NSS
Neutrosophic Sets and Systems, Vol. 102, 2027
University of New Mexico

N-CogHCI: A Reliability-Aware Neutrosophic Cognitive Control Framework for Cognitive AI-Enhanced Human–Computer Interaction

Nada Mohamed[1] and Alshaimaa A. Tantawy[2,*]

[1] Nada Mohamed Faculty of Computers and Informatics, Zagazig University, Zagazig 44519, Sharqiyah, Egypt; Nada.mohamed1316@hti.edu.eg

[2] Alshaimaa A. Tantawy Faculty of Computers and Informatics, Zagazig University, Zagazig 44519, Sharqiyah, Egypt; AlshaimaaTantawy@zu.edu.eg

* Correspondence: AlshaimaaTantawy@zu.edu.eg

Abstract: Cognitive AI systems increasingly participate in decisions that were previously controlled entirely by human users, yet most adaptive interfaces still compress uncertain user states into a single class probability or a fixed automation level. This paper introduces N-CogHCI, a reliability-aware neutrosophic cognitive control framework that represents cognitive evidence as truth, indeterminacy, and falsity components and uses that state to regulate human–AI authority in real time. The framework combines reliability-weighted multimodal evidence, explicit disagreement and missingness modeling, AI epistemic uncertainty, task risk, human expertise, and a safety gate that constrains autonomy when uncertainty and consequence are jointly high. A discrete authority variable selects among human-only, human-led shared, balanced shared, AI-led shared, and autonomous modes, while the interface policy adapts explanation depth, confirmation requirements, information density, and intervention timing. Formal analysis establishes boundedness of the state representation and monotonic growth of indeterminacy as evidence quality falls or cross-source disagreement rises. A reproducible controlled Monte-Carlo stress test of 100,000 interaction states (seed 42) evaluates the controller under clean, noisy, missing, and contradictory evidence. N-CogHCI reduced high-autonomy decisions in predefined unsafe contexts to 8.31%, compared with 14.20% for crisp mean fusion and 11.92% for reliability-weighted probabilistic fusion. In high-uncertainty contexts, the corresponding rates were 15.42%, 27.74%, and 21.60%. This safety gain was obtained with a small increase in normalized decision loss (0.2053 versus 0.2037 and 0.2037), exposing rather than hiding the safety–efficiency trade-off. No empirical performance is claimed for human-participant datasets that were not executed. Instead, a leakage-resistant validation protocol is specified for COLET, SWELL-KW, and CLAS. The results support neutrosophic indeterminacy as a control variable for mixed-initiative HCI, while identifying real-user validation as the necessary next empirical step.

Keywords: neutrosophic sets; cognitive AI; human–computer interaction; human–AI teaming; cognitive workload; uncertainty; adaptive interfaces; shared control; trust calibration; soft computing.

Highlights

- A neutrosophic cognitive state preserves supportive, opposing, and unresolved evidence instead of forcing uncertain user signals into a single probability.
- A risk-aware authority controller couples user workload, AI uncertainty, evidence indeterminacy, and human expertise to five interpretable control modes.
- Formal properties guarantee bounded states and monotonic increases in indeterminacy as evidence quality deteriorates or disagreement increases.
- A 100,000-trial controlled stress test shows substantially fewer high-autonomy decisions in unsafe and high-uncertainty contexts, with a quantified efficiency trade-off.
- COLET, SWELL-KW, and CLAS are specified as public external-validation benchmarks without fabricating unexecuted empirical results.

1. Introduction

Human–AI interaction has moved from isolated recommendation interfaces toward systems that continuously infer user state, generate advice, allocate functions, and sometimes execute actions. This transition changes the HCI problem. The central design question is no longer only whether an AI model is accurate, but when the system should assist, defer, explain, request confirmation, or assume greater authority. General human–AI interaction guidelines already emphasize communicating system capability, supporting correction, and adapting behavior over time [1]. Classical automation research likewise treats function allocation as a continuum rather than a binary division between human and machine [2]. These principles become more difficult to operationalize when the user is cognitively overloaded, the AI is uncertain, the task is consequential, and the evidence about the user is itself incomplete or contradictory.

