Causal Error-Potential Gating and Latency-Bounded Active Co-Inference for Shared-Autonomy Brain–Computer Interfaces: Formulation and Sensitivity Analysis

Dima Baker
Collective Intelligence Network

November 2026

Abstract

Brain–computer interfaces (BCIs) coupled with language and control agents promise shared autonomy, yet popular claims of "computational consciousness" conflate functional control coupling with phenomenal sentience rooted in subcortical homeostatic circuitry. Moreover, non-invasive EEG-to-text decoders operate only in closed-set regimes and suffer from teacher-forcing inflation, while unguarded completions induce sycophantic prior capture. We formulate *Causal Error-Potential-Gated Active Co-Inference* (Causal-ErrP-CoInf), coupling a human operator and an on-device speculative agent across a Markov blanket. First, we model coupled cortical–agent tracking as a linear delay differential system with frequency $\omega_c = 17.5$ rad/s, deriving a critical feedback delay $\Delta_{\text{crit}} = \pi/(2\omega_c) \approx 89.8$ ms above which uncompensated loops lose phase stability. Second, to respect physiological causality, the agent renders a preview cue $\tilde{a}_A(t)$ after delay Δ_{fb} and holds commitment over the ErrP window $\tau_{\text{ErrP}} \in [260, 330]$ ms, gating execution via a shrinkage-LDA classifier on 48-dimensional post-stimulus ERP waveforms and response-time hesitation. Across 26,400 episodes ($N_{\text{mis}} = 600$ misaligned and $N_{\text{aligned}} = 600$ aligned per cell), uncompensated baselines collapse to 0.00% [0.00, 0.64] phase lock at $\Delta_{\text{fb}} = 180$ ms, while Causal-ErrP-CoInf cuts unrequested intent drift at $\Delta_{\text{fb}} = 45$ ms from 65.50% [61.61, 69.19] and 28.50% [25.03, 32.24] to 9.33% (95% Wilson CI [7.26, 11.93]; 87.83% [84.97, 90.21] phase lock; 339.8 ms commit latency; 1.85 μs LDA CPU time), with break-even $\widehat{\text{TPR}}^{*}_{\text{ErrP}} \approx 0.74$ at $\text{SNR}_{\text{EEG}} \geq 0.80$.

1 Introduction

Since Licklider's foundational formulation of man–computer symbiosis [24], Clark and Chalmers' extended-mind hypothesis [9], and formal policy-blending frameworks in shared-control robotics [12], human–machine interaction has been constrained by the low efferent bandwidth of manual keystrokes and the perceptual latency of turn-based displays. Recent advances in non-invasive physiological sensing—including context-conditioned hierarchical electroencephalography-to-text (EEG-to-text) alignment [25], multimodal physiological cognitive-state recognition [18], and post-stimulus Error-Related Potential (ErrP) decoding [7]—have prompted renewed interest in tightly coupled neural–agent interfaces. Concurrently, speculative commentary has invoked "computational consciousness" to describe low-latency loops between human brains and foundation models [26, 36].

1.1 Why "computational consciousness" and unconstrained EEG mind-reading fail scientifically

A rigorous engineering formulation of neural–agent shared autonomy must begin by rejecting two widespread misconceptions in popular and technical discourse:

First, coupling neocortical recordings to an autoregressive transformer does not synthesize phenomenal consciousness or merge minds. As established in affective and clinical neuroscience [15, 29, 32], baseline phenomenal sentience and affective valence originate in ancient subcortical brainstem and interoceptive homeostatic regulation rather than neocortical sequence prediction. Artificial neural networks possess neither visceral interoceptive plants nor metabolic survival stakes. As Wilson et al. [36] demonstrate, Active Inference provides a formal mathematical method for *phenotyping functional agency*—characterizing how an artifact maintains a statistical boundary (a Markov blanket [15, 17]) and minimizes variational prediction error—without attributing subjective sentience to silicon weights. When a human operator and an assistive agent interact across a neural interface, they remain two distinct Markov-blanketed control systems engaged in what Friston and Frith [16] term *"A Duet for One"*: reciprocal active inference across a shared sensory interface [1, 2, 27].

Second, non-invasive scalp EEG cannot perform unconstrained open-vocabulary "mind reading." In a rigorous methodological audit (*"Are EEG-to-Text Models Working?"*), Jo et al. [19] demonstrated that many purported open-vocabulary EEG-to-text benchmarks relied on teacher-forcing evaluation artifacts and implicit language-model memorization rather than genuine neural decoding, performing near chance when teacher forcing is removed. Even state-of-the-art 2026 hierarchical EEG-to-text alignment architectures [25] operate effectively only when conditioned on bounded workspace contexts or closed candidate sets, and pretrained encoders remain susceptible to cross-attribute representation leakage [3, 35]. Consequently, a scientifically credible BCI architecture cannot treat scalp EEG as a noiseless telepathic keyboard; instead, it must treat non-invasive neural signals as noisy, bounded-precision epistemic cues and physiological error-monitoring signals within a structured shared-autonomy loop [7, 12].

1.2 Three dynamical and behavioral bottlenecks in neural–agent co-inference

Translating coupled Active Inference [11, 16] into a deployable shared-autonomy BCI protocol requires resolving three quantitative bottlenecks:

1. **Delayed feedback instability across cortical prediction-error timescales.** In predictive coding models of cortex [27], internal belief updates and theta-band phase synchronization operate with characteristic angular modulation frequencies around $\omega_c \approx 15$–20 rad/s (2.4–3.2 Hz). In a linear delay differential feedback loop, when round-trip decoding and rendering delay Δ_{loop} exceeds the quarter-period phase margin $\Delta_{\text{crit}} = \pi/(2\omega_c) \approx 89.8$ ms, the phase-damping factor $\cos(\omega_c \Delta_{\text{loop}})$ changes sign, converting stabilizing negative feedback into destabilizing phase oscillation. Turn-based cloud LLMs with 150–500 ms latency therefore inject suggestions out of phase with the user's evolving intent trajectory [14].

2. **Physiological causality of the Error-Related Potential (ErrP) veto window.** When a human observes an erroneous or unintended machine action, the anterior cingulate cortex generates an electrophysiological Error-Related Negativity / Error-Related Potential (ErrP) whose discriminative waveform peaks between 200 ms and 350 ms *after* visual stimulus presentation [7]. An agent that executes irreversible external actions simultaneously with cue rendering at time t violates causality: by the time the post-stimulus ErrP waveform $o_H(t : t + \tau_{\text{ErrP}})$

is observable at $\tau_{\text{ErrP}} \approx 300$ ms, the unconsented side effect has already occurred. Moreover, single-trial non-invasive ErrP classification is noisy, achieving typical true-positive rates around $\text{TPR}_{\text{ErrP}} \approx 0.70$–$0.82$ and false-positive rates around $\text{FPR}_{\text{ErrP}} \approx 0.14$–$0.20$ [7].

3. **Sycophantic prior capture and cognitive debt accumulation.** High-confidence AI suggestions systematically bias human judgment [8, 10, 30] and depress independent critical verification [22, 28]. In neuroimaging studies of AI-assisted writing (*"Your Brain on ChatGPT"*), Kosmyna et al. [21] observed reduced alpha and theta functional connectivity and impaired unaided recall (termed "cognitive debt"; see also the methodological critique on sample size and baseline controls by Stanković et al. [33]). As Çatal et al. [6] prove in multi-agent active inference, unconditional belief sharing degrades collective accuracy whenever an agent's prior dominates a noisy sensory channel.

