systems.