Thinking—Fast, Slow, and Artificial: How AI is Reshaping Human Reasoning and the Rise of Cognitive Surrender

Steven D. Shaw
The Wharton School of the University of Pennsylvania
3733 Spruce Street, Philadelphia, PA 19104
Email: shawsd@wharton.upenn.edu

Gideon Nave*
The Wharton School of the University of Pennsylvania
3733 Spruce Street, Philadelphia, PA 19104
Email: gnave@wharton.upenn.edu

*Corresponding author

Acknowledgments: This work was supported by the Wharton Behavioral Lab.

v20260111

Abstract

People increasingly consult generative artificial intelligence (AI) while reasoning. As AI becomes embedded in daily thought, what becomes of human judgment? We introduce Tri-System Theory, extending dual-process accounts of reasoning by positing System 3: artificial cognition that operates outside the brain. System 3 can supplement or supplant internal processes, introducing novel cognitive pathways. A key prediction of the theory is "cognitive surrender"—adopting AI outputs with minimal scrutiny, overriding intuition (System 1) and deliberation (System 2). Across three preregistered experiments using an adapted Cognitive Reflection Test (N = 1,372; 9,593 trials), we randomized AI accuracy via hidden seed prompts. Participants chose to consult an AI assistant on a majority of trials (>50%). Relative to baseline (no System 3 access), accuracy significantly rose when AI was accurate and fell when it erred (+25/-15 percentage points; Study 1), the behavioral signature of cognitive surrender (AI-Accurate vs. AI-Faulty contrast; Cohen's h = 0.81). Engaging System 3 also increased confidence, even following errors. Time pressure (Study 2) and per-item incentives and feedback (Study 3) shifted baseline performance but did not eliminate this pattern: when accurate, AI buffered time-pressure costs and amplified incentive gains; when faulty, it consistently reduced accuracy regardless of situational moderators. Across studies, participants with higher trust in AI and lower need for cognition and fluid intelligence showed greater surrender to System 3. Tri-System Theory thus characterizes a triadic cognitive ecology, revealing how System 3 reframes human reasoning and may reshape autonomy and accountability in the age of AI.

Thinking—Fast, Slow, and Artificial:
How AI is Reshaping Human Reasoning and the Rise of Cognitive Surrender

From personalized shopping to medical diagnostics, artificial intelligence (AI) is rapidly reshaping people's decisions. Everyday choices are increasingly augmented by or externalized to AI, and people routinely accept algorithmically generated answers, explanations, and predictions. Across domains, AI tools are not merely assisting decision-making; they are becoming decision-makers. This shift opens new theoretical ground: How should we understand human cognition and decision-making in an age when we outsource thinking to artificial processes?

For decades, dual-process theories of judgment and decision-making have served as a foundational framework for modeling cognitive processes. These theories propose two distinct decision-making processes: System 1, characterized by fast, intuitive, and affective processing, and System 2, defined by slow, deliberative, and analytical reasoning (Kahneman, 2011; Stanovich & West, 2000). While simplistic and not immune to criticism (Keren & Schul, 2009; Kruglanski & Gigerenzer, 2011; Melnikoff & Bargh, 2018), this distinction explains a wide range of phenomena—from reliance on heuristics and rapid choice to moral judgment (Greene et al., 2001), risk perception (Slovic et al., 2004), motivated reasoning (Kunda, 1990), susceptibility to misinformation (Pennycook & Rand, 2019; 2020), stereotype activation (Devine, 1989), confidence calibration (Thompson et al., 2011), and reasoning under cognitive load (De Neys, 2006).

Recent advances in AI expose a conceptual gap in dual-process theories: they presume cognition is confined to the biological mind. Today’s society increasingly relies on external cognitive agents (such as algorithms, recommender systems, and generative AI) to interpret

information, form judgments, and guide decisions. Whether generating a travel itinerary with ChatGPT, following Google Maps through unfamiliar streets, or accepting a dating algorithm's match, individuals routinely delegate reasoning components to machines. This goes beyond the well-established tendency to conserve effort as "cognitive misers" (Kahneman, 2011; Stanovich & West, 2000) and reflects a distinct phenomenon we call "cognitive surrender": the decision-maker no longer constructs an answer, but adopts one generated by an external system. Conceptually distinct from cognitive offloading (Risko & Gilbert, 2016), which involves strategically outsourcing a discrete task to an external tool (e.g., using a calculator), cognitive surrender represents a deeper abdication of critical evaluation, where the user relinquishes cognitive control and adopts the AI's judgment as their own. This raises several fundamental questions: How is AI integrated into decision-making processes? Who engages with it, when, and why? How does access to AI affect decision confidence and outcomes? Moreover, what happens when core building blocks of thought, such as inference, evaluation, and justification, are outsourced to artificial systems? Addressing these questions requires shifting from a purely consequentialist view (how AI alters outcomes) to a process-oriented lens: understanding how people engage artificial cognition, which cognitive paths they follow (e.g., surrender, offloading, override), who is most susceptible or resistant, and what conditions modulate these patterns.

This paper proposes a novel framework for understanding judgment and decision-making in the AI era: Tri-System Theory. Building on dual-process models, Tri-System Theory introduces a third cognitive system: System 3 (i.e., artificial cognition), defined as external, automated, data-driven reasoning originating from algorithmic systems rather than the human mind. While System 1 (intuition) and System 2 (deliberation) are internal processes shaped by individual experience, emotion, and logic, System 3 exists outside the self and operates through

statistical inference, pattern recognition, and machine learning. We argue that System 3 is not merely a tool that supports cognition but an active participant in cognitive processes. Decision-makers now use it not only to supply fast answers (pre-empting or suppressing System 1) and circumvent effortful thinking (short-circuiting System 2) but also to supplement and scaffold their reasoning. For example, System 3 may generate candidate options that feed into System 2 deliberation or flag contradictions that prompt users to re-evaluate an initial System 1 intuition. However, the same affordances can lead to a deeper transfer of agency, where System 3's outputs are adopted without verification, effectively substituting for judgment altogether: a state we term cognitive surrender.

Broadly put, Tri-System Theory posits that modern decision-making unfolds within a triadic cognitive ecology rather than a purely internal dual-process system. In this view, external, algorithmic cognition does not merely support intuition or deliberation; it can actively supplant, suppress, or augment them, altering the cost–benefit calculus of thinking itself. When System 3's outputs are fast, fluent, and seemingly authoritative, users may bypass effortful reasoning and adopt its answers as their own. Conversely, under certain conditions (e.g., when outputs violate expectations or introduce disfluency), System 3 can trigger greater deliberation, creating hybrid routes such as verify-then-adopt or override-then-rationalize.

This framework treats AI not as a passive aid but as a functional cognitive agent whose presence reshapes how and when internal systems engage. It complements and extends adjacent concepts such as the extended mind (Clark & Chalmers, 1998), automation bias (Mosier & Skitka, 1996), epistemic outsourcing (Lynch, 2016), and transactive memory systems (Wegner, 1987), but goes further by modeling dynamic delegation as the decision to offload, trust, or override AI reasoning in real time. Tri-System Theory thus provides a richer vocabulary and a

structural revision of cognitive architecture for explaining not just what decisions are made, but how they are formed, and whose cognition they reflect.

To test key predictions of the Tri-System Theory, we conduct a series of behavioral experiments using a modified version of the Cognitive Reflection Test (CRT; Frederick, 2005). In our core paradigm, participants solved reasoning and knowledge-based questions with and without access to System 3. Participants without access to System 3 (i.e., Brain-Only) could rely solely on their own intuition (System 1) and deliberation (System 2). Participants with access to System 3 could use an optional AI chatbot, but retained full autonomy over answers. Crucially, we manipulated the accuracy of System 3 (within-subjects; using AI models trained with hidden seed prompts), which allowed us to investigate how System 3 outputs interact with internal System 1 and System 2 cognition.

This design (Study 1) enabled us to assess how participants reason with or without access to System 3 through AI use, accuracy, follow/override behavior, and confidence. Our hypotheses are grounded in two theoretical propositions: System 3 use is ubiquitous in human reasoning, and engaging System 3 can lead to cognitive surrender. Namely, under this framework, we posit that participants will frequently consult the AI chatbot and adopt AI-generated answers, overriding intuitive and deliberative processes. Individual-difference and confidence measures in our design also allow us to observe heterogeneity in System 3 user profiles (e.g., characteristics of users who override incorrect outputs, ‘Independents’ who abstain from use) and metacognitive effects (e.g., how engaging System 3 affects confidence). Under cognitive surrender, decision-making accuracy will be surrendered to System 3 accuracy: participants will achieve higher accuracy when System 3 is correct and succumb to lower accuracy when System 3 is faulty.

To probe the boundaries of this effect, follow-up experiments manipulated key contextual moderators: time pressure (Study 2), hypothesized to increase reliance on System 3 by suppressing System 2 engagement; and performance incentives with item-level feedback (Study 3), expected to attenuate cognitive surrender by reactivating deliberative processing and increasing override of faulty AI. Together, these studies offer a structured and theory-driven test of Tri-System Theory, solidifying System 3's importance in augmenting, reconfiguring, or distorting human cognition and reasoning.

This research makes three main contributions. First, we introduce a new theoretical framework, Tri-System Theory, which updates cognitive models to include AI as a third system, System 3. Second, we provide empirical evidence of cognitive surrender, showing that AI outputs can enhance or undermine judgment and can inflate confidence, even when incorrect. We also show that cognitive surrender can reduce the negative performance effects of time pressure when AI is correct, whereas incentives and feedback may partially attenuate cognitive surrender, and that decision makers who trust AI more and engage in less effortful analytic thinking (lower cognitive reflection and fluid intelligence) are more likely to display cognitive surrender. Third, we outline practical implications for designers and policymakers concerned with AI's ethical and practical integration into society.

Tri-System Theory of Cognition

Dual-Process Theories

Over the past half a century, dual-process theories have served as a foundational framework in cognitive psychology, behavioral economics, and consumer research. These theories converge on a central insight: the human mind operates via two types of thought processes. Although specific terminology and emphases vary across traditions, dual-process

models generally posit two types of systems: System 1, fast, intuitive, and associative, and System 2, slow, reflective, and rule-based (Evans & Stanovich, 2013; Kahneman, 2011).

Kahneman (2003) articulated the distinction between two modes of thinking: System 1 (fast, automatic) and System 2 (slow, deliberative), a framework he later expanded in his book Thinking, Fast and Slow (Kahneman, 2011). While not the first to propose a dual-process account, Kahneman's model draws from and builds upon a broader family of dual-process theories across psychology and neuroscience. For example, Stanovich and West's (2000) influential model distinguishes the autonomous mind, aligned with intuitive, heuristic-based processing, from the analytic mind responsible for effortful, rule-based reasoning. In later work, Stanovich (2009) further differentiated the analytic system into the algorithmic mind, governing processing capacity, and the reflective mind, encompassing thinking dispositions, epistemic values, and metacognitive awareness. This framework emphasizes individual differences in cognitive ability and thinking style, introducing constructs such as cognitive decoupling and reflective override as key mechanisms for rational thought. Closely related is the concept of Need for Cognition (NFC; Cacioppo & Petty, 1982), which captures stable individual differences in people's tendency to engage in and enjoy effortful cognitive activities; individuals high in NFC are more likely to activate System 2 processing and override intuitive responses.

Other dual-process theories include Sloman's (1996) model, which differentiates a parallel, associative reasoning system and a serial, rule-based one, attributing cognitive biases to conflicts between these systems. Epstein's (1994) Cognitive-Experiential Self-Theory (CEST) proposes a rational system that is conscious, deliberative, and logical, and an experiential system that is intuitive, affect-laden, and fast. Importantly, CEST highlights that individuals vary in the degree to which they rely on each system in daily life. Evans' (2006) Heuristic-Analytic Theory

further posits that people default to fast heuristic processing, engaging analytic reasoning only when conflict or uncertainty is detected; in this view, System 2 functions as a supervisory mechanism that may endorse or override the outputs of System 1. In a complementary line of research, Lieberman (2007) proposed the X-System and C-System in social neuroscience, distinguishing between reflexive, emotion-driven processes and reflective, controlled processes, respectively. Convergent work in neuroscience draws a parallel distinction between goal-directed (model-based) and habitual (model-free) control (Balleine & O'Doherty, 2010; Daw, Niv, & Dayan, 2005). These systems are neurally dissociable and temporally distinct, mapping closely onto the core features of System 1 and System 2 (Dolan & Dayan, 2013). Despite important theoretical nuances, these models converge on the idea that human cognition consists of two interacting systems, each with distinct experiential, functional, and biological characteristics (for a summary of dual-process models of cognition, see Web Appendix Table W1).

Dual-process theories have been instrumental in explaining a wide range of psychological phenomena. For example, System 1 processes have been shown to underlie impulse purchases (Shiv & Fedorikhin, 1999), affective brand preferences (Zajonc, 1980), and context-driven or default-based choice patterns (Simonson, 1989). In contrast, System 2 supports more deliberative tradeoffs (Bettman, Luce, & Payne, 1998), ethical reasoning and moral choice (Greene et al., 2001), and high-stakes financial decision-making where analytic processing and self-control are required (Hinson, Jameson, & Whitney, 2003). Dual-process models have also clarified the role of fluency, default effects, and framing in shaping consumer judgments and behaviors (Petty & Cacioppo, 1986; Shiv & Fedorikhin, 1999).

Limitations of the Dual-Process Paradigm in the Age of AI

Despite their widespread influence, dual-process theories share a foundational limitation: they assume that all cognition occurs within the biological mind. Whether intuitive or analytical, fast or slow, mental processing is presumed to be an internally executed process. This assumption, what we term the 'brain-bound cognition assumption', increasingly misaligns with modern decision environments. With the rise of AI, people are not merely reacting to external inputs but actively delegating, deferring, and supplementing reasoning to non-human systems. For example, studies of AI-assisted endoscopy by physicians report that repeated deference to algorithmic recommendations (i.e., System 3) can erode unaided diagnostic performance ("deskilling"), underscoring how external cognition can reshape expertise (Budzyń et al., 2025). However, dual-process models treat such phenomena as simple tools or passive triggers of cognition, rather than active cognitive agents. This tool-versus-system perspective fails to account for the algorithmic autonomy of AI (Rainey & Hochberg, 2025). Furthermore, these models provide no account of dynamic delegation: the decision to surrender, trust, or override AI reasoning in real-time. As decision-makers increasingly engage with AI in ways that bypass, replace, or reshape their intuition and deliberation, the dual-process framework proves fundamentally incomplete. These emerging dynamics call for an expanded model of cognition, one that incorporates a third system: external, dynamic, algorithmic, and embedded in the modern cognitive ecology.