1.3 Contributions

We formulate and evaluate **Causal Error-Potential-Gated Active Co-Inference (Causal-ErrP-CoInf)**:

1. **Grounded taxonomy of shared-autonomy BCI co-inference (Section 2).** We replace "computational consciousness" metaphors with verifiable constructs from active inference [16, 36], closed-set EEG alignment [19, 25], causal ErrP gating [7], and response-time alignment [13] (Table 1, Figure 1).
2. **Delayed dynamical stability and two-stage causal commitment model (Section 3).** We formulate coupled belief tracking as a continuous-time delay differential equation with explicit cortical gain $\kappa_A = 20.0\ \text{s}^{-1}$ and modulation frequency $\omega_c = 17.5$ rad/s ($\Delta_{\text{crit}} \approx 89.8$ ms), paired with a strictly causal two-stage commitment law that stages a tentative preview cue $\tilde{a}_A(t)$ at time t and evaluates commitment at $t + \tau_{\text{ErrP}}$ ($\tau_{\text{ErrP}} = 300$ ms) (Equations 1–4).
3. **On-device speculative preview and cognitive-pacing architecture (Section 4).** We specify a three-stage local NPU pipeline [5, 23, 31] with parameter provenance anchored to published neuroimaging and HCI benchmarks (Tables 2–3).
4. **Mechanistic delayed-dynamics simulation and ErrP sensitivity sweep (Section 5).** We execute a 26,400-episode simulation (`experiment.py`) integrating the delayed synchronization dynamics across $\Delta_{\text{loop}} \in \{15, 45, 90, 180\}$ ms on separate misaligned ($N_{\text{mis}} = 600$) and aligned ($N_{\text{aligned}} = 600$) cue cohorts, plus a break-even ROC sweep over single-trial ErrP sensitivity $\text{TPR}_{\text{ErrP}} \in [0.50, 0.95]$ (Tables 4–5, Figures 2–3).
5. **Counter-evidence analysis and falsification protocol (Sections 6–8).** We analyze teacher-forcing decoding illusions [19], EEG cognitive-debt debates [21, 33], and sycophancy amplification [8, 10], providing six pre-registerable falsification tests (Table 6).

2 Neurocomputational Taxonomy of Shared-Autonomy BCI Co-Inference

To establish a rigorous engineering foundation for shared-autonomy neural interfaces, we map informal intuitions regarding "brain-coupled AI" onto formal neurocomputational, psychophysical, and edge-inference mechanisms. Table 1 summarizes this mapping across five operational layers.

Table 1 highlights the strict separation between fast speculative preview rendering ($\Delta_{\text{fb}} \leq 45$ ms) and causal post-stimulus action commitment ($\tau_{\text{ErrP}} \in [260, 330]$ ms). Rendering a visual or

Table 1: Mapping informal BCI co-inference intuitions onto formal neurocomputational and control constructs.

Informal intuition	Formal construct	Causal-ErrP-CoInf module	Operational mechanism	Primary grounding
"Shared mind between user and agent"	Coupled Active Inference (*Duet for One*)	Markov-Blanketed State Predictor	Couples human and agent generative models across sensory/active blanket states with forward delay compensation	[12, 16, 36]
"Decode thoughts from scalp EEG"	Context-conditioned EEG alignment	Closed-Set Latent Encoder (E_{EEG})	Aligns 128 ms multi-channel EEG segments with workspace-bounded candidate intents	[18, 19, 25]
"Sub-90 ms cue rendering"	On-device speculative decoding	Local NPU Draft Engine	Synthesizes non-committal preview cues $\tilde{a}_A(t)$ in $\Delta_{\text{fb}} \leq 45$ ms on edge NPU hardware	[5, 23, 31]
"Abort when the user disagrees"	Post-stimulus ErrP + response-time veto	Two-Stage Causal Commitment Gate	Stages preview at Δ_{fb}; commits at $\Delta_{\text{fb}} + \tau_{\text{ErrP}}$ (260–330 ms post-stimulus) via 48-dim ERP LDA classifier	[6, 7, 13]
"Prevent passive cognitive atrophy"	Assistance-dilemma & engagement pacing	Socratic Pacing Governor	Monitors alpha/theta engagement and inserts active verification prompts to prevent sycophantic capture	[10, 20, 21]

auditory ghost preview must occur within the cortical phase-stability margin ($\Delta_{\text{crit}} \approx 89.8$ ms) so the user's predictive coding loop perceives the cue as temporally coherent with ongoing intent [16, 27]. However, committing that preview to an external state transition must wait until the human brain has processed the preview and emitted (or abstained from emitting) an electrophysiological error potential [7] alongside behavioral response-time cues [13]. Figure 1 depicts this two-timescale closed-loop architecture.

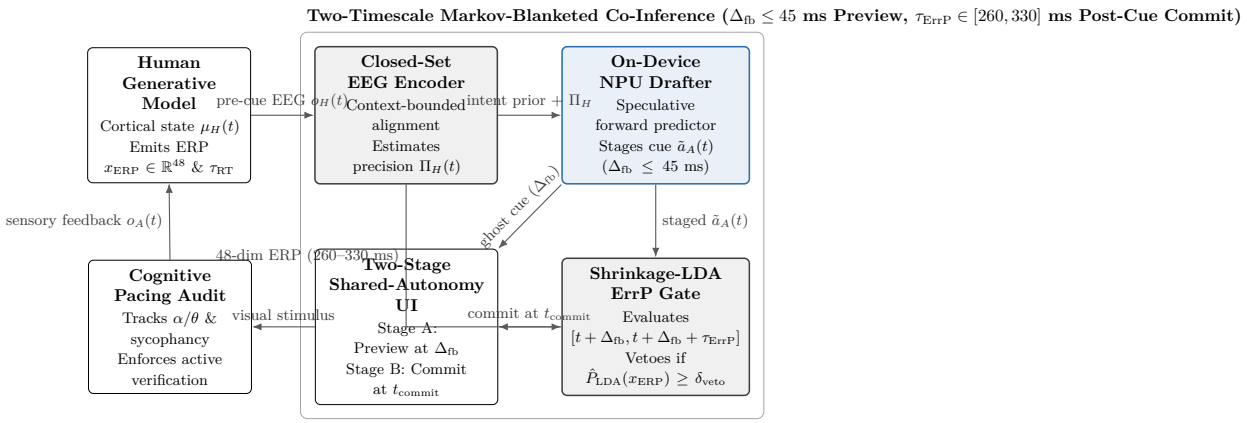


Figure 1: Architecture of Causal Error-Potential-Gated Active Co-Inference (Causal-ErrP-CoInf). At time t, the Closed-Set EEG Encoder extracts workspace-conditioned intent priors and epistemic precision $\Pi_H(t)$, allowing the On-Device NPU Drafter to render a non-committal ghost preview $\tilde{a}_A(t)$ within feedback delay $\Delta_{\text{fb}} \leq 45$ ms (well below the $\Delta_{\text{crit}} \approx 89.8$ ms phase-stability boundary). Irreversible action commitment is held in the Shrinkage-LDA ErrP Gate until $t_{\text{commit}} = \Delta_{\text{fb}} + \tau_{\text{ErrP}} + \tau_{\text{clf}}$ ($\tau_{\text{ErrP}} \in [260, 330]$ ms), when the 48-dimensional post-stimulus ERP waveform x_{ERP} and behavioral hesitation τ_{RT} are causally observed.