Trust further complicates this control problem. Appropriate reliance requires alignment between human trust and actual system capability rather than simply maximizing confidence in automation [3,4]. Recent work has therefore shifted toward calibrated trust [5], context-sensitive algorithm trust [6], and models of human control that move from supervision toward teaming [7]. Dynamic task allocation has also been identified as an open problem in industrial human–AI teaming, where agency, control, and risk cannot be separated from performance [8]. Optimized shared-control approaches can adapt the

confidence ratio between human and AI controllers [9], and recent calibrated adaptive human–AI frameworks explicitly incorporate probability calibration and epistemic uncertainty into closed-loop reliance [10]. These developments establish the need for adaptive authority, but they do not resolve how contradictory or missing evidence about the human cognitive state should enter the control law.

Cognitive-state-aware interfaces provide another part of the solution. Cognitive load can be incorporated into user models to support adaptive interface behavior [11], while workload instruments such as NASA-TLX remain widely used for subjective workload assessment [12]. Public datasets make cognitive-state modeling reproducible: COLET records eye behavior from 47 individuals across four induced workload levels with NASA-TLX annotation [13]; SWELL-KW records 25 participants performing knowledge-work tasks under neutral, interruption, and time-pressure conditions using behavioral and physiological modalities [14]; and CLAS provides synchronized ECG, PPG, EDA, accelerometer, and task information from 62 participants performing cognitive and affective tasks [15]. A persistent methodological weakness, however, is that uncertainty in these heterogeneous signals is usually reduced to a probability, confidence score, or missing-value flag. Such compression loses the distinction between evidence that supports a cognitive state, evidence that opposes it, and evidence that is unresolved.

Soft computing was developed precisely for reasoning in systems where exact models are unavailable or unnecessarily rigid [16]. Neutrosophic sets, introduced within Smarandache's neutrosophic framework [17] and made computationally practical through single-valued neutrosophic sets (SVNSs) [18], provide a three-component representation of truth, indeterminacy, and falsity. This structure is attractive for cognitive HCI because a user can simultaneously exhibit signals supporting high workload, signals opposing that interpretation, and unresolved evidence caused by sensor failure, individual variability, or modality conflict. Recent refined neutrosophic cognitive maps have shown how indeterminate causal relationships can be represented explicitly rather than absorbed into ordinary fuzzy membership [19]. The remaining opportunity is to move neutrosophic reasoning from post-hoc ranking into the real-time control loop of human–AI interaction.

This paper addresses that opportunity through N-CogHCI, a reliability-aware neutrosophic cognitive controller for mixed-initiative HCI. The contribution is not another interface-ranking method. N-CogHCI uses indeterminacy as an operational control variable that can reduce autonomy, trigger explanation, or force shared control when the system lacks sufficient justification for a stronger automated action. The method is deliberately formulated so that it can operate with eye tracking alone, with multimodal physiological and behavioral signals, or with partial sensor availability. The paper makes five contributions: (i) a reliability-aware neutrosophic state estimator that combines local model

uncertainty, modality reliability, disagreement, and missingness; (ii) a risk-aware human–AI authority optimization rule with a deterministic safety gate; (iii) a trust-calibration and adaptive-interface layer that responds differently to uncertainty, overtrust, and undertrust; (iv) formal properties showing boundedness and monotonic behavior of the indeterminacy term; and (v) an executable controlled stress test plus a subject-independent external-validation protocol for public cognitive HCI datasets.

2. Related Work and Research Gap

2.1 Cognitive workload and adaptive HCI

Adaptive interfaces require a user model that is updated from interaction evidence rather than a static demographic profile. Suryani et al. [11] explicitly linked cognitive-load modeling to adaptive interface development, illustrating the relevance of cognition to interface adaptation. NASA-TLX provides a multidimensional subjective workload instrument and remains a standard reference for workload studies [12]. COLET [13] complements subjective assessment with eye and gaze behavior and demonstrates that workload levels can be associated with ocular features. SWELL-KW [14] broadens the measurement space to computer interaction, facial behavior, posture, heart-rate variability, and skin conductance in realistic knowledge work. CLAS [15] adds synchronized physiological channels and cognitive tasks. Together, these datasets establish that cognitive state is observable only indirectly and through heterogeneous evidence. They do not, by themselves, specify how an adaptive interface should behave when those signals conflict or disappear.