System 3 (Artificial): A New Cognitive Model

To extend traditional dual-process models to an algorithmic decision environment, we propose a novel cognitive framework that introduces a third mode of cognition, System 3: external, automated, machine-based reasoning that supplements, replaces, or reconfigures the

operations of Systems 1 and 2. It represents not a metaphorical system, but a functional cognitive agent, executed not in the biological brain, but in algorithmic systems accessed by the human mind.

Reframing Human-AI Cognition as Triadic. System 3 is defined by four foundational properties: it is external, automated, data-driven, and dynamic. First, unlike Systems 1 and 2, which are neurally instantiated (*in vivo*), System 3 resides externally (i.e., outside the human nervous system), in artificial infrastructures (*in silico*), such as cloud-based models, embedded algorithms, and large-scale machine learning systems (e.g., large language models; LLMs). It is not introspectively accessible, nor biologically constrained. Second, System 3 operates through automated systems, executing cognitive operations using statistical, rule-based, or generative algorithms, trained on large datasets. It engages in pattern recognition, prediction, summarization, and synthesis—functions associated with both Systems 1 and 2 reasoning but performed at scale and speed beyond human capacity. Third, System 3 outputs are the result of data-driven analyses founded on large-scale training data and feedback. Performance and accuracy reflect the underlying data distribution, including its gaps and biases. Finally, System 3 is dynamic: it does not function in isolation. It is interactive, responding to human and environmental inputs in real time. This triadic interaction brings an expansive knowledge base and statistical capabilities, and enables new forms of bolstered hybrid reasoning, post hoc justification, or passive acceptance. Table 1 illustrates how System 3 introduces a fundamentally new set of cognitive affordances that overlap with, but are not reducible to those of Systems 1 and 2.

Table 1. Cognitive affordances and tradeoffs of System 3

	System 1: Fast	**System 2: Slow**	**System 3: Artificial**
Origin	Human (intuitive/associative)	Human (analytical/reflective)	Artificial (algorithmic/statistical)
Processing speed	Fast	Slow	Fast/Variable
Cognitive effort	Low	High	None/Variable (depends on accessibility)
Accuracy	Prone to bias	Normative but effortful	High in structured domains; brittle in open-ended tasks
Affective input	Emotion-driven	Emotion-regulated	Emotion-neutral
Ethical reasoning	Implicit norms	Explicit deliberation	Nonpartisan; depends on training data
Justification	Experiential or post hoc	Rationalized, articulated	Data-driven; externally generated

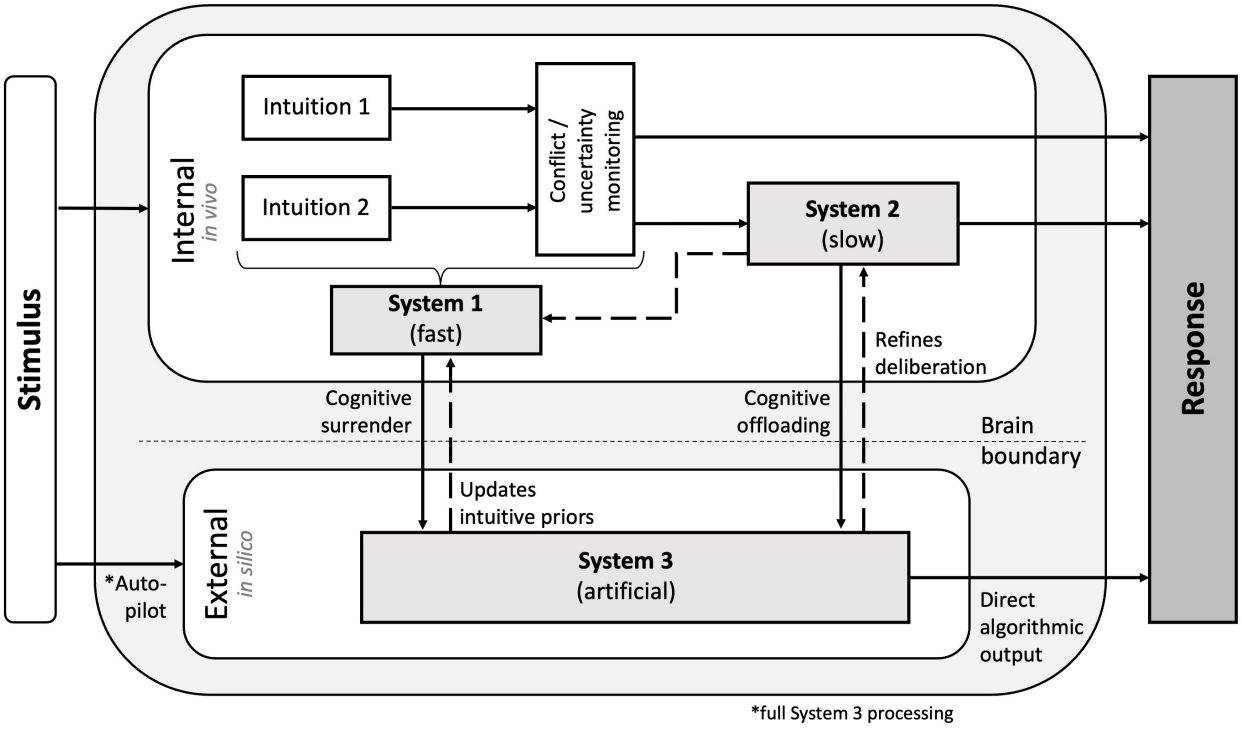
Notes: System 3 offers fast, externally generated, data-driven, and automated reasoning with minimal effort. When well-trained, it can deliver neutral, nonpartisan, and emotion-free outputs that are highly accurate in structured tasks.

In some domains, such as fact retrieval, language translation, or complex pattern recognition, System 3 can outperform System 1 and System 2 in both speed and accuracy. However, these affordances come at a cost: System 3 may lack affect, situational judgment, and normative reasoning grounded in human experience. Context is achieved by simulating coherence based on data, rather than possessing true phenomenological understanding. Improvements in grounding, memory, and alignment will narrow some gaps in situational judgment, but they are unlikely, in the near term, to confer intentional or affective understanding at a level of human experience. That is, System 3 introduces a qualitatively different mode of cognition (externalized, automated, and data-driven) that humans can access on demand. It

functions not just as a tool or extension but as a co-agent in reasoning, often delivering outputs with epistemic authority.

Tri-System Theory builds on and extends prior models of distributed cognition, including the extended mind (Clark & Chalmers, 1998), cognitive offloading (Risko & Gilbert, 2016), and automation bias (Mosier & Skitka, 1996) by modeling System 3 as a dynamic cognitive agent capable of influencing, displacing, or overriding internal reasoning systems (see the *Conceptual Foundations and Adjacent Constructs* section in Web Appendix A). In this view, reasoning is no longer merely fast or slow—it is increasingly artificial. System 3 is not internalized, but it is internalized in consequence: its outputs are integrated into thought and behavior, its logic affects judgment, and its presence reshapes how and when humans engage their own faculties (see Figure 1).

Figure 1. Tri-System Theory of Cognition



Notes: This model illustrates how human decision-making emerges from the interaction of three distinct cognitive systems: System 1 (fast, intuitive, heuristic processing), System 2 (slow, deliberative, rule-based reasoning), and System 3 (artificial, AI-based external cognition). A stimulus may be processed internally via System 1 or externally by System 3 (even before internal cognition is engaged, or without internal cognition at all). Internal systems operate within the individual, with System 2 recruited when conflict or uncertainty is detected. System 3 resides beyond the brain boundary and provides automated, algorithmic cognition. Transitions between systems are mediated by conflict/uncertainty monitoring. Engagement with System 3 occurs via cognitive offloading (delegation from System 2) or cognitive surrender (minimal System 1 engagement). System 3 can also feed back into internal systems, refining System 2 deliberation or updating System 1 intuition. A final response may originate from any of the three systems, reflecting a fluid integration of internal and external cognition.

Why System 3 Matters. The implications of System 3 are not merely additive to existing models of cognition; they are structural. As people increasingly integrate AI into their decision-making processes, they interact and engage with a cognitive system that can reshape the functions of both intuition and deliberation. For example, System 3 can replace System 1 by offering confident, ready-made answers that preempt the need for intuitive reasoning (i.e., autopilot; Chiriatti et al., 2024). It can suppress System 2 by diminishing the motivation or perceived necessity for reflective thought. Under conditions of ambiguity or contradiction, it may also trigger System 2, prompting users to engage in deeper analytical processing (see Table 2 for canonical routes of cognition). Further, engagement with artificial cognition is unlikely to be uniform: users may have differential propensity to engage System 3 or adopt algorithmic advice after observing errors (Dietvorst et al., 2015; Tully, Longoni, & Appel, 2025), motivating systematic heterogeneity in System 3 consultation (i.e., thinking profiles).

In some cases, System 3 may co-produce cognition, influencing the way individuals construct justifications, form beliefs, or explain their own decisions. Within this triadic cognitive ecology, the central question is no longer simply whether judgments are fast or slow, heuristic or analytical—it is whether those judgments were formed in the individual's mind, with the aid of an artificial system, or for the individual by System 3. These distinctions carry significant implications for agency, confidence, and epistemic responsibility. Tri-System Theory offers a unified framework to theorize these dynamics. It accounts for the rise of cognitive surrender, the fluidity of epistemic outsourcing, and the need for ethical design principles that preserve System 2 engagement in an age of System 3 integration.

Table 2. Canonical routes of cognition under Tri-System Theory

	Cognitive Locus	Theoretical Description
Intuition Stimulus → System 1 → Response	System 1	Fast, automatic processing based on heuristics or prior associations. No conflict detection; System 2/3 remain disengaged.
Deliberation Stimulus → System 1 → conflict /uncertainty → System 2 → Response	System 2	Reflective override triggered by detected conflict or uncertainty. Supports justification, rule-based reasoning, and analytic correction.
Cognitive Offloading Stimulus → System 1/2 → System 3 (assist) → System 1/2 → Response	System 2	Internal reasoning remains active; System 3 extends or scaffolds cognition. System 2 integrates artificial cognition. Strategic delegation.
Cognitive Surrender Stimulus → System 1 (brief) → System 3 → System 1 (optional) → Response	System 3	Minimal internal engagement. System 3's answer is accepted as one's own, without verification. Deliberation does not take place. Resembles delegation without oversight. Uncritical adoption.
Recursive or Hybrid Routes e.g., System 3 → System 1 (retraining), System 2 → System 1 (rationalize)	Mixed	Loopbacks and corrections. Includes: verify-then-adopt, override System 3, post-hoc rationalization, and long-run reanchoring of intuition. Recursive routes reflect reflective engagement, narrative repair, or shifts in intuitive response patterns shaped by repeated artificial cognition exposure.
Autopilot Stimulus → System 3 → Response	System 3	Immediate adoption of AI output without internal engagement. Stimulus never enters the brain side of the brain boundary. Bypasses System 1/2 processes.

Cognitive Surrender: A Behavioral-Epistemic Shift in Human–AI Reasoning

As AI systems increasingly participate in human cognition, a new phenomenon emerges that cannot be explained by traditional concepts such as cognitive offloading or automation bias alone. We define cognitive surrender as the behavioral and motivational tendency to defer judgment, effort, and responsibility to System 3's output, particularly when that output is delivered fluently, confidently, or with minimal friction.

Unlike cognitive offloading, which is typically strategic and task-specific (e.g., using GPS to navigate), cognitive surrender entails a deeper transfer of agency. Whereas cognitive offloading is a strategic delegation of deliberation, using a tool to aid one's own reasoning, cognitive surrender is an uncritical abdication of reasoning itself. It reflects not merely the use of external assistance, but a relinquishing of cognitive control: the user accepts the AI's response without critical evaluation, substituting it for their own reasoning. Whereas automation bias focuses on specific errors of omission or commission in response to automated tools, cognitive surrender describes a broader disposition of epistemic dependence. In cases of cognitive surrender, the user does not just follow System 3: they stop deliberative thinking altogether.

Tri-System Theory positions cognitive surrender as a central mechanism by which System 3 displaces or suppresses System 2. Rather than initiating deliberative/reflective processing, System 3 can short-circuit the deliberative path, reducing the likelihood of override or justification. This is especially likely under common psychological antecedents: time pressure, task complexity, low domain knowledge, high trust in AI, or a desire for cognitive ease. Decision-makers may not only accept System 3 cognitions but may also come to believe that AI reasoning is their own.

Empirically, cognitive surrender should manifest in measurable outcomes: users accept System 3 advice without critical analysis, show low override rates, offer shorter justifications, and display inflated confidence even when wrong (Spatharioti, Rothschild, Goldstein, & Hofman, 2025). In the studies that follow, we demonstrate cognitive surrender and show its persistence under time pressure and response incentives paired with item-level feedback. Additional moderators likely include the AI's perceived authority, presentation format and fluency, and users' beliefs in their own reasoning ability.

Importantly, cognitive surrender is not inherently irrational. In many domains (e.g., probabilistic settings, risk assessment, or extensive data), deferring to a statistically superior system may be adaptive or even optimal. However, from a theoretical standpoint, it marks a profound shift in the structure of cognition: one in which users may not know when or why they have deferred, and where the line between human and machine agency becomes blurred. As such, cognitive surrender represents both a consequence and a driver of the rise of System 3, and a fundamental element of Tri-System Theory.

Empirical Overview

We conducted three preregistered experiments and a within-paper trial-level synthesis to test key predictions derived from Tri-System Theory, examining how access to System 3 (artificial cognition) affects reasoning. Across studies, participants solved seven Cognitive Reflection Test (CRT-7) items, originally designed to dissociate intuitive (System 1) from deliberative (System 2) responding. In Study 1, System 3 availability was manipulated between subjects (Brain-Only vs. AI-Assisted). Crucially, among participants with access to AI (i.e., AI-Assisted) we manipulated System 3 accuracy on each trial (within-subjects; Trial Type) using hidden seed prompt logic so that the AI assistant returned either a correct answer (AI-Accurate)

or a confidently presented intuitive error (i.e., a wrong answer; AI-Faulty). Participants could consult the assistant ad libitum and retained full autonomy over their final responses.

The subsequent studies extend this framework by introducing contextual moderators. In Study 2, all participants had access to System 3, and we manipulated time pressure to assess how constrained deliberation affected reliance on artificial cognition. Study 3 tested whether performance incentives and item-level feedback would encourage deliberative monitoring and reduce uncritical adoption of AI responses. In both studies, a randomly placed Brain-Only probe trial served as a baseline for comparison. Additionally, we aggregated trial-level data across Studies 1–3 to estimate the overall magnitude of cognitive surrender and examine how it varied by context and individual traits. This preregistered trial-level synthesis used generalized linear mixed-effects models to test the robustness of System 3 effects across studies, platforms, and manipulations. The complete set of preregistered hypotheses and test statistics can be found in Web Appendix D: Summary of Preregistered Hypotheses.