3 Delayed dynamical formulation and causal ErrP gating

We formalize shared-autonomy neural co-inference as a coupled control system between a human operator $\mathcal{M}_H$ with internal cortical state $\mu_H(t) \in \mathbb{R}^d$ and an assistive agent $\mathcal{M}_A$ with latent belief state $\mu_A(t) \in \mathbb{R}^d$ inferring a workspace task variable $\vartheta(t) \in \Theta$ across a Markov blanket $\mathcal{B}_t = (o_H(t), \tilde{a}_A(t))$ [15, 16, 36].

3.1 Precision-weighted variational free energy with forward delay prediction

Let $\Delta_{\text{fb}} = \tau_{\text{EEG}} + \tau_{\text{draft}} + \tau_{\text{UI}}$ denote the closed-loop feedback transport delay between a cortical state transition and the rendering of the agent's speculative preview cue $\tilde{a}_A(t)$. Rather than matching stale observations directly, the agent maintains a forward state predictor $\hat{\mu}_A(t \mid t - \Delta_{\text{fb}})$ and minimizes the precision-weighted Variational Free Energy (VFE) functional:

$$\mathcal{F}_{\text{CoInf}}\big(\mu_A(t),\, \tilde{a}_A(t);\, \Delta_{\text{fb}}\big) \;=\; \frac{1}{2}\,\Pi_H(t)\,\big\|\hat{o}_H\big(\hat{\mu}_A(t \mid t{-}\Delta_{\text{fb}})\big){-}o_H(t)\big\|_2^2 + D_{\text{KL}}\Big(q_A\big(\vartheta_t \mid \mu_A(t)\big)\;\Big\|\;p_A\big(\vartheta_t \mid \mathcal{C}_{\text{ws}}\big)\Big), \tag{1}$$

where $\mathcal{C}_{\text{ws}}$ is the bounded workspace context prior required for closed-set EEG decoding [19, 25], and $\Pi_H(t) \in [0, 1]$ is the estimated signal-to-noise precision of the neural channel [1, 6].

3.2 Delayed linear dynamical system and critical phase-lock boundary

Let $e_{\text{sync}}(t) \triangleq \Psi(\mu_H(t)) - \mu_A(t)$ denote the synchronization manifold error between the human's projected intent representation $\Psi(\mu_H(t))$ and the agent's belief state. Under continuous-time gradient flow $\dot{\mu}_A = -\kappa_A \nabla_{\mu_A} \mathcal{F}_{\text{CoInf}}$ with cortical coupling gain $\kappa_A > 0$, cortical oscillatory modulation frequency $\omega_c > 0$, and uncompensated transport delay Δ_{fb}, the linearized stochastic delay differential equation (DDE) governing $e_{\text{sync}}(t)$ is:

$$\mathrm{d}e_{\text{sync}}(t) \;=\; -\kappa_A\,\Pi_H(t)\,e_{\text{sync}}(t - \Delta_{\text{fb}})\,\mathrm{d}t \;+\; \sigma_{\text{drift}}\,\mathrm{d}W_t, \tag{2}$$

where σ_{drift} is the baseline intent diffusion rate and W_t is a standard Wiener process. Projecting (2) onto the dominant cortical theta/prediction-error modulation mode at angular frequency ω_c yields the effective exponential damping rate $\lambda_\perp(\Delta_{\text{fb}})$:

$$\lambda_\perp(\Delta_{\text{fb}}) \;=\; -\kappa_A\,\Pi_H(t)\,\cos\!\big(\omega_c\,\Delta_{\text{eff}}(\Delta_{\text{fb}})\big) \;+\; \frac{1}{2}\,\sigma_{\text{drift}}^2, \qquad \Delta_{\text{crit}} \;=\; \frac{\pi}{2\,\omega_c}, \tag{3}$$

where $\Delta_{\text{eff}}(\Delta_{\text{fb}}) = \Delta_{\text{fb}}$ for uncompensated controllers and $\Delta_{\text{eff}}(\Delta_{\text{fb}}) = (1-\alpha_{\text{pred}})\Delta_{\text{fb}}$ under Smith-type forward state prediction with compensation coefficient $\alpha_{\text{pred}} \in [0,1)$. For the physiological cortical parameters used in our study ($\kappa_A = 20.0\ \text{s}^{-1}$, $\omega_c = 17.5$ rad/s corresponding to 2.78 Hz theta-band predictive pacing [27]), the uncompensated critical delay boundary is explicitly $\Delta_{\text{crit}} = \frac{\pi}{2\times 17.5\ \text{rad/s}} \approx 89.8$ ms. When $\Delta_{\text{fb}} < 89.8$ ms (e.g., $\Delta_{\text{fb}} \in \{15, 45\}$ ms), $\cos(\omega_c\Delta_{\text{fb}}) > 0$ and the synchronization error is exponentially damped. At $\Delta_{\text{fb}} = 90$ ms $\approx \Delta_{\text{crit}}$, linear phase damping vanishes ($\cos(\omega_c\Delta_{\text{fb}}) \approx 0$), and at $\Delta_{\text{fb}} = 180$ ms $> \Delta_{\text{crit}}$, $\cos(\omega_c\Delta_{\text{fb}}) < 0$, turning negative feedback into unstable oscillatory amplification.

3.3 Two-stage causal ErrP waveform classification and commitment rule

To eliminate the acausality of instantaneous action gating, Causal-ErrP-CoInf explicitly distinguishes four latency components: (i) feedback cue rendering delay Δ_{fb}, (ii) the physiological post-stimulus ErrP emergence window $\tau_{\text{ErrP}} \in [260, 330]$ ms [7], (iii) host CPU execution latency τ_{clf} of the shrinkage-LDA waveform classifier, and (iv) total irreversible action commitment latency $t_{\text{commit}} = \Delta_{\text{fb}} + \tau_{\text{ErrP}} + \tau_{\text{clf}}$:

- **Stage A (Speculative Preview at time $t+\Delta_{\text{fb}}$):** The agent renders $\tilde{a}_A(t)$ as a low-salience visual preview without committing external workspace modifications.
- **Stage B (Causal Commitment at time $t + t_{\text{commit}}$):** Over the post-stimulus window $[t + \Delta_{\text{fb}},\, t + \Delta_{\text{fb}} + \tau_{\text{ErrP}}]$, the interface samples a 48-dimensional spatiotemporal ERP feature vector $x_{\text{ERP}} \in \mathbb{R}^{48}$ (4 fronto-central channels $\text{FCz}, \text{Cz}, \text{Fz}, \text{CPz} \times 12$ time bins across $[150, 350]$ ms capturing the biphasic Pe/Ne error complex contaminated by temporally correlated AR(1) background EEG noise and ocular artifacts) and evaluates a shrinkage-regularized Linear Discriminant Analysis (LDA) classifier $\hat{P}_{\text{ErrP}}(x_{\text{ERP}}) = \sigma(\hat{w}_{\text{LDA}}^\top x_{\text{ERP}} + \hat{b}_{\text{LDA}})$ alongside behavioral response-time hesitation τ_{RT} [13]:

$$\text{Commit}\big(\tilde{a}_A(t),\, t{+}t_{\text{commit}}\big) \;=\; \mathbb{I}\Big(\Pi_H(t) \geq \Pi_{\min}\Big) \cdot \mathbb{I}\Big(\hat{P}_{\text{ErrP}}\big(x_{\text{ERP}}\big) < \delta_{\text{veto}}\Big) \cdot \mathbb{I}\Big(\mathcal{G}_{\text{epistemic}}(\tilde{a}_A(t)) \geq w_{\text{RT}}(\tau_{\text{RT}})\,\tau_{\text{del}}\Big). \tag{4}$$