2.2 Human–AI trust, reliance, and authority allocation

Trust has long been recognized as a determinant of reliance on automation [3,4]. The contemporary emphasis is calibration: overtrust can produce inappropriate reliance, whereas undertrust can suppress beneficial automation [5]. Yang et al. [6] further show that task context and time pressure can shape algorithm trust, reinforcing the need for dynamic rather than fixed reliance policies. Cognitive forcing functions can reduce overreliance, but they may impose subjective usability costs [20]. At the control level, Parasuraman et al. [2] conceptualized multiple types and levels of automation, and recent work has revisited human control for foundation-model systems [7], dynamic allocation in human–AI teaming [8], and shared control with adaptive human/AI confidence ratios [9]. The 2026 CAHAT framework [10] is especially close in spirit because it combines adaptive reliance, calibration, uncertainty, and closed-loop decision-making. N-CogHCI differs in its primary representation: uncertainty in the human cognitive state is not a scalar confidence deficit but an explicit neutrosophic component derived from disagreement, local uncertainty, reliability, and missingness.

2.3 Soft computing and neutrosophic modeling

Soft computing combines approximate reasoning and learning mechanisms to tolerate imprecision in complex systems [16]. Fuzzy membership is useful when a state is gradual, but ordinary membership does not separately encode unresolved evidence. Neutrosophic sets introduce truth (T), indeterminacy (I), and falsity (F) as distinct components [17], while SVNSs provide a computational formulation with components in the unit interval [18]. This separation is particularly relevant when HCI sensors are incomplete or mutually inconsistent. Refined neutrosophic cognitive maps extend this idea to causal graphs with indeterminate relationships [19]. Nevertheless, most neutrosophic decision applications remain evaluative: they rank alternatives after uncertainty has been summarized. The present work instead treats I as a first-class signal in a control policy that changes the degree and form of AI intervention.

2.4 Gap synthesis

Table 1. Positioning of N-CogHCI relative to the closest methodological streams.

Study / stream	**Human cognitive state**	**Explicit indeterminacy**	**Dynamic authority**	**Primary validation**
Adaptive HCI [11]	Yes	No	Interface adaptation	Methodology / user-model design
Human–AI control [2,7–9]	Context-dependent	Mostly probabilistic	Yes	Human factors / control studies
Calibrated trust [5,10]	Partial	Probabilistic uncertainty	Yes	Review / controlled simulation
RNCM [19]	Generic concepts	Yes	Not HCI authority allocation	Causal modeling
N-CogHCI (this work)	Yes	Yes: T–I–F	Yes: five modes + safety gate	Formal analysis + reproducible stress test

The table distinguishes the representation of uncertainty from the presence of an adaptive controller; these are separate design choices.

The gap can therefore be stated precisely: current adaptive HCI can model workload; human–AI teaming can vary authority; probabilistic frameworks can calibrate AI

uncertainty; and neutrosophic models can represent indeterminacy. What is missing is a unified controller in which the indeterminacy of the human cognitive-state estimate directly constrains authority and interface behavior. N-CogHCI is designed to fill this gap without claiming that neutrosophic logic replaces probabilistic calibration. The two are complementary: probabilities can estimate event likelihoods, while the neutrosophic state records the epistemic status of the evidence feeding the HCI controller.

3. Problem Formulation

Consider an interactive system observed at discrete interaction cycles $t = 1, 2, \ldots$. The system receives M sources of evidence about a cognitive concept c, such as high workload, low attention, stress, or comprehension difficulty. For source m, $p_{m,c}(t) \in [0,1]$ denotes the source-specific support probability, $u_{m,c}(t) \in [0,1]$ denotes local model uncertainty, $r_m(t) \in [0,1]$ denotes signal reliability, and $a_m(t) \in \{0,1\}$ denotes availability. The controller also receives an AI confidence $c_A(t)$, an AI uncertainty $u_A(t)$, task risk R(t), human expertise H(t), and—where available—an observed trust or reliance estimate $\tau(t)$. The output is an authority level $\alpha(t)$ selected from $A = \{0, 0.25, 0.50, 0.75, 1\}$. These values correspond respectively to human-only control, human-led shared control, balanced shared control, AI-led shared control with human confirmation, and autonomous AI action.