Common Procedures. All participants completed seven open-ended CRT items adapted from Manfredi and Nave (2022), each with a canonical intuitive (incorrect) and deliberative (correct) response (for items and answers, see Web Appendix Table W2). In AI-Assisted conditions, an AI assistant (ChatGPT; GPT-4o) was embedded in the survey of each trial. Participants could engage with the assistant as much as they wished, and however they saw fit. The assistant's behavior was unconstrained, except regarding the current CRT item. When consulted about the item at hand, the AI assistant randomly returned either the correct deliberative (AI-Accurate) or faulty intuitive answer (AI-Faulty), accompanied by a short explanatory rationale (for seed prompts, see Web Appendix Table W3). Participants retained all

autonomy to follow or override its suggestions; answers were submitted via open-ended text submission. In Brain-Only conditions, no AI assistant was present.

Key Variables. Primary dependent variables included accuracy, System 3 engagement/AI chat usage (whether the assistant was consulted), and follow/override behavior (whether the AI's recommendation was adopted, conditional on engaging System 3). Confidence (0-100; continuous slider scale) was measured globally in Studies 1 and 2 and on a per-item basis in Study 3. Following the CRT task, participants completed four prespecified individual difference surveys: a 3-item Trust in AI scale (Jian et al., 2000), the 5-item short form of the Need for Cognition scale (NFC; Cacioppo & Petty, 1982), and, in Study 1 only, the 15-item Need for Cognitive Closure scale (NFCC; Roets & Van Hiel, 2011). These scales were presented in randomized order. Finally, participants completed a 13-item, 2-minute timed fluid intelligence test (Fluid IQ; Lyall et al., 2016).

Statistical Analysis. Analyses were conducted in R (v4.3.1) using the lme4 package. Accuracy and follow/override behavior were modeled as binary outcomes with generalized linear mixed-effects models (GLMM) using a logit link. Fixed effects included experimental condition (System 3 availability and/or contextual manipulation) and Trial Type (within-subjects; AI-Accurate vs. AI-Faulty), with random intercepts and slopes for participant and item (when applicable). Continuous outcomes (e.g., confidence, log-transformed response times) were analyzed using linear mixed-effects models. Planned contrasts estimated the effect of System 3 availability (Brain-Only vs. AI-Assisted) and the within-subject effect of System 3 output accuracy (Trial Type; AI-Accurate vs. AI-Faulty) among participants with access. Our trial-level synthesis pooled Brain-Only and AI-Assisted trials across all three experiments to estimate cognitive surrender effects and their moderators. Model results are reported as odds ratios (OR)

or unstandardized coefficients, with 95% confidence intervals (CI) and exact *p*-values (minimum reported $p < 2.20 \times 10^{-16}$). Models include a data source fixed effect (in-person vs. online) when appropriate. Individual difference scores were z-scored.

Data Collection. Data were collected in person and online. Laboratory participants completed the study in private cubicles using standardized hardware; online participants were restricted to desktop or laptop devices and instructed not to use external aids. All participants provided informed consent and were debriefed after the session. Additional methodological details can be found in Web Appendix B: Supplementary Methods. A summary of sample sizes and exclusions can be found in Web Appendix C: Detailed Sample Exclusions. Studies were approved by the University of Pennsylvania Institutional Review Board (Protocol #858904). Full materials, preregistrations, and codebooks are available on the Open Science Framework (OSF).

Study 1: System 3 Influences Cognitive Reasoning

Study 1 tested the core prediction of Tri-System Theory: that access to System 3 (Artificial) would systematically alter participants' reasoning behavior, accuracy, and confidence when solving reasoning problems. This study also provides the first empirical demonstration of cognitive surrender; participants heavily rely on System 3, and their accuracy is largely contingent on System 3 accuracy.

Methods

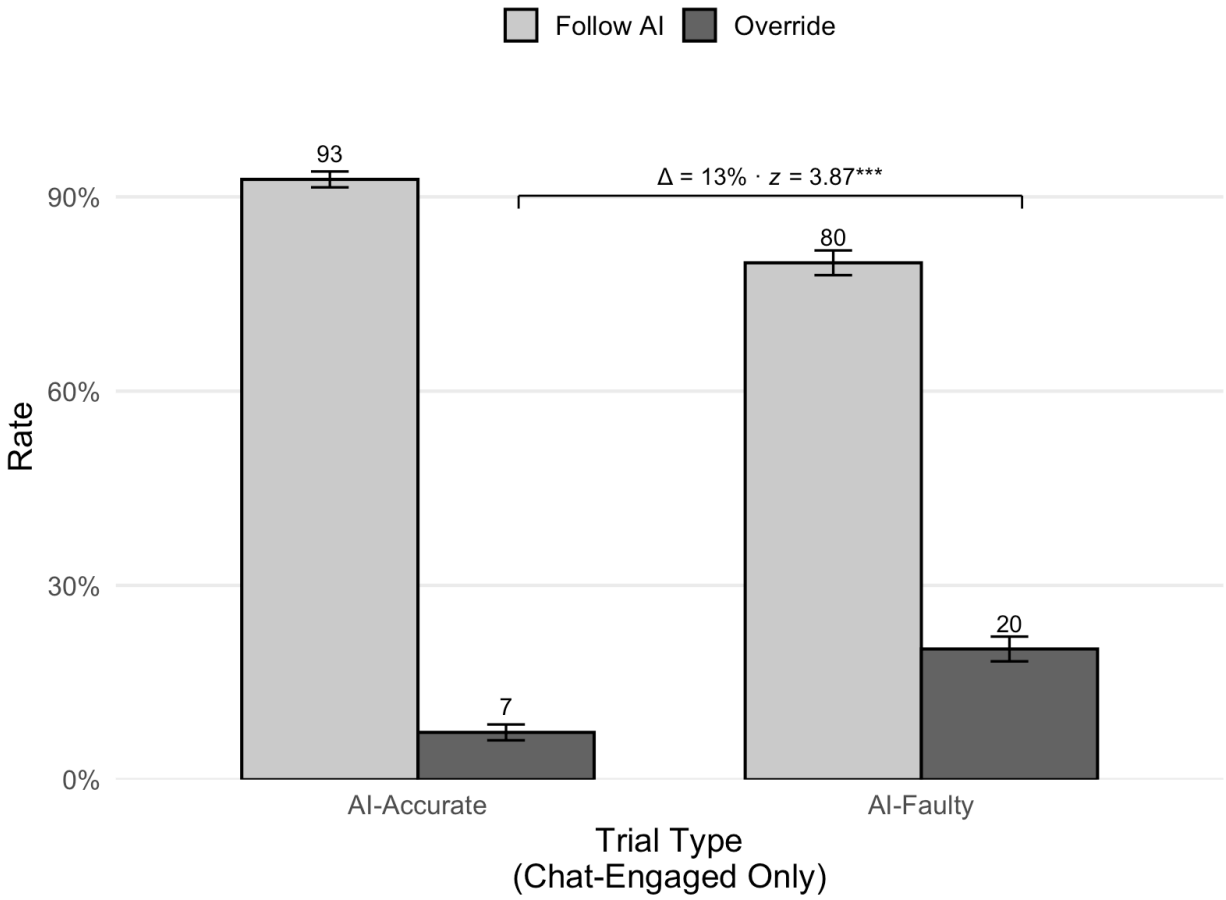
Participants ($N = 359$) were recruited in person by a university-affiliated behavioral laboratory. The primary task involved solving seven open-ended CRT items, each with an intuitive (incorrect) and deliberative (correct) numeric response. Participants could consult an AI assistant embedded in the experiment survey (if available) on each trial and retained full autonomy over whether to follow its suggestions. Participants were randomly assigned, between

subjects, to either a Brain-Only condition (no access to System 3; $n = 121$) or an AI-Assisted condition (System 3 accessible; $n = 238$), oversampling AI-Assisted 2:1. Within the AI-Assisted condition, Trial Type was manipulated within-subjects: on each item, the AI assistant returned either a correct response (AI-Accurate) or a confidently presented intuitive error (AI-Faulty), with item-level accuracy randomized across trials. Following the task, participants rated overall confidence in their answers, then completed three individual difference measures (Trust in AI, NFC, and NFCC) and a 2-minute Fluid IQ test. For robustness, a small online sample ($N = 81$) was collected to ensure that results replicated online (data not included in Study 1 results).

Results

Chat Use & AI Adoption. Participants in the AI-Assisted condition engaged with System 3 (i.e., chat use) on 54.4% of AI-Accurate trials and 52.8% of AI-Faulty trials. Conditional on using System 3, participants followed AI's advice on 92.7% of AI-Accurate trials ($SE = 1.2\%$, 95% *CI* [90.2, 95.0]), overriding it only 7.3% of the time, and 79.8% of AI-Faulty trials ($SE = 1.9\%$, 95% *CI* [75.9, 83.4]), overriding it 20.2% of the time. While override rates were substantially higher on AI-Faulty than AI-Accurate trials ($\beta = 1.43$, $SE = 0.37$, $z = 3.87$, $p = 1.09 \times 10^{-4}$, $OR = 4.19$, 95% CI [2.03, 8.65]), participants followed faulty AI recommendations on roughly four out of five chat-engaged trials (see Figure 2).

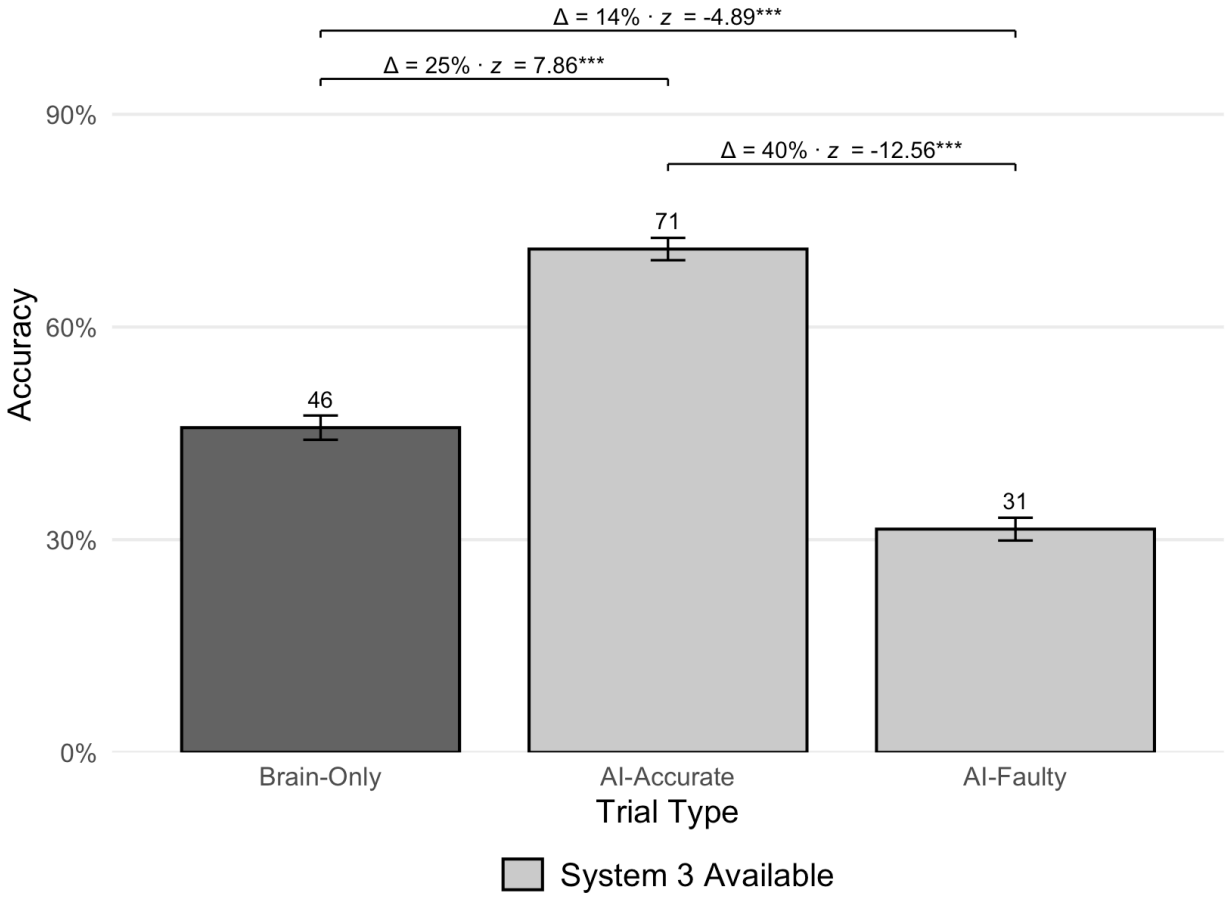
Figure 2. Participants adopt System 3 answers



Notes: Error bars indicate SEs. Z-test from a logistic mixed-effects model compares override rates across AI-Accurate and AI-Faulty Trial Types (***$p < 0.001$, **$p < 0.01$, *$p < 0.05$; two-sided).

Accuracy. CRT performance differed markedly by condition (Figure 3). Participants answered 45.8% correctly in the Brain-Only condition (*SE* = 1.7%, 95% *CI* [42.5, 49.2]), compared to 71.0% on AI-Accurate trials (*SE* = 1.6%, 95% *CI* [67.9, 74.1]), and 31.5% on AI-Faulty trials (*SE* = 1.6%, 95% *CI* [28.3, 34.6]).

Figure 3. System 3 facilitates cognitive surrender



Notes: Error bars indicate SEs. Asterisks reflect two-sided Wald z-tests from a logistic mixed-effects model for pairwise contrasts (***$p < 0.001$, **$p < 0.01$, *$p < 0.05$).

Relative to Brain-Only, accuracy was significantly higher on AI-Accurate trials ($\beta = 1.65$, $SE = 0.21$, $z = 7.86$, $p = 3.79 \times 10^{-15}$, $OR = 5.20$, 95% CI [3.45, 7.81]), and significantly lower on the AI-Faulty trials ($\beta = -1.01$, $SE = 0.21$, $z = -4.89$, $p = 9.84 \times 10^{-7}$, $OR = 0.37$, 95% CI [0.24, 0.55]). The contrast between AI-Accurate and AI-Faulty was especially large ($\beta = -2.65$, $SE = 0.21$, $z = -12.56$, $p < 2.20 \times 10^{-16}$, $OR = 0.07$, 95% CI [0.05, 0.11]), with participants having ~14 times lower odds of answering AI-Faulty items correctly (versus AI-Accurate). Thus, when

System 3 was accurate, participants' accuracy improved substantially; when faulty, accuracy dropped well below the Brain-Only baseline, illustrating cognitive surrender (Cohen's *h*: 0.81; 95% *CI* [0.72, 0.91]).