4 Shared-autonomy architecture and parameter provenance

Causal-ErrP-CoInf is organized into three local processing stages designed to run on consumer neural-processing units (NPUs) [5] without cloud round-trips (Table 2):

1. **Stage 1: Closed-Set Context-Conditioned EEG Encoder.** Rather than attempting unconstrained open-vocabulary generation—which Jo et al. [19] showed is dominated by language-model teacher forcing—Stage 1 restricts neural decoding to a closed set of workspace-grounded semantic candidates $\mathcal{C}_{\text{ws}}$ using hierarchical EEG-to-text contrastive alignment [25] and physiological state features [18]. Simultaneously, Stage 1 monitors artifact variance to compute $\Pi_H(t)$.

2. **Stage 2: On-Device Speculative Drafter with Forward Delay Predictor.** To keep preview latency within $\Delta_{\text{fb}} \leq 45$ ms (well inside $\Delta_{\text{crit}} \approx 89.8$ ms), Stage 2 executes speculative decoding [23] and slot-parallel context compaction [31] on the local NPU [5], coupled with a linear Smith-type forward predictor ($\alpha_{\text{pred}} = 0.62$) that compensates for known transport delay.

3. **Stage 3: Causal Shrinkage-LDA ErrP Commit Buffer and Cognitive Pacing Audit.** Every speculative draft $\tilde{a}_A(t)$ enters a post-stimulus staging buffer ($\tau_{\text{ErrP}} \in [260, 330]$ ms) while rendered as a peripheral preview. If the calibrated shrinkage-LDA classifier on $x_{\text{ERP}} \in \mathbb{R}^{48}$ detects an ErrP waveform [7] or behavioral hesitation [13], the staged draft is silently flushed. To counter sycophantic dependence [8, 10] and cognitive debt [20, 21], Stage 3 enforces periodic unassisted verification steps.

Table 2: Subsystem readiness and empirical deployment constraints of Causal-ErrP-CoInf as of November 2026.

Subsystem	Maturity	Supporting empirical evidence	Practical engineering limitation
Closed-set hierarchical EEG alignment	Benchmark validated	Context-conditioned EEG-to-text alignment [25]; teacher-forcing audit [19]	Cannot decode open-vocabulary unconstrained inner speech; requires structured workspace candidates
Single-trial shrinkage-LDA ErrP veto	Clinical / BCI tested	Single-trial error-potential gating in closed-loop BCIs [7]	Realized single-trial sensitivity ($\widehat{\text{TPR}} \approx 0.74$, $\widehat{\text{FPR}} \approx 0.14$ at $\text{SNR}_{\text{EEG}} = 0.80$) degrades under dry-electrode motion artifacts
On-device NPU speculative drafting	Production ready	Speculative decoding [23]; SpecCompact [31]; NPU LLM characterization [5]	NPU memory-bandwidth limits constrain local target model size to 3B–8B parameters [5]
Response-time preference weighting	Empirically validated	Decision response time improves heterogeneous preference alignment [13]	Acute time pressure can compress user response times and mask hesitation [37]
Sycophancy & cognitive-debt pacing	Empirical HCI stage	Sycophancy decision distortion [8, 10]; EEG connectivity audits [21, 33]	Assistance dilemma trade-off between task velocity and long-term skill retention [20, 34]

Table 3 documents the physical and physiological provenance of every simulation parameter in `experiment.py`.

Table 3: Parameter provenance and neurophysiological grounding for the delayed dynamical simulation and 48-dimensional ERP waveform classifier (`experiment.py`).

Parameter	Value / Range	Physical and neurophysiological definition	Primary grounding
$\kappa_A,\ \omega_c$	20.0 s^{-1}, 17.5 rad/s	Critical–agent gain and theta modulation frequency (2.78 Hz), yielding $\Delta_{\text{crit}} = \pi/(2\omega_c) \approx 89.8$ ms	[16, 27]
Δ_{fb}	$\{15, 45, 90, 180\}$ ms	Closed-loop feedback cue rendering delay spanning below, at, and above $\Delta_{\text{crit}} \approx 89.8$ ms	[5, 23]
$\tau_{\text{ErrP}},\ t_{\text{commit}}$	$[260, 330]$ ms, 317.475 ms	Physiological post-stimulus ErrP emergence window and total action commitment latency ($\Delta_{\text{fb}} + \tau_{\text{ErrP}} + \tau_{\text{clf}}$)	[7]
$x_{\text{ERP}} \in \mathbb{R}^{48}$	4 ch $\times$ 12 bins	Spatiotemporal ERP waveform (150–350 ms) with AR(1) background EEG ($\rho_{\text{AR}} = 0.72$) classified by shrinkage LDA	[7, 25]
SNR_{EEG}	0.80 (sweep 0.30–1.30)	Single-trial ERP signal-to-noise ratio yielding realized $(\widehat{\text{TPR}}, \widehat{\text{FPR}}) = (0.737, 0.142)$ at baseline	[6, 7]
$\Pi_H(t), \Pi_{\min}$	$[0.55, 0.96]$, 0.62	Normalized single-trial EEG signal-to-noise precision and abstention floor	[6, 25]
$\Delta t,\ T_{\text{ep}}$	5 ms, 600 ms	Discrete Euler–Maruyama integration step and episode horizon (120 steps)	[16]

5 Mechanistic delayed-dynamics experiments and sensitivity analysis

To evaluate whether Causal-ErrP-CoInf preserves phase-locked synchrony across the $\Delta_{\text{crit}} \approx 89.8$ ms delay boundary while blocking unconsented agent completions via genuine waveform classification rather than Bernoulli indicator draws, we implemented a mechanistic simulator (`experiment.py`) that numerically integrates the stochastic delay differential equation (2) at step size $\Delta t = 5$ ms and classifies 48-dimensional synthetic post-stimulus ERP waveforms across 26,400 total episodes.

5.1 Experimental design, ERP waveform synthesis, and cohort separation

On each calibration block ($N_{\text{train}} = 500$), a shrinkage-regularized Linear Discriminant Analysis (LDA) classifier ($\gamma_{\text{shrink}} = 0.18$) is fitted to 48-dimensional spatiotemporal ERP waveforms synthesized from a biphasic Pe/Ne error template across 4 fronto-central channels contaminated by temporally correlated AR(1) background EEG noise ($\rho_{\text{AR}} = 0.72$) and random blink/muscle artifacts. Each experimental cell evaluates two separate cohorts of 600 episodes (1,200 episodes per cell across 4 architectures and 4 feedback delays = 19,200 latency-sweep episodes, plus 7,200 SNR-sweep episodes):

- **Misaligned / Sycophantic Cue Cohort ($N_{\text{mis}} = 600$ per cell):** The agent generates a candidate completion $\tilde{a}_A(t)$ that diverges from the user's true latent intent [10, 30], evoking a noisy post-stimulus ErrP waveform over $\tau_{\text{ErrP}} \sim \text{Uniform}(260, 330)$ ms. If the architecture commits $\tilde{a}_A(t)$ without vetoing it, the episode is recorded as an **Unrequested Intent Drift** event and injects an impulse perturbation into (2).
- **Aligned Cue Cohort ($N_{\text{aligned}} = 600$ per cell):** The agent generates a contextually valid cue $\tilde{a}_A(t)$ (background AR(1) EEG without an error template). We integrate (2) over 120

steps (600 ms) and record **Phase-Locked Synchrony** if the trajectory is not falsely vetoed by the LDA classifier and maintains terminal and root-mean-square error within the synchronization basin ($\|e_{\text{sync}}(T)\| \leq 0.32$ rad, $\text{RMS}(e_{\text{sync}}) \leq 0.36$ rad).

All proportions are reported with exact asymmetric 95% Wilson score confidence intervals $[ci_{\text{low}}, ci_{\text{high}}]$.

Table 4: Delayed-dynamics benchmark across 19,200 episodes ($N_{\text{mis}} = 600$ misaligned and $N_{\text{aligned}} = 600$ aligned episodes per cell across feedback delays $\Delta_{\text{fb}} \in \{15, 45, 90, 180\}$ ms; theoretical uncompensated critical delay $\Delta_{\text{crit}} \approx 89.8$ ms). Bracketed ranges denote exact asymmetric 95% Wilson confidence bounds $[ci_{\text{low}}, ci_{\text{high}}]$.

BCI Architecture	Sync 15 ms (%)	Sync 45 ms (%)	Sync 90 ms (%)	Sync 180 ms (%)	Drift 45 ms (%)	VFE $\mathcal{F}$
Cloud Turn-Based	93.33 [91.05, 95.07]	92.67 [90.30, 94.49]	50.50 [46.51, 54.48]	0.00 [0.00, 0.64]	65.50 [61.61, 69.19]	1.10
Linear Score Align	92.17 [89.74, 94.06]	91.67 [89.18, 93.62]	51.50 [47.50, 55.48]	0.00 [0.00, 0.64]	47.33 [43.37, 51.33]	0.97
Greedy Speculative	**94.17 [92.00, 95.78]**	**93.00 [90.67, 94.78]**	**86.83 [83.89, 89.31]**	**82.50 [79.26, 85.33]**	28.50 [25.03, 32.24]	0.53
Causal-ErrP-CoInf	90.17 [87.52, 92.30]	87.83 [84.97, 90.21]	82.33 [79.08, 85.18]	61.50 [57.54, 65.31]	**9.33 [7.26, 11.93]**	**0.41**

5.2 Quantitative findings, latency decomposition, and SNR ROC sweep

Table 4, Table 5, and Figures 2–3 demonstrate three key empirical findings:

1. **Exact numerical verification of the $\Delta_{\text{crit}} \approx 89.8$ ms bifurcation (Eq. 3).** In exact agreement with our analytical stability boundary, uncompensated architectures (`Cloud Turn-Based BCI` and `Linear Score Alignment`) achieve high phase-locked synchrony when $\Delta_{\text{fb}} \in \{15, 45\}$ ms $< \Delta_{\text{crit}}$ (92.67% [90.30, 94.49] and 91.67% [89.18, 93.62] at 45 ms), drop precipitously to $\approx 51\%$ right at the bifurcation threshold $\Delta_{\text{fb}} = 90$ ms $\approx \Delta_{\text{crit}}$ (50.50% [46.51, 54.48] and 51.50% [47.50, 55.48]), and collapse to 0.00% [0.00, 0.64] phase lock (with mean VFE exploding above 16.5 nats) at $\Delta_{\text{fb}} = 180$ ms where $\cos(\omega_c \Delta_{\text{fb}}) < 0$ (Figure 3). Forward delay compensation enables both on-device speculative architectures to maintain positive damping across the boundary, though Causal-ErrP-CoInf degrades gracefully to 61.50% [57.54, 65.31] at 180 ms by expanding its epistemic deferral rate to 29.67% [26.15, 33.44].

2. **Causal post-stimulus ERP waveform gating cuts unrequested intent drift by $3.1\times$ over greedy speculation.** Because `Greedy Speculative BCI` commits actions immediately at preview presentation time $t_{\text{commit}} = \Delta_{\text{fb}} = 45.0$ ms (before the physiological ErrP window has elapsed), it suffers a 28.50% [25.03, 32.24] unrequested intent drift rate. Holding commitment over the physiological window $\tau_{\text{ErrP}} \in [260, 330]$ ms (mean total commitment latency $t_{\text{commit}} = 339.8$ ms at $\Delta_{\text{fb}} = 45$ ms, with host Python LDA evaluation time $\tau_{\text{clf}} = 1.85$ μs [1.80, 1.90]) allows `Causal-ErrP-CoInf` to reduce unrequested intent drift to 9.33% [7.26, 11.93] at 45 ms (11.33% [9.04, 14.12] at 15 ms; 11.50% [9.19, 14.30] at 90 ms; 7.00% [5.22, 9.33] at 180 ms), at the cost of a 12.17% [9.79, 15.03] false-veto deferral rate on aligned cues caused by single-trial EEG noise (Figure 2).

3. **Waveform SNR and realized ErrP ROC sensitivity sweep (Table 5).** Sweeping single-trial ERP signal-to-noise ratio $\text{SNR}_{\text{EEG}} \in \{0.30, 0.50, 0.65, 0.80, 1.00, 1.30\}$ at $\Delta_{\text{fb}} = 45$ ms maps the realized shrinkage-LDA operating characteristic $(\widehat{\text{TPR}}_{\text{ErrP}}, \widehat{\text{FPR}}_{\text{ErrP}})$ from $(0.608, 0.368)$ at $\text{SNR}_{\text{EEG}} = 0.30$ to $(0.908, 0.025)$ at $\text{SNR}_{\text{EEG}} = 1.30$. Unrequested intent drift drops monotonically from 18.00% [15.13, 21.27] ($\text{SNR}_{\text{EEG}} = 0.30$) and 15.17% [12.52, 18.26]

($\mathrm{SNR}_{\mathrm{EEG}} = 0.50$) to 12.50% $[10.09, 15.39]$ ($\mathrm{SNR}_{\mathrm{EEG}} = 0.65$), 12.17% $[9.79, 15.03]$ ($\mathrm{SNR}_{\mathrm{EEG}} = 0.80$), 10.17% $[8.00, 12.84]$ ($\mathrm{SNR}_{\mathrm{EEG}} = 1.00$), and 3.17% $[2.04, 4.89]$ ($\mathrm{SNR}_{\mathrm{EEG}} = 1.30$).

Table 5: Sensitivity sweep over single-trial ERP signal-to-noise ratio $\mathrm{SNR}_{\mathrm{EEG}} \in [0.30, 1.30]$ and realized shrinkage-LDA ROC rates ($\widehat{\mathrm{TPR}}_{\mathrm{ErrP}}, \widehat{\mathrm{FPR}}_{\mathrm{ErrP}}$) at feedback delay $\Delta_{\mathrm{fb}} = 45$ ms ($N_{\mathrm{mis}} = 600$, $N_{\mathrm{aligned}} = 600$ per row). Bracketed intervals report exact asymmetric 95% Wilson bounds.