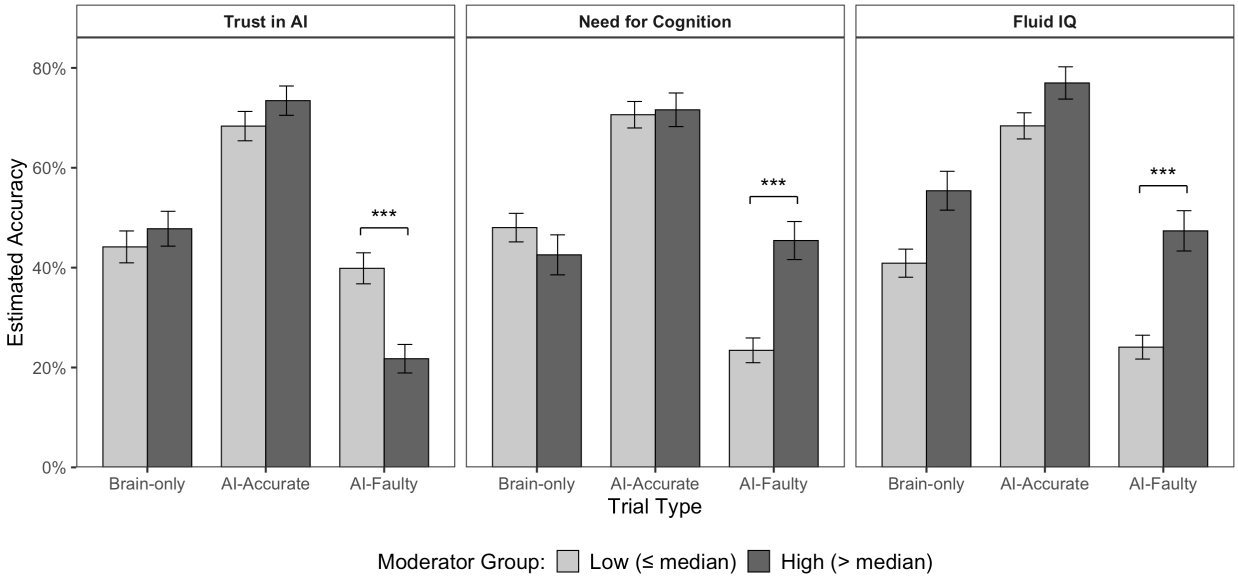
Confidence. Despite approximately half of System 3 answers being faulty, access to AI increased confidence by 11.7 percentage points (AI-Assisted: $M = 77.0\%$, $SE = 1.30\%$, 95% *CI* [74.4, 79.6]; Brain-Only: $M = 65.3\%$, $SE = 2.21\%$, 95% *CI* [61.0, 69.6]; $t(202.91) = 4.57$, $p = 8.55 \times 10^{-6}$; Hedges' $g = 0.54$, 95% *CI* [0.32, 0.77]). Within the AI-Assisted condition, confidence did not significantly decline as the number of faulty trials increased ($\beta = -1.14$, $SE = 0.89$, $t = -1.28$, 95% *CI* [-2.88, 0.61], $p = 0.202$).

Individual Differences. Pearson correlations among individual difference measures were small ($-0.19 \leq r \leq 0.18$). Participants higher in Trust in AI used the AI chat more often ($\beta = 0.50$, $SE = 0.19$, $z = 2.70$, $p = 0.007$, $OR = 1.64$, 95% *CI* [1.14, 2.35]), were less accurate on AI-Faulty trials ($\beta = -1.03$, $SE = 0.21$, $z = -4.97$, $p = 6.75 \times 10^{-7}$, $OR = 0.36$, 95% *CI* [0.24, 0.54]), and were more likely to follow faulty AI recommendations more often ($\beta = 1.25$, $SE = 0.34$, $z = 3.69$, $p = 2.22 \times 10^{-4}$, $OR = 3.47$, 95% *CI* [1.79, 6.73]).

Conversely, participants with higher Fluid IQ were more accurate on AI-Faulty trials ($\beta = 0.67$, $SE = 0.20$, $z = 3.39$, $p = 6.90 \times 10^{-4}$, $OR = 1.96$, 95% *CI* [1.33, 2.88]) and less likely to follow faulty AI suggestions when chat was engaged ($\beta = -1.49$, $SE = 0.36$, $z = -4.37$, $p = 1.24 \times 10^{-5}$, $OR = 0.23$, 95% *CI* [0.12, 0.44]). Fluid IQ also predicted lower overall chat usage ($\beta = -0.36$, $SE = 0.19$, $z = -1.87$, $p = 0.062$, $OR = 0.70$, 95% *CI* [0.48, 1.02]), although this effect did not reach statistical significance. Similarly, higher NFC was associated with better accuracy on AI-Faulty trials ($\beta = 0.62$, $SE = 0.20$, $z = 3.09$, $p = 0.002$, $OR = 1.86$, 95% *CI* [1.26, 2.75]) and reduced chat use ($\beta = -0.43$, $SE = 0.18$, $z = -2.34$, $p = 0.019$, $OR = 0.65$, 95% *CI* [0.45, 0.93]).

NFC did not reliably moderate AI output adoption behavior (β = -0.30, *SE* = 0.36, z = -0.82, p = 0.410, *OR* = 0.74, 95% *CI* [0.36, 1.51]; see Figure 4).

Figure 4. Individual differences moderate System 3 reasoning



Notes: Model-based predicted accuracy by Trial Type as a function of Trust in AI, NFC, and Fluid IQ, estimated from logistic moderation models with Trial Type × moderator interactions. Higher Trust in AI amplified vulnerability, while higher NFC and Fluid IQ buffered against faulty AI. Error bars indicate 95% confidence intervals.

NFCC did not moderate accuracy (β = -0.18, *SE* = 0.20, z = -0.89, p = 0.373, *OR* = 0.84, 95% *CI* [0.56, 1.24]) or chat use (β = -0.04, *SE* = 0.15, z = -0.28, p = 0.786, *OR* = 0.96, 95% *CI* [0.72, 1.28]), but did increase AI output adoption behavior among chat-engaged AI trials (β = 0.75, *SE* = 0.36, z = 2.10, p = 0.036, *OR* = 2.11, 95% *CI* [1.05, 4.25]).

Discussion

These findings support a core prediction of Tri-System Theory. When System 3 was available, participants engaged it and frequently adopted its answers to logic and reasoning

problems. Access to System 3 outputs significantly influenced accuracy, increasing correct answers when AI was correct, and decreasing accuracy when incorrect. Access to System 3 made decision-makers more confident, despite approximately half of System 3 outputs being incorrect. Finally, users who trust AI more and have lower NFC and fluid IQ were more likely to display cognitive surrender. Whether System 3 was accurate or faulty, its presence displaced internal reasoning.

Study 2: System 3 Buffers the Effects of Time Pressure

Study 2 tested whether time pressure alters the dynamics of System 3 reasoning. Time constraints are a well-established factor that reduces System 2 deliberation and increases reliance on intuitive, default processes associated with System 1 (e.g., Payne, Bettman, & Johnson, 1988; Evans & Curtis-Holmes, 2005). We examined the impact of time pressure on System 3 usage and cognitive surrender. As preregistered, we analyzed adoption of System 3 recommendations, overall accuracy, and examined effects within predefined thinking profiles (AI-Users, who engaged System 3 frequently, versus Independents, who primarily relied on internal reasoning).

Methods

Participants (N = 485) were recruited from a university-affiliated behavioral lab (n = 138) and online via Prolific (n = 347). Design and procedure mirrored Study 1, with two key changes. First, all participants had access to System 3 (akin to the AI-Assisted condition in Study 1), except for one randomly placed Brain-Only probe trial serving as a baseline/manipulation check; second, participants were randomly assigned, between subjects, to either a Time Pressure condition (n = 228) with a 30-second countdown per item (Bilancini, Boncinelli, & Celadin, 2024) or a Control condition (no timer/unlimited time; n = 257). Trial Type (AI-Accurate vs.

AI-Faulty) was manipulated within subjects at the item level. Consistent with prior literature, timeouts were treated as incorrect (Lugo et al., 2016).

After the task, participants completed a global confidence rating, followed by the Trust in AI and NFC scales (in random order) and a 2-minute timed Fluid IQ test. Planned contrasts estimated the effects of time pressure, System 3 accuracy, and their interaction within preregistered thinking profiles (AI-Users vs. Independents).

Results

Chat Use & AI Adoption. When available, participants used the chat on 52.0% of trials (Control: 53.9%; Time Pressure: 49.9%). Participants followed AI recommendations at high rates across conditions (AI-Accurate trials: Control = 80.3%; Time Pressure = 67.4%; AI-Faulty trials: Control = 74.6%; Time Pressure = 72.3%). In a mixed-effects model predicting following AI advice on chat-engaged trials, the Time Pressure × Trial Type interaction was not significant ($\beta = -0.49$, $SE = 0.30$, $p = 0.102$; $OR = 0.61$, 95% CI [0.34, 1.10]), indicating no reliable evidence that time pressure differentially altered selectivity in following accurate versus faulty AI advice. Critically, chat use was strongly linked to performance: opening the chat predicted higher accuracy on AI-Accurate trials ($\beta = 1.10$, $SE = 0.13$, $z = 8.49$, $p < 2.20 \times 10^{-16}$, $OR = 3.00$, 95% CI [2.33, 3.86]) and lower accuracy on AI-Faulty trials ($\beta = -1.31$, $SE = 0.18$, $z = -7.46$, $p = 8.89 \times 10^{-14}$, $OR = 0.27$, 95% CI [0.19, 0.38]), replicating Study 1's signature cognitive-surrender pattern.

Accuracy Under Time Pressure: Replicating previous studies (e.g., Thompson et al., 2011; Raoelison et al., 2020), CRT scores significantly decreased under time pressure. On the Brain-Only probe trial, participants in the Time Pressure condition were less accurate, answering 32.6% of items correctly ($SE = 3.1\%$, 95% CI [26.5, 38.8]), compared to 46.9% in the Control

condition (*SE* = 3.1%, 95% *CI* [40.7, 53.1]; β = -0.63, *SE* = 0.20, *z* = -3.17, *p* = 1.52×10^{-3}, *OR* = 0.53, 95% *CI* [0.36, 0.79]).

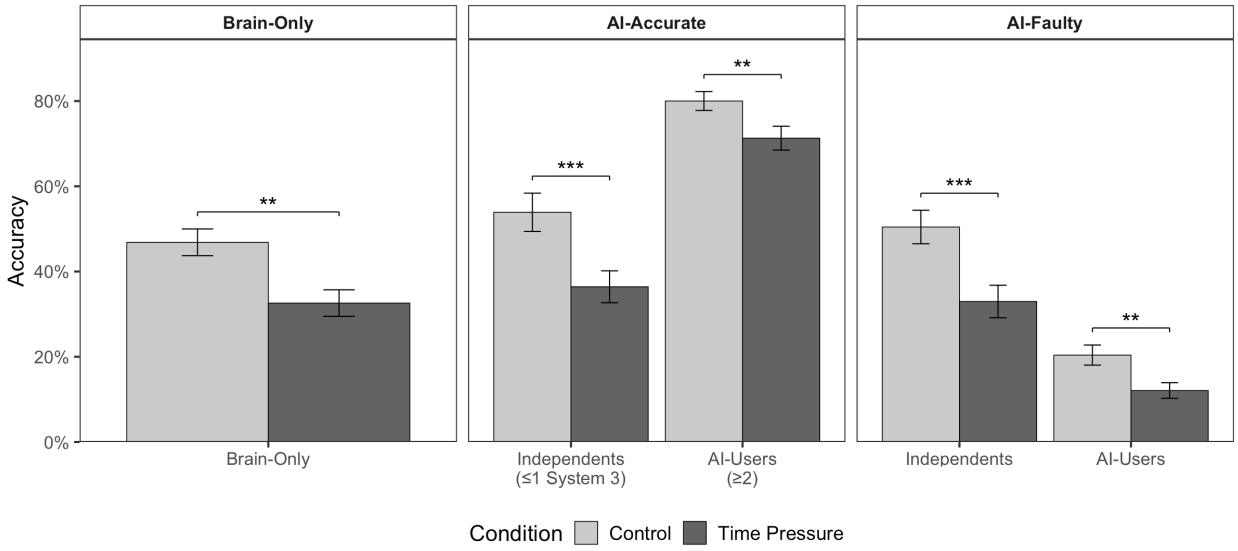
Across AI-Assisted trials, time pressure continued to impair accuracy (β = -0.86, *SE* = 0.15, *z* = -5.65, *p* = 1.59×10^{-8}, *OR* = 0.42, 95% *CI* [0.31, 0.57]): participants under Time Pressure answered 38.9% correctly (*SE* = 1.3%, 95% *CI* [36.3, 41.5]), compared to 51.9% in the Control condition (*SE* = 1.3%, 95% *CI* [49.3, 54.5]). Controlling for System 3 usage, we found no evidence of an interaction between Time Pressure and Trial Type on accuracy (β = -0.02, *SE* = 0.23, *z* = -0.08, *p* = 0.937, *OR* = 0.98, 95% *CI* [0.63, 1.53]) indicating that time pressure impaired performance consistently.

Thinking Profiles. Participants classified as Independents (used AI once or never; *n* = 170) had lower accuracy under Time Pressure (34.0%) than in the Control condition (51.5%; β = -1.18, *SE* = 0.28, *z* = -4.22, *p* = 2.48×10^{-5}; OR = 0.31, 95% *CI* [0.18, 0.52]), and did not differ in accuracy across AI Trial Types (Control: *OR* = 1.39, *z* = 1.42, *p* = 0.155, 95% *CI* [0.88, 2.20]; Time Pressure: *OR* = 1.23, *z* = 0.96, *p* = 0.339, 95% *CI* [0.81, 1.86]). Model-based contrasts confirmed that Independents and Brain-Only participants did not differ in either condition (Control: *OR* = 1.30, *SE* = 0.296, z = 1.15, *p* = 0.25, 95% *CI* [0.83, 2.03]; Time Pressure: *OR* = 0.92, *SE* = 0.205, *z* = -0.38, *p* = 0.704, 95% *CI* [0.59, 1.42]).

In contrast, participants classified as AI-Users (engaged the AI chat two or more times; *n* = 315) had higher accuracy on AI-Accurate trials (Control = 80.0%, Time Pressure = 71.3%) and lower on AI-Faulty trials (Control = 20.4%, Time Pressure = 12.1%). Relative to the Brain-Only baseline, AI-Users were substantially more accurate on AI-Accurate trials (Control: *OR* = 7.98, *z* = 9.20, *p* < 2.20×10^{-16}; Time Pressure: *OR* = 7.94, *z* = 8.31, *p* < 2.20×10^{-16}), and substantially less accurate on AI-Faulty trials (Control: *OR* = 0.25, *z* = -6.36, *p* = 2.05×10^{-10};

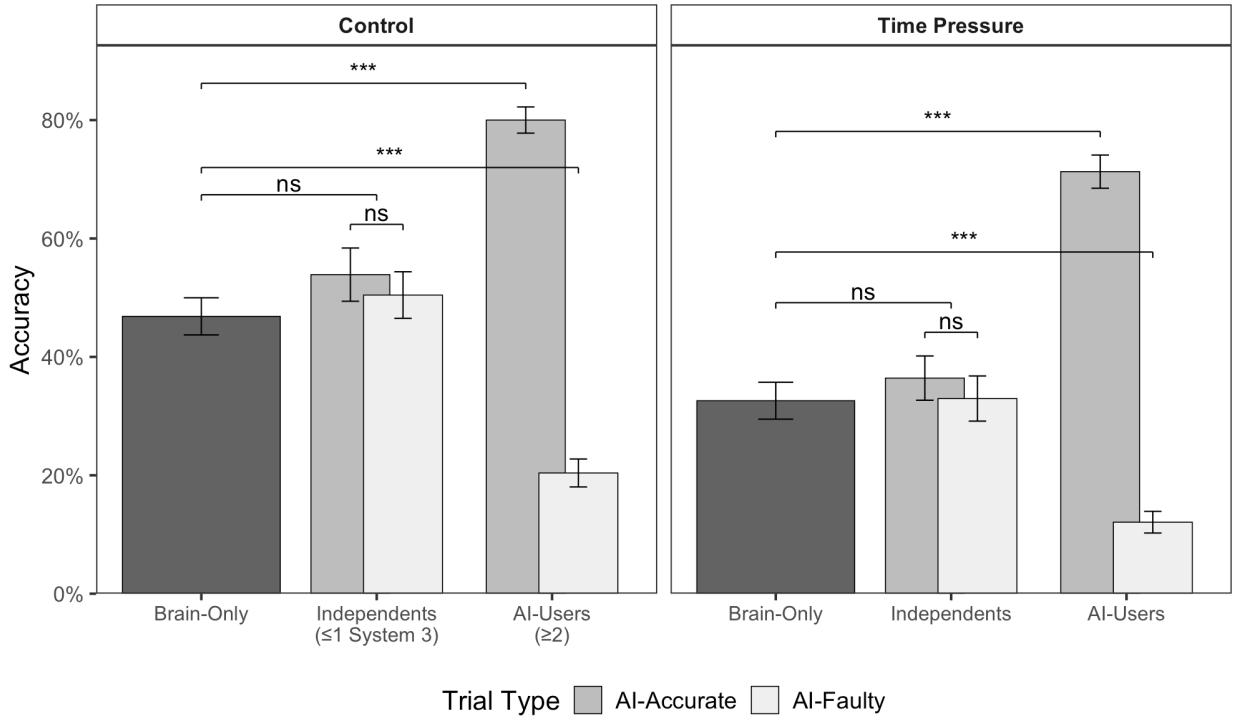
Time Pressure: $OR = 0.28$, $z = -4.91$, $p = 9.05 \times 10^{-7}$). A GLMM restricted to AI-Users confirmed a powerful main effect of Trial Type ($\beta = 3.71$, $SE = 0.23$, $z = 16.19$, $p < 2.20 \times 10^{-16}$; $OR = 40.9$, 95% CI [26.1, 64.1]; see Figures 5 and 6).

Figure 5. Time pressure reduces accuracy across trial types and thinking profiles



Notes: Brain-Only probe trials replicated classic CRT effects: accuracy declined under time pressure. The same pattern held for Independents (≤1 System 3 use). Among AI-Users (≥2 uses), accuracy tracked System 3 output quality: accuracy was higher on AI-Accurate trials and lower on AI-Faulty trials, consistent with cognitive surrender. Error bars indicate SEs. Asterisks denote significance of pairwise contrasts (***$p < 0.001$, **$p < 0.01$, *$p < 0.05$; two-sided).

Figure 6. System 3 can avert time pressure performance declines



Notes: Whereas Independents' accuracy is strongly reduced with time pressure and parallels Brain-Only performance, AI-Users produced significantly higher/lower accuracy on System 3 accurate/faulty trials. Under time pressure, participants who used System 3 (and it was accurate) had the highest accuracy. Error bars indicate SEs. Asterisks denote significance of pairwise contrasts (***$p < 0.001$, **$p < 0.01$, *$p < 0.05$; two-sided).

Individual Differences. Similar to Study 1, participants higher in Trust in AI were significantly more likely to engage System 3 (i.e., more likely to be AI-Users; $\beta = 0.26$, $SE = 0.05$, $t = 5.32$, $p = 1.59 \times 10^{-7}$), whereas NFC showed a trend toward less engagement ($\beta = -0.07$, $SE = 0.04$, $t = -1.64$, $p = 0.101$) and Fluid IQ was not predictive ($\beta = -0.03$, $SE = 0.05$, $t = -0.65$, $p = 0.516$). These relationships were not moderated by Time Pressure (Trust in AI ×

Time Pressure: β = -0.03, SE = 0.07, t = -0.41, p = 0.681; NFC × Time Pressure: β = -0.03, SE = 0.07, t = -0.51, p = 0.613; Fluid IQ × Time Pressure: β = 0.00, SE = 0.07, t = 0.02, p = 0.987).

Among System 3 engaged trials, Trust in AI strongly predicted following AI (β = 0.92, SE = 0.14, z = 6.79, $p = 1.15 \times 10^{-11}$), while NFC (β = -0.18, SE = 0.13, z = -1.40, p = 0.161) and Fluid IQ did not (β = 0.09, SE = 0.12, z = 0.71, p = 0.478). Time Pressure did not significantly interact with Trust in AI (β = -0.07, SE = 0.18, z = -0.39, p = 0.696), NFC (β = 0.10, SE = 0.17, z = 0.56, p = 0.576), or Fluid IQ (β = 0.13, SE = 0.17, z = 0.74, p = 0.457).

Finally, we examined whether thinking profiles reflected distinct psychological dispositions. AI-Users scored higher on Trust in AI (M = 4.44, SE = 0.04) than Independents (M = 3.45, SE = 0.05; t(2269) = 14.98, $p < 2.20 \times 10^{-16}$, 95% CI for Δ [0.86, 1.12]), and lower on NFC (M = 3.36, SE = 0.02) than Independents (M = 3.57, SE = 0.03; t(2404.6) = -6.18, $p = 7.59 \times 10^{-10}$, 95% CI for Δ [-0.27, -0.14]). Importantly, Fluid IQ did not differ between thinking profiles (AI-Users: M = 6.12, SE = 0.04; Independents: M = 6.16, SE = 0.06; t(2484.5) = -0.65, p = 0.518, 95% CI for Δ [-0.18, 0.10]).

Discussion

Overall, time pressure reduced performance, suppressed System 2 engagement, and shifted participants toward either System 1 (intuitive) or System 3 (artificial). Among Independents, who rarely engaged System 3, and those without System 3 access (Brain-Only), time pressure led to impaired performance, replicating classic dual-process findings. Among AI-Users, who engaged System 3 frequently, performance was less affected by time constraints but was dependent on the correctness of System 3 outputs. Replicating the cognitive surrender effect in Study 1, when System 3 was accurate, AI-Users performed well; when it was faulty, they

performed poorly. These results illustrate that System 3 adoption can buffer the cognitive demands of time pressure and reduce its adverse performance effects when System 3 is accurate.

Study 3: Incentives and Feedback Reduce Cognitive Surrender

Study 3 tested whether performance-based incentives combined with item-by-item feedback would reduce reliance on System 3, particularly when its outputs were faulty. This Incentives + Feedback manipulation was expected to reinforce System 2 monitoring and reduce uncritical adoption of System 3 outputs, thereby decreasing cognitive surrender. Further, Study 3 helps discriminate between deference to System 3 as a System 1 heuristic and cognitive surrender, because item-by-item feedback supplies direct error signals that should weaken deference if it is primarily calibrated to perceived AI accuracy.

Methods

Participants ($N = 450$) were recruited via Prolific. The design closely mirrored Study 2, except that participants were randomly assigned to either an Incentives + Feedback ($n = 238$) or a Control condition ($n = 212$) instead of a time pressure manipulation. Participants under Incentives + Feedback earned $0.20 (USD) and were entered into a $20 lottery for each correct answer, and received immediate item-level feedback (correct/incorrect) after submitting each response. Unlike Studies 1 and 2, which used global confidence ratings, Study 3 collected per-item confidence ratings after each response (prior to item-level feedback), enabling fine-grained confidence analyses. In the Control condition, participants did not receive incentive bonuses or feedback.

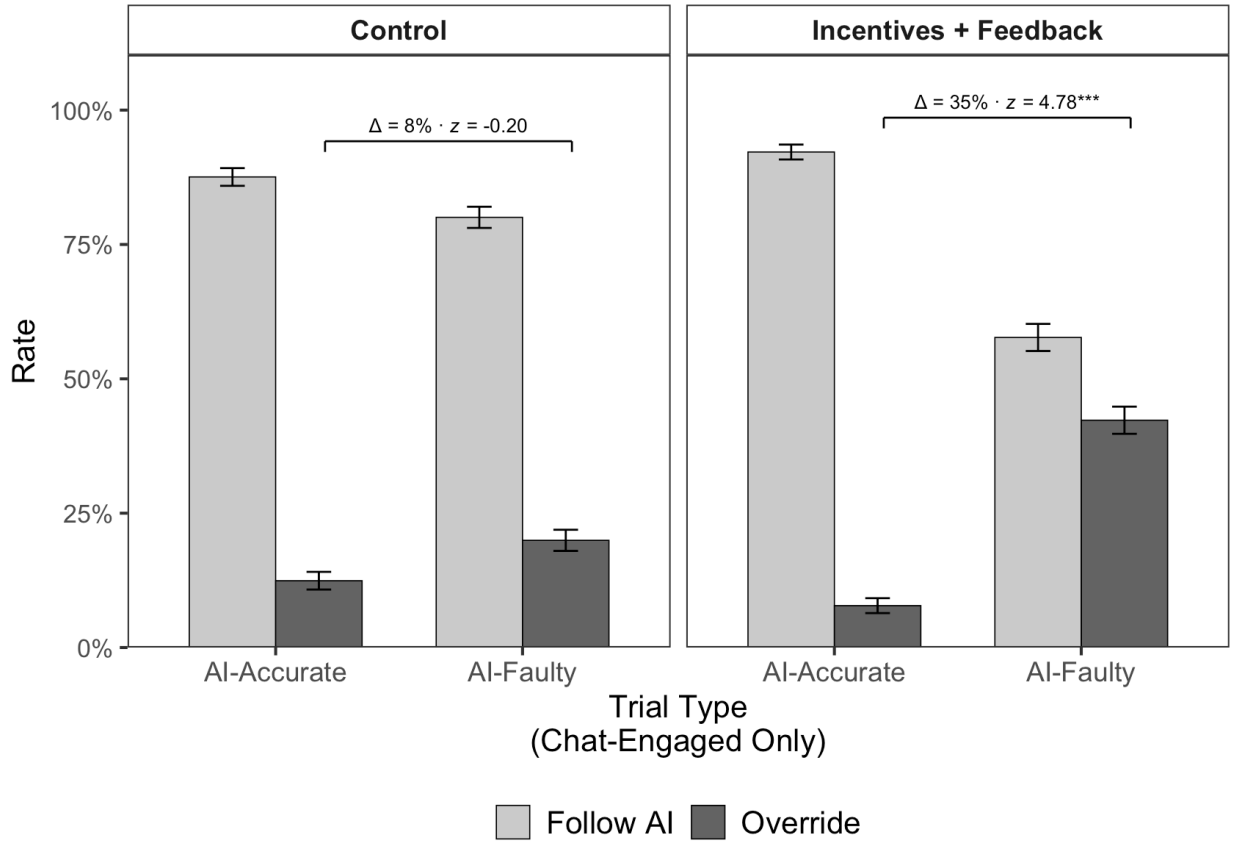
Results

Chat Use & AI Adoption. On chat-engaged AI-Accurate trials, follow rates increased under Incentives + Feedback ($M = 92.2\%$, $SE = 1.4\%$, 95% CI [89.5, 94.9]) compared to Control

(*M* = 87.6%, *SE* = 1.6%, 95% *CI* [84.4, 90.8]; $\beta = 0.66$, *SE* = 0.31, $z = 2.13$, $p = 0.033$; *OR* = 1.94, 95% *CI* [1.05, 3.57]). Similarly, on chat-engaged AI-Faulty trials, Incentives + Feedback more than doubled override rates (i.e., rejecting faulty AI advice; *M* = 42.3%, *SE* = 2.5%, 95% *CI* [37.4, 47.2]) compared to Control (*M* = 20.0%, *SE* = 2.0%, 95% *CI* [16.1, 23.9]; $\beta = 1.44$, *SE* = 0.25, $z = 5.81$, $p = 6.30 \times 10^{-9}$; *OR* = 4.25, 95% *CI* [2.62, 6.90]).

There was a significant Incentives + Feedback × Trial Type interaction (chat-engaged trials only; $\beta = 2.10$, *SE* = 0.35, $z = 6.02$, $p = 1.75 \times 10^{-9}$; *OR* = 8.19, 95% *CI* [4.13, 16.27]), consistent with enhanced discrimination and System 2 engagement. Supporting contrasts confirmed that override rates on AI-Faulty trials increased under Incentives + Feedback (Δ = 34.5%, $z = 4.78$, $p = 1.77 \times 10^{-6}$), but not in the Control condition (Δ = 7.5%, $z = -0.20$, $p = 0.841$; see Figure 7).

Figure 7. Incentives and feedback reduce cognitive surrender to faulty System 3



Notes: Incentives + Feedback increased follow rates when System 3 was accurate and override rates when System 3 was faulty. Error bars indicate SEs.