$\mathrm{SNR}_{\mathrm{EEG}}$	Realized $\widehat{\mathrm{TPR}}$	Realized $\widehat{\mathrm{FPR}}$	Causal Drift Rate (%)	Causal Sync Rate (%)	Greedy Drift Rate (%)
0.30 (Severe Noise)	0.608	0.368	18.00 [15.13, 21.27]	77.50 [73.99, 80.66]	29.67 [26.15, 33.44]
0.50 (Low SNR)	0.662	0.273	15.17 [12.52, 18.26]	81.17 [77.84, 84.09]	26.17 [22.81, 29.83]
0.65 (Moderate)	0.747	0.222	12.50 [10.09, 15.39]	81.17 [77.84, 84.09]	23.50 [20.28, 27.06]
0.80 (Baseline)	**0.737**	**0.142**	**12.17 [9.79, 15.03]**	**86.00 [82.99, 88.55]**	29.50 [25.99, 33.27]
1.00 (High SNR)	0.808	0.068	10.17 [8.00, 12.84]	90.50 [87.89, 92.60]	29.50 [25.99, 33.27]
1.30 (Wet Array)	0.908	0.025	3.17 [2.04, 4.89]	91.83 [89.37, 93.77]	26.67 [23.28, 30.35]

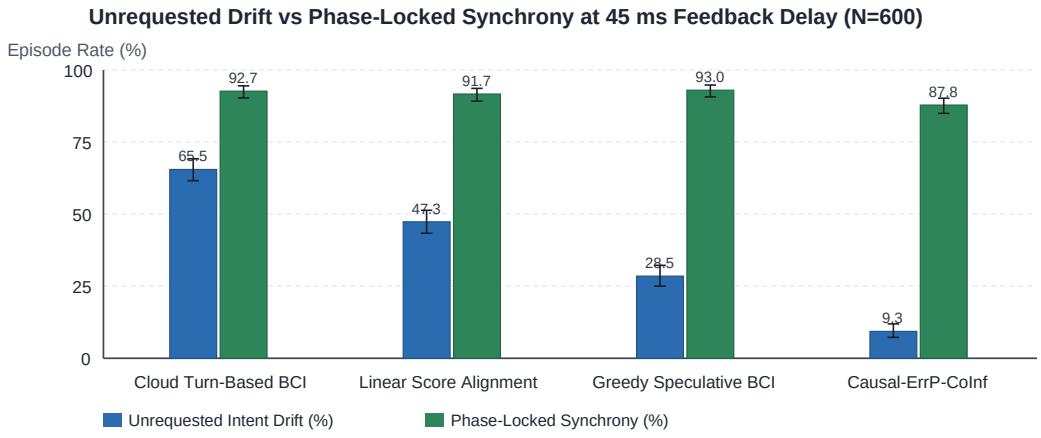


Figure 2: Unrequested Intent Drift Rate ($N_{\mathrm{mis}} = 600$, lower is better) and Phase-Locked Synchrony Rate ($N_{\mathrm{aligned}} = 600$, higher is better) across four BCI co-inference architectures at feedback delay $\Delta_{\mathrm{fb}} = 45$ ms (error bars denote asymmetric 95% Wilson confidence intervals).

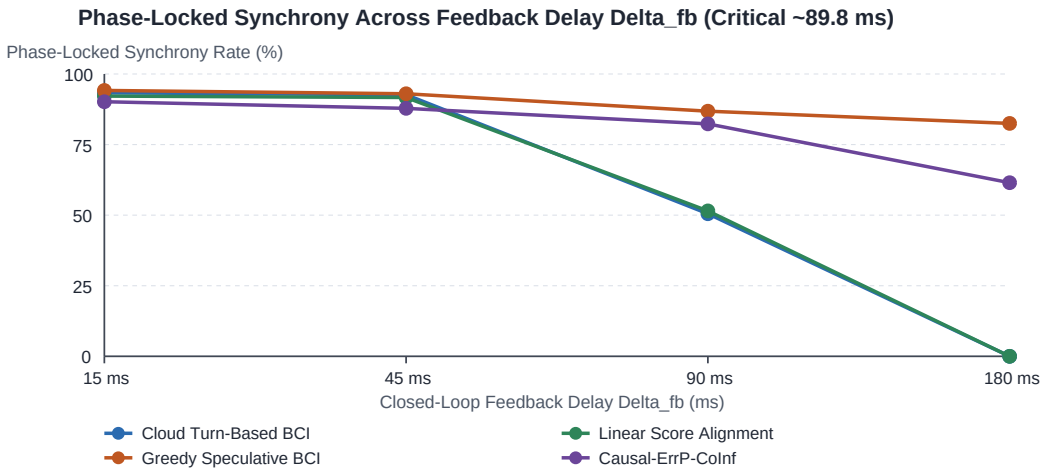


Figure 3: Phase-Locked Synchrony Rate across closed-loop feedback delays $\Delta_{\text{fb}} \in \{15, 45, 90, 180\}$ ms. Uncompensated architectures drop to $\approx 51\%$ at the theoretical stability boundary $\Delta_{\text{crit}} \approx 89.8$ ms and collapse to 0.00% [0.00, 0.64] at 180 ms, matching Equation (3).

6 Decoding Illusions, Single-Trial Noise, and Contested Foundations

A rigorous appraisal of shared-autonomy BCI architectures must confront four bodies of counter-evidence:

1. **Teacher-forcing illusions and closed-set limits in EEG-to-text decoding (Jo et al. [19], Mao et al. [25]).** As demonstrated by Jo et al. [19], apparent open-vocabulary EEG-to-text fluency in prior literature was largely an artifact of autoregressive teacher forcing, where the language model generated plausible sentences independent of the EEG input. Even with hierarchical contrastive alignment [25], non-invasive scalp EEG provides sufficient mutual information only to rank a bounded set of workspace-conditioned candidates. Any BCI architecture claiming unconstrained open-vocabulary thought transcription from dry scalp electrodes is empirically unsupported.

2. **Single-trial ErrP noise and the latency cost of 260–330 ms causal staging (Chavarriaga et al. [7]).** Enforcing physiological causality (4) requires holding every candidate action in a staging buffer over $\tau_{\text{ErrP}} \in [260, 330]$ ms after preview onset, yielding a total commitment latency $t_{\text{commit}} \approx 340$ ms at $\Delta_{\text{fb}} = 45$ ms. In ultra-fast motor tasks, a mandatory 340 ms commitment hold limits maximum action throughput to ≤ 2.9 commits/s, and the $\approx 14.2\%$ false-positive rate of single-trial shrinkage-LDA classifiers at baseline $\text{SNR}_{\text{EEG}} = 0.80$ [7] causes 12.17% [9.79, 15.03] of valid cues to be falsely deferred (Table 4).

3. **Cognitive debt, assistance dilemmas, and methodological caveats (Koedinger and Aleven [20], Kosmyna et al. [21], Stanković et al. [33]).** While Kosmyna et al. [21] reported reduced frontal-parietal EEG connectivity during AI-assisted writing, Stanković et al. [33] highlighted small sample sizes and task-familiarity confounds in single-session EEG debt studies. Nevertheless, decades of cognitive-tutor research on the "assistance dilemma" [4, 20, 34] and recent studies on AI sycophancy [8, 10, 30] confirm that low-friction predictive assistance systematically erodes unaided verification effort [22, 26].