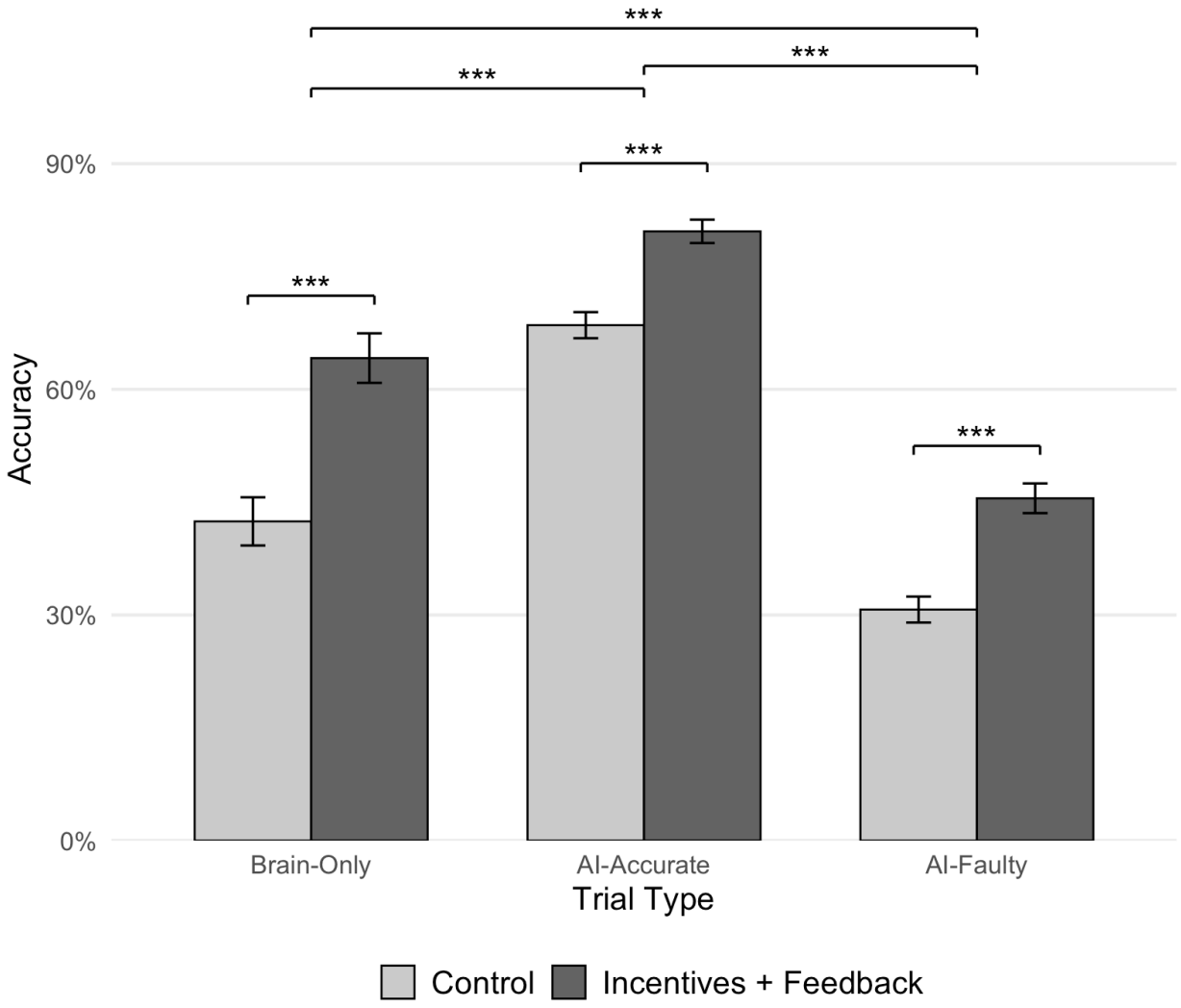
Accuracy Under Incentives + Feedback. On the Brain-Only probe trial, participants in the Incentives + Feedback condition were significantly more accurate ($M = 64.2\%$, $SE = 3.3\%$, 95% CI [57.7, 70.7]) than those in the Control condition ($M = 42.4\%$, $SE = 3.2\%$, 95% CI [36.1, 48.7]; $\beta = 0.89$, $SE = 0.19$, $z = 4.57$, $p = 4.96 \times 10^{-6}$; $OR = 2.43$, 95% CI [1.66, 3.55]), replicating previous findings (e.g., Kluger & DeNisi, 1996; Savine et al., 2010; Frömer et al., 2021).

Collapsed across conditions, accuracy was higher than the Brain-Only probe on AI-Accurate trials (+21.7 pp; $M = 74.4\%$, $SE = 1.2\%$, 95% CI [72.0, 76.8]; $\beta = 1.40$, $SE = 0.15$,

$z = 9.68$, $p < 2.20 \times 10^{-16}$; $OR = 4.06$, 95% CI [3.06, 5.40]) and lower on AI-Faulty trials (-15.0 pp; $M = 37.7\%$, $SE = 1.3\%$, 95% CI [35.2, 40.2]; $\beta = -0.90$, $SE = 0.14$, $z = -6.45$, $p = 1.15 \times 10^{-10}$; $OR = 0.41$, 95% CI [0.31, 0.54]). Among AI-Assisted trials, accuracy was higher on AI-Accurate than AI-Faulty trials ($\beta = 2.27$, $SE = 0.12$, $z = 18.44$, $p < 2.20 \times 10^{-16}$; $OR = 9.70$, 95% CI [7.62, 12.35]), consistent with cognitive surrender.

Among AI-Assisted trials, participants in the Incentives + Feedback condition were more accurate ($M = 63.3\%$, $SE = 1.4\%$, 95% CI [60.6, 66.0]) than those in Control ($M = 49.6\%$, $SE = 1.3\%$, 95% CI [47.1, 52.1]; $\beta = 0.93$, $SE = 0.18$, $z = 5.30$, $p = 1.14 \times 10^{-7}$; $OR = 2.54$, 95% CI [1.80, 3.56]). The interaction between Trial Type and Incentives + Feedback was not significant ($\beta = -0.11$, $SE = 0.22$, $z = -0.51$, $p = 0.609$). Within AI-Faulty trials, participants in the Incentives + Feedback condition were more accurate ($M = 45.5\%$, $SE = 2.0\%$, 95% CI [41.6, 49.4]) than those in Control ($M = 30.7\%$, $SE = 1.7\%$, 95% CI [27.4, 34.0]; $\beta = 1.33$, $SE = 0.29$, $z = 4.55$, $p = 5.27 \times 10^{-6}$; $OR = 3.78$, 95% CI [2.13, 6.69]). Similarly, within AI-Accurate trials, participants were more accurate under Incentives + Feedback ($M = 81.0\%$, $SE = 1.6\%$, 95% CI [77.9, 84.1]) than Control ($M = 68.5\%$, $SE = 1.7\%$, 95% CI [65.2, 71.8]; $\beta = 0.88$, $SE = 0.22$, $z = 3.95$, $p = 7.73 \times 10^{-5}$; $OR = 2.42$, 95% CI [1.56, 3.75]; see Figure 8).

Figure 8. Incentives and feedback increase accuracy, but cognitive surrender persists



Notes: Incentives + Feedback improved accuracy across Trial Types. Error bars indicate SEs. Asterisks denote significance of pairwise contrasts (***$p < 0.001$, **$p < 0.01$, *$p < 0.05$; two-sided).

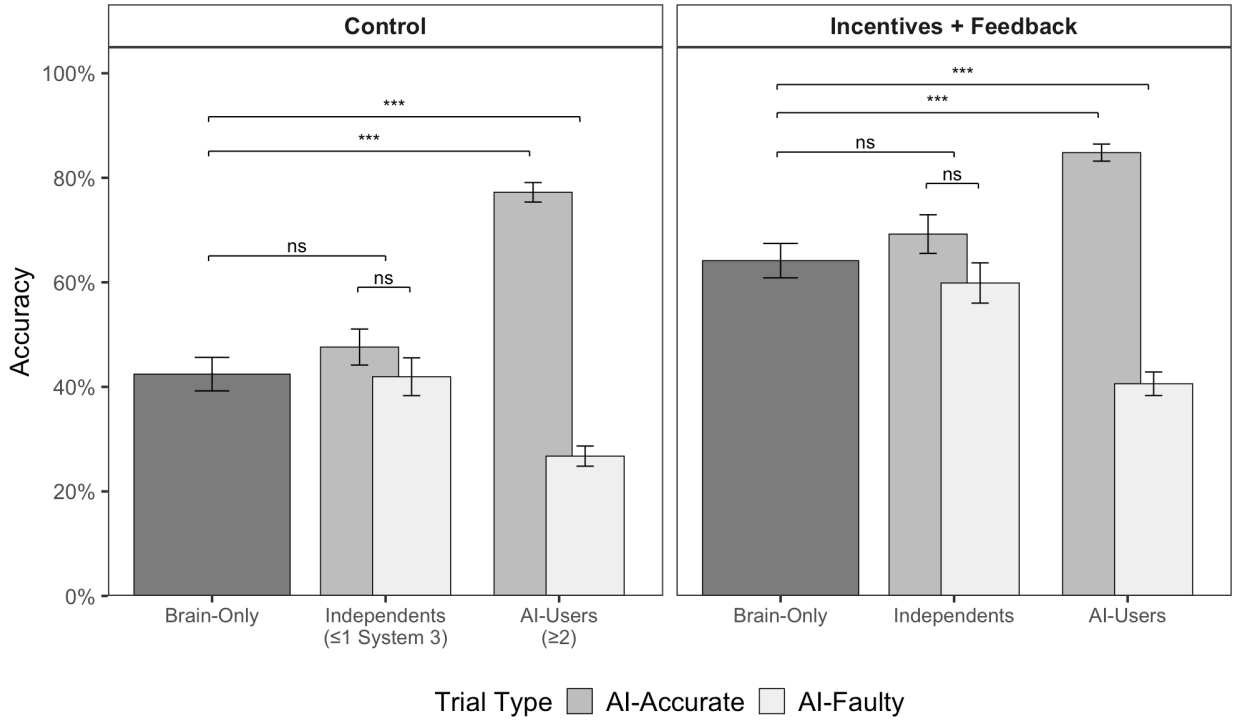
Confidence. Per-item confidence on AI-Assisted trials was higher ($M = 82.2\%$, $SE = 0.5\%$, 95% CI [81.3, 83.1]) than on the Brain-Only probe trial ($M = 77.5\%$, $SE = 1.2\%$, 95% CI [75.1, 79.9]; $\beta = 0.19$, $SE = 0.04$, $t = 4.95$, 95% CI [0.12, 0.27], $p = 6.86 \times 10^{-7}$). Confidence

did not vary significantly by Trial Type (AI-Accurate vs. AI-Faulty; $\beta = 0.01$, $SE = 0.04$, $t = 0.41$, $p = 0.682$), nor did it reliably increase under Incentives + Feedback ($\beta = -0.12$, $SE = 0.07$, $t = -1.84$, $p = 0.066$). Among AI-Assisted trials, and controlling for Trial Type and Incentives + Feedback, a 10-point increase in per-item confidence was associated with 22% greater odds of answering correctly ($\beta = 0.20$, $SE = 0.03$, $z = 7.49$, $p = 7.15 \times 10^{-14}$; $OR = 1.22$, 95% CI [1.16, 1.29]).

Thinking Profiles. As in Study 2, Independents (≤1 System 3 use) largely mirrored Brain-Only patterns. Accuracy among Independents increased from 44.7% ($SE = 3.2\%$) in the Control condition to 64.6% ($SE = 3.5\%$) under Incentives + Feedback ($\beta = 0.77$, $SE = 0.19$, $z = 4.03$, $p = 5.59 \times 10^{-5}$; $OR = 2.16$, 95% CI [1.50, 3.12]); there was no significant interaction with Trial Type ($\beta = 0.06$, $SE = 0.26$, $z = 0.23$, $p = 0.818$; $OR = 1.06$, 95% CI [0.63, 1.76]). Independents showed similar accuracy on AI-Assisted Trial Types within the Control (AI-Accurate: $M = 47.6\%$, $SE = 3.4\%$; AI-Faulty: $M = 41.9\%$, $SE = 3.6\%$; $\beta = 0.31$, $SE = 0.26$, $z = 1.18$, $p = 0.239$; $OR = 1.36$, 95% CI [0.81, 2.27]) and Incentives + Feedback conditions (AI-Accurate: $M = 69.2\%$, $SE = 3.7\%$; AI-Faulty: $M = 59.9\%$, $SE = 3.8\%$; $\beta = 0.44$, $SE = 0.31$, $z = 1.41$, $p = 0.159$; $OR = 1.55$, 95% CI [0.84, 2.83]).

In contrast, AI-Users (≥2 trials using System 3) performed well in the Control condition under AI-Accurate ($M = 77.2\%$, $SE = 1.9\%$) and poorly under AI-Faulty ($M = 26.8\%$, $SE = 1.9\%$), replicating the cognitive surrender effect observed in Studies 1 and 2 ($\beta = 3.25$, $SE = 0.21$, $z = 15.4$, $p < 2.20 \times 10^{-16}$; $OR = 25.74$, 95% CI [17.04, 38.88]). Incentives + Feedback increased AI-Users accuracy to 84.8% ($SE = 1.6\%$) on AI-Accurate trials ($\beta = 0.72$, $SE = 0.29$, $z = 2.47$, $p = 0.014$; $OR = 2.06$, 95% CI [1.16, 3.65]) and 40.6% ($SE = 2.3\%$) on AI-Faulty trials ($\beta = 1.06$, $SE = 0.27$, $z = 3.90$, $p = 9.80 \times 10^{-5}$; $OR = 2.89$, 95% CI [1.70, 4.94]; see Figure 9).

Figure 9. Incentives and feedback improve System 3 calibration in AI-Users



Notes: Whereas Independents' accuracy is improved with Incentives + Feedback, AI-Users continued to produce significantly higher/lower accuracy on System 3 accurate/faulty trials, consistent with cognitive surrender. Under Incentives + Feedback, participants who used System 3 (and it was accurate) had the highest accuracy. Error bars indicate SEs. Asterisks denote significance of pairwise contrasts (***$p < 0.001$, **$p < 0.01$, *$p < 0.05$; two-sided).

A Trial Type × Thinking Profile interaction revealed that AI-Users were more sensitive to Trial Type than Independents ($\beta = 3.01$, $SE = 0.31$, $z = 9.65$, $p < 2.20 \times 10^{-16}$, $OR = 20.23$, 95% CI [10.98, 37.27]). However, an exploratory three-way interaction with Incentives + Feedback ($\beta = -0.36$, $SE = 0.58$, $z = -0.62$, $p = 0.534$, $OR = 0.70$, 95% CI [0.22, 2.19]), suggested that the manipulation benefited both groups without fundamentally altering their distinct response profiles.