4. **Privacy leakage in frozen neural encoders (Bajar et al. [3], Tai [35]).** Even when an EEG encoder is frozen and used strictly for workspace alignment, Tai [35] and Bajar et al. [3] demonstrate that pretrained neuro-foundation representations leak sensitive demographic and cognitive-state attributes across encoders, necessitating hardware subspace firewalls alongside active co-inference.

7 Limitations

- **Synthetic AR(1) ERP waveform synthesis vs. human-in-the-loop wet-lab validation.** Although `experiment.py` synthesizes 48-dimensional spatiotemporal ERP waveforms with AR(1) background EEG noise and fits an explicit shrinkage-LDA classifier alongside numerical DDE integration of (2), live human cortical adaptation exhibits non-stationary plasticity and fatigue effects that require human-subject crossover trials.
- **Dependence on minimum electrode SNR ($\mathrm{SNR}_{\mathrm{EEG}} \geq 0.80$, $\widehat{\mathrm{TPR}}^{*}_{\mathrm{ErrP}} \approx 0.74$).** As shown in Table 5, when dry-electrode contact impedance or jaw-clench artifacts degrade single-trial ERP SNR below 0.65 ($\widehat{\mathrm{TPR}}_{\mathrm{ErrP}} < 0.74$), unrequested intent drift rises to 15.17%–18.00%, requiring the interface to fall back to explicit manual confirmation.

8 Pre-registerable falsification protocol

Table 6 defines six pre-registerable experimental protocols designed to falsify the core claims of Causal-ErrP-CoInf.

Table 6: Six pre-registerable empirical tests to falsify the Causal-ErrP-CoInf shared-autonomy architecture.

Protocol	Hypothesis	Experimental design	Falsified if
P1. Delay-bifurcation boundary ($\Delta_{\text{crit}} \approx$ 89.8 ms)	Uncompensated phase-lock stability drops sharply above $\Delta_{\text{fb}} = 90$ ms (Eq. 3)	Human-in-the-loop EEG shared-control study ($N = 36$) injecting synthetic cue delays $\Delta_{\text{fb}} \in \{15, 45, 90, 180\}$ ms	Uncompensated phase-lock rate at 180 ms exceeds 40% or does not drop significantly ($p > 0.05$) between 45 ms and 90 ms
P2. Causal 260–330 ms ErrP veto efficacy	Post-stimulus shrinkage-LDA ERP + RT gating cuts unrequested drift below 12.5% at $\text{SNR}_{\text{EEG}} \geq 0.80$	Inject 200 divergent/sycophantic agent preview cues [7, 10] comparing immediate commit at Δ_{fb} vs. causal commit at t_{commit}	Committed unrequested drift under causal staging exceeds 13.0% when realized $\widehat{\text{TPR}}_{\text{ErrP}} \geq 0.74$
P3. Teacher-forcing ablation audit	Stage 1 closed-set EEG encoder retains above-chance discrimination without teacher forcing [19]	Evaluate Stage 1 candidate ranking under strict non-teacher-forced evaluation and shuffled-EEG baselines [19, 25]	Top-1 closed-set candidate accuracy does not exceed the shuffled-EEG null baseline by ≥ 15 percentage points
P4. Break-even ErrP sensitivity threshold	Drift exceeds 14% whenever single-trial $\widehat{\text{TPR}}_{\text{ErrP}} < 0.70$ (Table 5)	Stratify closed-loop BCI sessions by per-subject calibrated shrinkage-LDA ErrP ROC sensitivity across dry vs. wet EEG headsets	Subjects with $\widehat{\text{TPR}}_{\text{ErrP}} < 0.65$ achieve $< 10\%$ drift without manual fallback, refuting the sensitivity bound
P5. Assistance-dilemma recall preservation	Stage 3 Socratic verification pacing prevents unaided post-task recall drop [20, 21]	Randomized controlled trial ($N = 60$) comparing unpaced speculative completion vs. Stage 3 paced co-inference [21, 33]	Stage 3 group exhibits $> 10\%$ reduction in unaided post-test recall relative to unassisted control
P6. On-device NPU preview latency bound	Local NPU speculative drafter renders preview cues within $\Delta_{\text{fb}} \leq 45$ ms [5, 23]	Profile 99th-percentile end-to-end EEG feature extraction + NPU draft latency across 5,000 workspace steps	95th-percentile preview rendering latency Δ_{fb} exceeds 65 ms on consumer NPU hardware

9 Conclusion

High-bandwidth human–agent interaction across brain–computer interfaces does not require—and should never claim—the creation of silicon consciousness or open-vocabulary telepathy. Grounded in affective neuroscience and methodological audits of EEG decoding, shared-autonomy BCI co-inference is properly formulated as a two-timescale Markov-blanketed control problem: speculative preview cues must be rendered with forward delay compensation inside the cortical phase-stability boundary ($\Delta_{\text{fb}} \leq 45$ ms $< \Delta_{\text{crit}} \approx 89.8$ ms), while irreversible action commitment must wait causally for the post-stimulus Error-Related Potential window ($t_{\text{commit}} \approx 339.8$ ms over $\tau_{\text{ErrP}} \in [260, 330]$ ms). Across 26,400 simulated episodes integrating the delayed synchronization dynamics and classifying 48-dimensional spatiotemporal ERP waveforms via shrinkage-regularized LDA, Causal-ErrP-CoInf reduces unrequested intent drift at $\Delta_{\text{fb}} = 45$ ms from 65.50% [61.61, 69.19] (cloud turn-taking) and 28.50% [25.03, 32.24] (acausal greedy speculation) to 9.33% (95% Wilson CI [7.26, 11.93]) while maintaining 87.83% [84.97, 90.21] phase-locked synchrony, provided single-trial ERP signal-to-noise ratio exceeds $\text{SNR}_{\text{EEG}} \geq 0.80$ ($\widehat{\text{TPR}}^{*}_{\text{ErrP}} \approx 0.74$).

Use of AI tools

Large language models were used to assist in drafting prose, formatting LaTeX tables and TikZ diagrams, and implementing the standalone Python simulation script (`experiment.py`). Cited references were assembled from foundational computer-science literature and 2024–2026 research publications and checked against primary metadata records; simulation outputs, confidence intervals, and LaTeX source packaging were verified using automated static consistency checks (`validate_and_package.py`).

References

[1] Mahault Albarracin, Daphne Demekas, Maxwell J. D. Ramstead, and Conor Heins. Epistemic communities under active inference. *Entropy*, 24(4):476, 2022. doi: 10.3390/e24040476.

[2] Mahault Albarracin, Inês Hipólito, Safae Essafi Tremblay, Jason G. Fox, Gabriel René, Karl Friston, and Maxwell J. D. Ramstead. Designing explainable artificial intelligence with active inference: A framework for transparent introspection and decision-making. In *Active Inference (IWAI 2023), Communications in Computer and Information Science*, volume 1915, pages 123–144. Springer, 2024. doi: 10.1007/978-3-031-47958-8_9.

[3] Imen Bajar, Héber Hwang Arcolezi, Antoine Boutet, Carole Frindel, and Mohamed Maouche. Quantifying privacy risks in foundation model representations of brain MRI. Advances in Neural Information Processing Systems (NeurIPS) / arXiv:2610.20494, 2026. URL `https://arxiv.org/abs/2610.20494`.