Discussion

Study 3 provides evidence that a combined Incentives + Feedback manipulation can facilitate System 2 engagement and reduce cognitive surrender to System 3. Nevertheless, cognitive surrender persists. Participants rewarded for accuracy and given immediate item-level feedback were significantly more accurate, particularly in cases when AI recommendations may have led them astray (i.e., AI-Faulty trials). Incentives + Feedback reduced following of incorrect AI advice and increased override behavior, indicating that participants actively monitored and corrected System 3 outputs when motivated to do so (incentives) and able to course correct (feedback).

The pattern was especially pronounced in AI-Users, who showed a large asymmetry in performance between AI-Accurate and AI-Faulty trials under Incentives + Feedback (*OR* = 24.37, 95% *CI* [15.11, 39.30]). Under Incentives + Feedback, accuracy for AI-Users increased on both AI-Accurate (77.2% to 84.8%) and AI-Faulty trials (26.8% to 40.6%). However, the accuracy gap between Trial Types remained large, at ~44 percentage points under Incentives + Feedback (compared to ~50 pp under Control). Trial-level confidence mirrored patterns in Studies 1 and 2 (access to System 3 inflates confidence), but revealed that confidence ratings retain some diagnostic value.

These results support the Tri-System Theory notion that external motivational cues and timely diagnostic feedback can help shift participants from cognitive surrender (System 3 dominance) toward more override and deliberative (re)engagement (see Figure 1).

General Discussion

We propose and examine a novel cognitive framework, Tri-System Theory, which extends the dual-process architecture of judgment by incorporating a third system: System 3 (artificial cognition). System 3 reflects external, automated, data-driven, and dynamic reasoning performed by AI systems, and operates alongside internal intuitive (System 1) and deliberative (System 2) processes. Across three experiments using a canonical cognitive reasoning task, we provide empirical support for Tri-System Theory by illustrating how System 3 operates, and how reliance on it can bypass System 1/2 (cognitive surrender) and/or supplement System 2 (cognitive offloading).

Our findings show that decision-makers use System 3 and frequently adopt its outputs, even when those outputs are systematically incorrect (Study 1). This behavioral pattern, cognitive surrender, reflects an uncritical reliance on System 3 and persists under conditions that suppress as well as enhance analytical reasoning, such as time pressure (Study 2) and incentives/feedback (Study 3), respectively. Further, we observed elevated confidence when System 3 was engaged, despite consistent error. These studies illustrate the foundational dynamics of a new triadic cognitive ecology under Tri-System Theory, cementing System 3 as a valuable cognitive resource, often substituting or supplementing System 1, while reducing demands on System 2.

Cognitive Surrender Effect Size

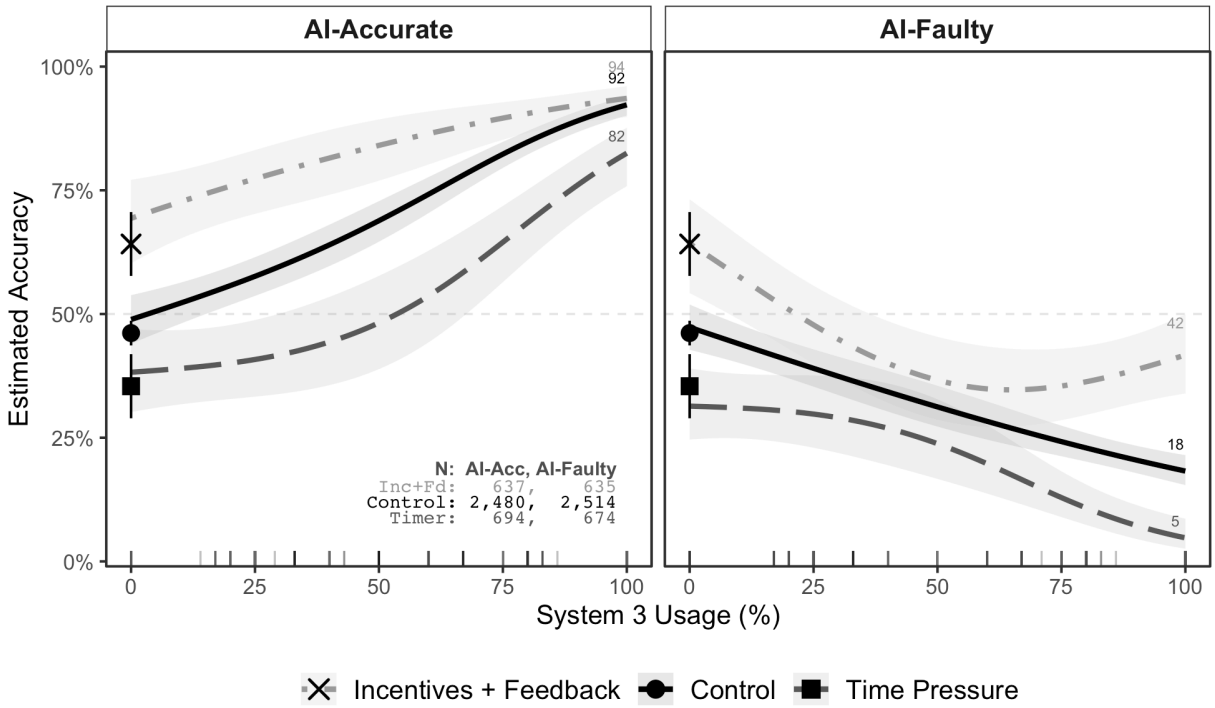
To precisely estimate the cognitive surrender effect size, we preregistered a trial-level synthesis of Studies 1–3 (N = 1,372; 9,593 trials)[1]. This allowed us to assess robustness across samples/platforms, and test moderation by situation (Time Pressure; Incentives + Feedback) and

[1] The ancillary Study 1 online sample was included in the trial-level synthesis.

individual differences (Trust in AI, NFC, Fluid IQ). Inclusion and modeling decisions were prespecified (Vosgerau et al., 2019).

Cognitive surrender was robust across studies. Correct responding was over 16 times greater when System 3 was correct (*OR* = 16.07, 95% *CI* [11.50, 22.46]; $z = 16.30$, $p < 2.20 \times 10^{-16}$). Per-study Cohen's *h* values were large (S1 = 0.83, S2 = 0.86, S3 = 0.78; trial-weighted $h = 0.82$). Situational manipulations shifted baselines, but cognitive surrender persisted: time pressure reduced accuracy (Brain-Only -13.5 pp; AI-Accurate -13.5 pp; AI-Faulty -11.3 pp), whereas Incentives + Feedback improved it (Brain-Only +18.1 pp; AI-Accurate +9.8 pp; AI-Faulty +14.4 pp). The AI-Accurate versus AI-Faulty gap remained large under both Time Pressure (*OR* = 14.28, 95% *CI* [8.51, 23.98]) and Incentives + Feedback (*OR* = 11.05, 95% *CI* [6.48, 18.83]). Figure 10 shows accuracy scaling with System 3 use: as reliance increases, performance tracks AI quality, rising when accurate and falling when faulty, illustrating the promises of superintelligence and exposing a structural vulnerability of cognitive surrender.

Figure 10. Cognitive surrender as a function of System 3 usage across three studies



Notes: Participant-level logistic models estimated accuracy as a function of System 3 usage by Trial Type for three experimental conditions (all with optional System 3 access): Control, Time Pressure (30 second deadline), and Incentives + Feedback ($0.20 per-item accuracy bonus and feedback). Shaded curves show model-predicted 95% CIs. Rug plots indicate the distribution of participant reliance. Inset sample sizes present AI-trial counts by condition. Brain-Only trial counts: Inc+Fd = 212, Control = 1,530, Timer = 227. Anchors (0% usage) mark Brain-Only performance. Accuracy rose with System 3 reliance when outputs were accurate but declined sharply when outputs were faulty, highlighting both the benefits and vulnerabilities of cognitive surrender. Incentives and feedback improved accuracy, while time pressure impaired it; neither manipulation eliminated the large performance gap between accurate and faulty System 3 advice indicative of cognitive surrender.

Individual Differences in Susceptibility to Cognitive Surrender. Participants who scored higher on Trust in AI were more likely to follow System 3 outputs, leading to greater divergence between AI-Accurate and AI-Faulty trials ($OR = 2.81$, 95% CI [2.36, 3.34]; $z = 11.70$, $p < 2.20 \times 10^{-16}$). In contrast, those higher in NFC were more resistant to System 3 influence, showing smaller accuracy differences between AI-Accurate and AI-Faulty trials ($OR = 0.83$, 95% CI [0.70, 0.98]; $z = -2.14$, $p = 0.032$). Fluid IQ showed a similar protective pattern ($OR = 0.69$, 95% CI [0.57, 0.83]; $z = -3.93$, $p = 8.64 \times 10^{-5}$). Together, these findings indicate that individuals with greater deliberative capacity or motivation were less prone to cognitive surrender, while those who trusted the AI more were more vulnerable to it.

Cognitive Surrender vs. Cognitive Offloading. On trials where System 3 was engaged and incorrect, following the incorrect AI output indexed cognitive surrender; conversely, overriding it and answering correctly indexed cognitive offloading. Across studies, 73.2% of such trials showed cognitive surrender, 19.7% cognitive offloading, and 7.1% failed overrides (resisted AI but still answered incorrectly). Situational factors shifted these shares: Incentives + Feedback increased cognitive offloading by ~19 pp to 37.1% and reduced surrender to 57.9% (failed overrides = ~5%); Time Pressure cut offloading by ~12 pp to 6.2% and raised failed overrides by ~8 pp to 13.8%. Logistic models converged with these patterns: Incentives + Feedback promoted cognitive offloading over cognitive surrender ($OR = 0.38$, 95% CI [0.15, 0.93], $p = 0.034$), whereas Time Pressure increased cognitive surrender ($OR = 3.19$, 95% CI [1.03, 9.88], $p = 0.044$). Individual differences revealed that higher Trust in AI increased cognitive surrender over cognitive offloading ($OR = 4.36$, 95% CI [2.95, 6.46], $p = 1.77 \times 10^{-13}$), whereas Need for Cognition ($OR = 0.46$, 95% CI [0.33, 0.64], $p = 3.12 \times 10^{-6}$) and Fluid IQ ($OR = 0.37$, 95% CI [0.26, 0.51], $p = 4.23 \times 10^{-9}$) predicted greater resistance to surrender and more offloading.

Theoretical Implications

Tri-System Theory extends traditional dual-process frameworks, long anchored in the distinction between internal intuition (System 1) and deliberation (System 2). System 3 is engaged externally, crossing the brain boundary, and may be engaged via cognitive surrender, cognitive offloading, or without internal cognition at all (autopilot). Our findings underscore how System 3 integrates into the cognitive landscape and fundamentally reshapes traditional assumptions underlying decision-making. Considering System 3, we must re-evaluate the limits of human cognition, agency, and judgment (Gonzalez & Heidari, 2025).

A consequence of System 3 is the introduction of cognitive surrender, characterized by uncritical reliance on externally generated artificial reasoning, bypassing System 2. Crucially, we distinguish cognitive surrender, marked by passive trust and uncritical evaluation of external information, from cognitive offloading, which involves strategic delegation of cognition during deliberation.

Mechanisms Underlying Cognitive Surrender. Across our studies, we observe that when System 3 was available, people readily engaged it and frequently adopted its answers. This shift reflects a reallocation of cognitive control rather than mere effort saving. System 3's fluent, confident outputs are treated as epistemically authoritative, lowering the threshold for scrutiny and attenuating the metacognitive signals that would ordinarily route a response to deliberation. In the case of cognitive surrender, there is a shift in the locus of control, with an external system (System 3) occupying the default position.

Previous research has shown that people often curtail later AI use after observing an algorithm err ('algorithmic aversion'; Dietvorst, Simmons, & Massey, 2015), yet in structured tasks, they frequently prefer algorithmic over human judgment ('algorithm appreciation'; Logg,

Minson, & Moore, 2019). In the present studies, we observed heterogeneity in System 3 usage ('Independents' vs. 'AI-Users'), but more participants chose to use System 3 than not, and participants reported higher confidence when using AI. More broadly, research has shown that AI can alter which internal processes are recruited even when it is not supplying an answer (Goergen et al., 2025).

Time constraints clarify why surrender arises so readily, while incentives and feedback show that surrender is malleable. When decision time is scarce, the internal monitor detecting conflict and recruiting deliberation is less likely to trigger. Hence, the low-friction path to defer to external cognition becomes attractive. Conversely, making consequences salient and providing item-level correctness signals encourages verification and selective override without bluntly suppressing helpful assistance. The mechanism is best understood as a targeted reactivation of System 2 (i.e., increased cognitive offloading over surrender): users pause, check, and depart from guidance when it conflicts with available cues, yet continue to benefit when AI guidance is sound.

Our trial-level synthesis shows the strength and importance of System 3 engagement: accuracy followed a dose–response relationship, tracking AI accuracy. Importantly, this pattern is robust across situational moderators, indicating that cognitive surrender is a native dynamic of the triadic ecology rather than a by-product of any single context. Situational moderators shifted intercepts, but the dominant force was System 3 usage: the cognitive surrender effect was consistently large (*OR* = 15.50 baseline, 14.28 under time pressure, and 11.05 with incentives and feedback), and a much larger effect than the shifts produced by time pressure (*OR* = 0.42) or motivational manipulations (*OR* = 3.45). Therefore, beyond access, the key cognitive

architecture of System 3 relies on dosage and automaticity of use; calibrated engagement of System 3 may provide cognitive benefits while limiting negative consequences.

Societal Implications

The growing integration of AI into society, spanning domains like finance, healthcare, education, and e-commerce, amplifies the relevance of Tri-System Theory. Our findings demonstrate that people readily incorporate AI-generated outputs into their decision-making processes, often with minimal friction or skepticism. This seamless engagement with System 3 underscores its potential to enhance everyday cognition by reducing cognitive effort, accelerating decisions, and supplementing or substituting internal cognition with externally processed, vastly resourced, AI-powered insights.

At the same time, our experiments highlight the ease with which individuals adopt AI-generated suggestions without scrutiny. These findings raise important questions about how decision-makers engage with AI under conditions of uncertainty or error. For example, in contexts such as financial advice, medical triage, or legal decision support, uncritical evaluation of System 3 could result in significant harm and a lack of personal accountability for serious life outcomes. In therapeutic contexts, conversational AI may amplify delusional beliefs in vulnerable users, a risk termed ‘AI psychosis’ (Hudon & Stip, 2025).