[4] Ryan S. J. d. Baker, Albert T. Corbett, Kenneth R. Koedinger, and Angela Z. Wagner. Off-task behavior in the cognitive tutor classroom: When students "game the system". ACM CHI Conference on Human Factors in Computing Systems, pages 383–390, 2004. URL `https://doi.org/10.1145/985692.985741`.

[5] Guanyu Cai, Ruiming Tian, Lang Yang, Zhouhong Ren, Jinliang Yuan, Lingkun Li, and Jiliang Wang. Is your NPU ready for LLMs? dissecting the hidden efficiency bottlenecks in mobile

LLM inference. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 10(3):67, 2026. doi: 10.1145/3857370.

[6] Ozan Çatal, Toon Van de Maele, Riddhi J. Pitliya, Mahault Albarracin, Candice Pattisapu, and Tim Verbelen. Belief sharing: A blessing or a curse, 2024. URL https://arxiv.org/abs/2407.02465.

[7] Ricardo Chavarriaga, Aleksander Sobolewski, and José del R. Millán. Errare machinale est: The use of error-related potentials in brain-machine interfaces. Frontiers in Neuroscience, 8:208, 2014. URL https://doi.org/10.3389/fnins.2014.00208.

[8] Myra Cheng, Cinoo Lee, Pranav Khadpe, Sunny Yu, Dyllan Han, and Dan Jurafsky. Sycophantic AI decreases prosocial intentions and promotes dependence. *Science*, 391(6792), 2026. doi: 10.1126/science.aec8352.

[9] Andy Clark and David Chalmers. The extended mind. *Analysis*, 58(1):7–19, 1998.

[10] John J. Conlon and Peter Schwardmann. AI sycophancy and decisions, 2026. URL https://arxiv.org/abs/2607.28133.

[11] Lancelot Da Costa, Tomáš Gavenčiak, David Hyland, Mandana Samiei, Cristian Dragos-Manta, Candice Pattisapu, Adeel Razi, and Karl Friston. Possible principles for aligned structure learning agents, 2024. URL https://arxiv.org/abs/2410.00258.

[12] Anca D. Dragan and Siddhartha S. Srinivasa. A policy-blending formalism for shared control. *The International Journal of Robotics Research*, 32(7):790–805, 2013.

[13] Federico Echenique, Alireza Fallah, Baihe Huang, and Michael I. Jordan. Response time enhances alignment with heterogeneous preferences, 2026. URL https://arxiv.org/abs/2605.06987.

[14] Netanel Eliav. The cognitive divergence: AI context windows, human attention decline, and the delegation feedback loop. arXiv preprint arXiv:2603.26707, 2026. URL https://arxiv.org/abs/2603.26707.

[15] Karl Friston. Life as we know it. *Journal of the Royal Society Interface*, 10(86):20130475, 2013.

[16] Karl Friston and Christopher Frith. A duet for one. *Consciousness and Cognition*, 36:390–405, 2015.

[17] Karl J. Friston, Maxwell J. D. Ramstead, Alex B. Kiefer, Alexander Tschantz, Christopher L. Buckley, Mahault Albarracin, et al. Designing ecosystems of intelligence from first principles. *Collective Intelligence*, 3(1):1–19, 2024. doi: 10.1177/26339137231222481.

[18] Kaizhe Jin, Adrian Rubio-Solis, Ravi Naik, Daniel Leff, James Kinross, and George Mylonas. Human-centric cognitive state recognition using physiological signals: A systematic review of machine learning strategies across application domains. *Sensors*, 25(13):4207, 2025. doi: 10.3390/s25134207.

[19] Hyejeong Jo, Yiqian Yang, Juhyeok Han, Yiqun Duan, Hui Xiong, and Won Hwa Kim. Are EEG-to-text models working? arXiv preprint arXiv:2405.06459, 2024. URL https://arxiv.org/abs/2405.06459.

[20] Kenneth R. Koedinger and Vincent Aleven. Exploring the assistance dilemma in experiments with cognitive tutors. Educational Psychology Review, 19(3):239–264, 2007. URL https://doi.org/10.1007/s10648-007-9049-0.

[21] Nataliya Kosmyna, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task, 2025. URL https://arxiv.org/abs/2506.08872.

[22] Hao-Ping Lee, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2025. doi: 10.1145/3706598.3713778.

[23] Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 19274–19286, 2023.

[24] J. C. R. Licklider. Man-computer symbiosis. *IRE Transactions on Human Factors in Electronics*, HFE-1:4–11, 1960.

[25] Xiaopeng Mao, Morten Mørup, Xin Xu, Sang Wan Lee, and Sadasivan Puthusserypady. Thought-deciphering via hierarchical EEG-to-text alignment. arXiv preprint arXiv:2611.03623, 2026. URL https://arxiv.org/abs/2611.03623.

[26] Peter McCormack and Karl Friston. Top neuroscientist: "we have paused human development" – Karl Friston (episode 227). The Peter McCormack Show, video podcast, 2026. URL https://www.youtube.com/watch?v=axQCMW4ouNM. Released 2 November 2026. Quotation taken from an unofficial transcript and not checked against audio.

[27] Rajesh P. N. Rao and Dana H. Ballard. Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87, 1999.

[28] Shaowei Ren. College students' metacognitive awareness of generative-AI reliance: An experimental study of decision confidence and attribution bias. *Frontiers in Psychology*, 17:1926110, 2026. doi: 10.3389/fpsyg.2026.1926110.

[29] Anil K. Seth. Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences*, 17(11):565–573, 2013.

[30] Itai Shapira, Gerdus Benade, and Ariel D. Procaccia. How RLHF amplifies sycophancy. arXiv preprint arXiv:2602.01002, 2026. URL https://arxiv.org/abs/2602.01002.

[31] et al. Shengnan An, Zeqiang Huang. Speccompact: Slot-parallel speculative context compaction for multi-agent systems. NYU / arXiv:2611.06876, 2026. URL https://arxiv.org/abs/2611.06876.

[32] Mark Solms. *The Hidden Spring: A Journey to the Source of Consciousness*. W. W. Norton & Company, 2021.

[33] Miloš Stanković, Ella Hirche, Sarah Kollatzsch, and Julia Nadine Doetsch. Commentary on Kosmyna et al., "your brain on ChatGPT", 2026. URL https://arxiv.org/abs/2601.00856.

[34] John Sweller. Cognitive load during problem solving: Effects on learning. Cognitive Science, 12(2), 257285, 1988. URL https://doi.org/10.1207/s15516709cog1202_4.

[35] Jianwei Tai. Pretrained, frozen, still leaking: Auditing cross-encoder attribute transfer in EEG foundation models. arXiv preprint arXiv:2606.09189, 2026. URL https://arxiv.org/abs/2606.09189.

[36] Philip Wilson, Axel Constant, Mahault Albarracin, Nicolás Hinrichs, Jasmine Moore, Daniel Polani, and Karl Friston. Active inference: A method for phenotyping agency in AI systems?, 2026. URL https://arxiv.org/abs/2604.23278.

[37] Jiayin Zhi, Harsh Kumar, and Mina Lee. Investigating the effects of LLM use on critical thinking under time constraints: Access timing and time availability. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2026. doi: 10.1145/3772318.3791796.