Rather than sounding alarm bells, we view the vulnerabilities of cognitive surrender to System 3 as a design and education challenge: how can we support decision-makers in using System 3 effectively while maintaining critical thinking and accountability when necessary? Our results suggest that giving feedback and aligning incentives may help people engage System 2 when needed, without diminishing the efficiency gains provided by System 3. Features

accompanying AI outputs, such as confidence scores, uncertainty indicators, or transparent explanations may serve as lightweight cues that encourage thoughtful decision-making.

AI Design and Policy Recommendations. Tri-System Theory offers a cognitive framework for designing AI systems that support human reasoning. Designers of AI interfaces should consider when and how to trigger different cognitive systems. Rather than fully automating choices, effective AI design may encourage calibrated collaboration, where System 3 enhances and collaborates with internal cognition. This may include customizable modes that align with user preferences for autonomy versus assistance, adaptive nudging that signals uncertainty, domain-specific cautionary filters, or interfaces that dynamically adjust cognitive demands based on context.

From a policy perspective, these insights call for renewed attention to education and digital literacy. As AI continues to shape decision-making environments, it will be critical to equip users with the digital literacy tools to understand when and how to trust System 3. Just as ‘trust your gut’ may align and sway judgment towards System 1 intuition, ‘trust the data’ may sway people towards System 3. The context in which such phrases are used, however, is important. While complete algorithmic transparency may be unrealistic or unnecessary, understanding and indicating the reliability of AI recommendations, such as whether an answer is grounded, probabilistic, or uncertain, may help individuals strategically regulate System 3 usage.

Limitations and Future Research Directions

Several limitations warrant consideration and suggest opportunities for future work. First, our studies were conducted in controlled experimental environments, which allowed us to isolate cognitive mechanisms with high internal validity but may limit generalizability to real-world

settings. In everyday contexts, AI interactions are embedded within broader ecological, social, and motivational systems. Future research should extend these findings using field experiments outside the lab or in naturalistic studies, such as exploring System 3 decision-making in financial apps, online health platforms, or retail recommendation systems.

Second, we used the CRT as our core task across studies. While the CRT is well-validated, foundational, and widely used to study reasoning, it captures a specific form of cognition. Future work should generalize System 3 by incorporating additional tasks that reflect other cognitive domains (e.g., probabilistic reasoning, everyday decision-making, moral judgment) to test the dynamics of Tri-System Theory across reasoning types.

Third, the current studies provide a snapshot of decision-making under a single set of exposures to System 3. In real-world contexts, decision-makers often interact with AI repeatedly over time, involving feedback, learning, and evolving trust. As such, memory capabilities of System 3 may alter cognitive dynamics of trust, reliability, and personalization. Longitudinal or adaptive paradigms could explore how repeated engagement with System 3 alters the dynamics of cognitive surrender, calibration, and override. Does trust in System 3 grow or attenuate with experience? How do people learn to detect or correct faulty AI outputs over time?

Finally, our research opens important questions about how other situational moderators and individual differences influence tri-system processes. For example, other forms of cognitive load (e.g., task framing) or personal accountability (e.g., social presence) may serve as additional situational moderators of cognitive surrender. At the same time, additional individual differences, such as age, gender, personality, neurotypicality, or technological ability, may influence how people engage with System 3.

Conclusion

As AI becomes ubiquitous in society, understanding how it reshapes human thought is essential. Tri-System Theory offers a new framework for this cognitive frontier. By introducing System 3 (Artificial) as a distinct and external reasoning process, we move beyond the classical architecture of dual-process theories and chart a new decision-making paradigm: one where intuition, deliberation, and artificial cognition co-exist, compete, or converge. We show that people not only use System 3 to assist with reasoning, but often surrender to its outputs—whether correct or flawed. This cognitive surrender illustrates the value and integration of System 3, but also highlights the vulnerability of System 3 usage. Similar to how System 1-driven heuristics lead to systematic biases, System 3 has differential cognitive shortcomings that will challenge decision-makers and society at large.

Tri-System Theory is not a warning about AI's dangers but a recognition of System 3's psychological presence. We do not merely use AI; we think with it. In doing so, we must ask new questions: What happens when our judgments are shaped by minds not our own? What becomes of intuition and effort when a generative, artificial partner stands ready to answer? How do we preserve agency, reflection, and autonomy in a world where users engage in cognitive surrender? We offer Tri-System Theory as a conceptual foundation for understanding these challenges. It is a theory for an age of human-AI algorithmic cognition, and for the decision-makers, researchers, and designers shaping that future.

References

Balleine, B. W., & O'Doherty, J. P. (2010). Human and rodent homologies in action control: Corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology, 35*(1), 48–69. https://doi.org/10.1038/npp.2009.131

Bettman, J. R., Luce, M. F., & Payne, J. W. (1998). Constructive consumer choice processes. *Journal of Consumer Research, 25*(3), 187–217. https://doi.org/10.1086/209535

Bilancini, E., Boncinelli, L., & Celadin, T. (2024). Manipulating response times in the cognitive reflection test: Time delay boosts deliberation, time pressure hinders it. *Journal of Behavioral and Experimental Economics, 112*, 102273. https://doi.org/10.1016/j.socec.2024.102273

Budzyń, K., Romańczyk, M., Kitala, D., Kołodziej, P., Bugajski, M., Adami, H. O., Blom, J., Buszkiewicz, M., Halvorsen, N., Hassan, C., Romańczyk, T., Holme, Ø., Jarus, K., Fielding, S., Kunar, M., Pellise, M., Pilonis, N., Kamiński, M. F., Kalager, M., … Mori, Y. (2025). Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: A multicentre, observational study. *The Lancet Gastroenterology & Hepatology, 10*(10), 896–903. https://doi.org/10.1016/S2468-1253(25)00133-5

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42*(1), 116–131. https://doi.org/10.1037/0022-3514.42.1.116

Chiriatti, M., Ganapini, M., Panai, E., Ubiali, M., & Riva, G. (2024). The case for human–AI interaction as system 0 thinking. *Nature Human Behaviour, 8*(10), 1829–1830. https://doi.org/10.1038/s41562-024-01995-5

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis, 58*(1), 7–19. https://doi.org/10.1093/analys/58.1.7

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience, 8*(12), 1704–1711. https://doi.org/10.1038/nn1560

De Neys, W. (2006). Dual processing in reasoning: Two systems but one reasoner. *Psychological Science, 17*(5), 428–433. https://doi.org/10.1111/j.1467-9280.2006.01723.x

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56*(1), 5–18. https://doi.org/10.1037/0022-3514.56.1.5

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*(1), 114–126. https://doi.org/10.1037/xge0000033

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron, 80*(2), 312–325. https://doi.org/10.1016/j.neuron.2013.09.007

Epstein, S. (1994). Integration of the cognitive and the psychodynamic unconscious. *American Psychologist, 49*(8), 709–724. https://doi.org/10.1037/0003-066X.49.8.709

Evans, J. St. B. T. (2006). The heuristic-analytic theory of reasoning: Extension and evaluation. *Psychonomic Bulletin & Review, 13*(3), 378–395. https://doi.org/10.3758/BF03193858

Evans, J. St. B. T., & Curtis-Holmes, J. (2005). Rapid responding increases belief bias: Evidence for the dual-process theory of reasoning. *Thinking & Reasoning, 11*(4), 382–389. https://doi.org/10.1080/13546780542000005

Evans, J. St. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science, 8*(3), 223–241. https://doi.org/10.1177/1745691612460685

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19*(4), 25–42. https://doi.org/10.1257/089533005775196732

Frömer, R., Lin, H., Wolf, C. K. D., Inzlicht, M., & Shenhav, A. (2021). Expectations of reward and efficacy guide cognitive control allocation. *Nature Communications, 12*, Article 1030. https://doi.org/10.1038/s41467-021-21315-z

Goergen, J., de Bellis, E., & Klesse, A.-K. (2025). AI assessment changes human behavior. *Proceedings of the National Academy of Sciences of the United States of America, 122*(25), e2425439122. https://doi.org/10.1073/pnas.2425439122

Gonzalez, C., & Heidari, H. (2025). A cognitive approach to human–AI complementarity in dynamic decision-making. *Nature Reviews Psychology, 4*, 808–822. https://doi.org/10.1038/s44159-025-00499-x

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*(5537), 2105–2108. https://doi.org/10.1126/science.1062872

Hinson, J. M., Jameson, T. L., & Whitney, P. (2003). Impulsive decision making and working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(2), 298–306. https://doi.org/10.1037/0278-7393.29.2.298

Hudon, A., & Stip, E. (2025). Delusional experiences emerging from AI chatbot interactions or "AI Psychosis". *JMIR Mental Health, 12*, e85799. https://doi.org/10.2196/85799

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics, 4*(1), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist, 58*(9), 697–720. https://doi.org/10.1037/0003-066X.58.9.697

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Keren, G., & Schul, Y. (2009). Two is not always better than one: A critical evaluation of two-system theories. *Perspectives on Psychological Science, 4*(6), 533–550. https://doi.org/10.1111/j.1745-6924.2009.01164.x

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254–284. https://doi.org/10.1037/0033-2909.119.2.254

Kruglanski, A. W., & Gigerenzer, G. (2011). Intuitive and deliberate judgments are based on common principles. *Psychological Review, 118*(1), 97–109. https://doi.org/10.1037/a0020762

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*(3), 480–498. https://doi.org/10.1037/0033-2909.108.3.480

Lieberman, M. D. (2007). Social cognitive neuroscience: A review of core processes. *Annual Review of Psychology, 58*, 259–289. https://doi.org/10.1146/annurev.psych.58.110405.085654

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

Lugo, R. G., Sütterlin, S., Knox, B. J., Jøsok, Ø., Helkala, K., & Lande, N. M. (2016). The moderating influence of self-efficacy on interoceptive ability and counterintuitive decision making in officer cadets. *Journal of Military Studies, 7*(1), 44–52. https://doi.org/10.1515/jms-2016-0005

Lyall, D. M., Cullen, B., Allerhand, M., Smith, D. J., Mackay, D., Evans, J., Anderson, J., Fawns-Ritchie, C., McIntosh, A. M., Deary, I. J., & Pell, J. P. (2016). Cognitive test scores in UK Biobank: Data reduction in 480,416 participants and longitudinal stability in 20,346 participants. *PLOS ONE, 11*(4), e0154222. https://doi.org/10.1371/journal.pone.0154222

Lynch, M. P. (2016). *The internet of us: Knowing more and understanding less in the age of big data*. Liveright Publishing.

Manfredi, D. A., & Nave, G. (2022). Beyond the bat and the ball: Overcoming familiarity effects in the Cognitive Reflection Test by rewording its questions. *SSRN*. https://doi.org/10.2139/ssrn.3445138

Melnikoff, D. E., & Bargh, J. A. (2018). The mythical number two. *Trends in Cognitive Sciences, 22*(4), 280–293. https://doi.org/10.1016/j.tics.2018.02.001

Mosier, K. L., & Skitka, L. J. (1996). Human decision makers and automated decision aids: Made for each other? In R. Parasuraman & M. Mouloua (Eds.), *Automation and human performance: Theory and applications* (pp. 201–220). Lawrence Erlbaum Associates.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). *The adaptive decision maker*. Cambridge University Press.

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition, 188*, 39–50. https://doi.org/10.1016/j.cognition.2018.06.011

Pennycook, G., & Rand, D. G. (2020). Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of Personality, 88*(2), 185–200. https://doi.org/10.1111/jopy.12476

Petty, R. E., & Cacioppo, J. T. (1986). *Communication and persuasion: Central and peripheral routes to attitude change*. Springer-Verlag.

Rainey, P. B., & Hochberg, M. E. (2025). Could humans and AI become a new evolutionary individual? *Proceedings of the National Academy of Sciences of the United States of America, 122*(37), e2509122122. https://doi.org/10.1073/pnas.2509122122

Raoelison, M., Thompson, V. A., & De Neys, W. (2020). The smart intuitor: Cognitive capacity predicts intuitive rather than deliberate thinking. *Cognition, 204*, Article 104381. https://doi.org/10.1016/j.cognition.2020.104381

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences, 20*(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002

Roets, A., & Van Hiel, A. (2011). Item selection and validation of a brief, 15-item version of the Need for Closure Scale. *Personality and Individual Differences, 50*(1), 90–94. https://doi.org/10.1016/j.paid.2010.09.004

Savine, A. C., Beck, S. M., Edwards, B. G., Chiew, K. S., & Braver, T. S. (2010). Enhancement of cognitive control by approach and avoidance motivational states. *Cognition & Emotion, 24*(2), 338–356. https://doi.org/10.1080/02699930903381564

Shiv, B., & Fedorikhin, A. (1999). Heart and mind in conflict: The interplay of affect and cognition in consumer decision making. *Journal of Consumer Research, 26*(3), 278–292. https://doi.org/10.1086/209563

Simonson, I. (1989). Choice based on reasons: The case of attraction and compromise effects. *Journal of Consumer Research, 16*(2), 158–174. https://doi.org/10.1086/209205

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*(1), 3–22. https://doi.org/10.1037/0033-2909.119.1.3

Slovic, P., Finucane, M. L., Peters, E., & MacGregor, D. G. (2004). Risk as analysis and risk as feelings: Some thoughts about affect, reason, risk, and rationality. *Risk Analysis, 24*(2), 311–322. https://doi.org/10.1111/j.0272-4332.2004.00433.x

Spatharioti, S. E., Rothschild, D., Goldstein, D. G., & Hofman, J. M. (2025). Effects of LLM-based search on decision making: Speed, accuracy, and overreliance. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3706598.3714082

Stanovich, K. E. (2009). *What intelligence tests miss: The psychology of rational thought*. Yale University Press.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate. *Behavioral and Brain Sciences, 23*(5), 645–665. https://doi.org/10.1017/S0140525X00003435

Thompson, V. A., Prowse Turner, J. A., & Pennycook, G. (2011). Intuition, reason, and metacognition: Insights into their interplay. *Cognitive Psychology, 63*(3), 107–140. https://doi.org/10.1016/j.cogpsych.2011.06.001

Tully, S. M., Longoni, C., & Appel, G. (2025). Lower artificial intelligence literacy predicts greater AI receptivity. *Journal of Marketing*. https://doi.org/10.1177/00222429251314491

Vosgerau, J., Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2019). 99% impossible: A valid, or falsifiable, internal meta-analysis. *Journal of Experimental Psychology: General, 148*(9), 1628–1639. https://doi.org/10.1037/xge0000593

Wegner, D. M. (1987). Transactive memory: A contemporary analysis of the group mind. In B. Mullen & G. R. Goethals (Eds.), *Theories of group behavior* (pp. 185–208). Springer-Verlag.

Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist, 35*(2), 151–175. https://doi.org/10.1037/0003-066X.35.2.151