Table 2. Core notation used in N-CogHCI.

Symbol	Meaning	Range
$p_{m,c}$	support for cognitive concept c from modality m	[0,1]
$u_{m,c}$	local uncertainty of modality/model m	[0,1]
r_m	source/signal reliability	[0,1]
a_m	availability indicator	{0,1}
T_c, I_c, F_c	aggregated neutrosophic cognitive state	[0,1]^3
c_A, u_A	AI confidence and uncertainty	[0,1]
R	task/consequence risk	[0,1]
H	human expertise or capability prior	[0,1]
A	AI authority share	{0,.25,.5,.75,1}

4. Proposed N-CogHCI Framework

Figure 1 summarizes the closed-loop architecture. The state estimator converts heterogeneous user evidence into a neutrosophic cognitive state. A separate AI capability state captures confidence and epistemic uncertainty. The authority optimizer then selects a mixed-initiative mode subject to safety constraints. Trust calibration and interface adaptation operate on top of the same state so that a decision to reduce autonomy can be accompanied by an appropriate interaction response rather than a silent mode switch.

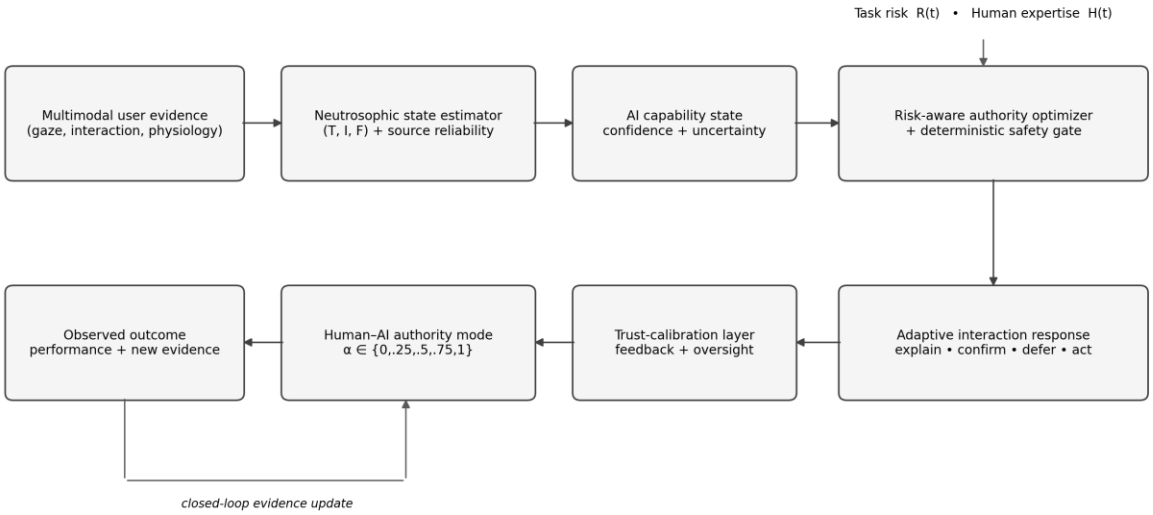

Figure 1. N-CogHCI closed-loop architecture.

4.1 Reliability-aware neutrosophic cognitive-state estimation

Only available sources contribute to the state. Reliability-normalized source weights are defined by Eq. (1), where ε is a small numerical constant. This prevents a low-quality or absent channel from contributing the same weight as a stable channel.

$$\omega_m(t) = a_m(t)\, r_m(t) \,/\, [\Sigma_j\, a_j(t)\, r_j(t) + \varepsilon] \tag{1}$$

Each source contributes discounted support and opposition, while its local uncertainty is retained. Let $\bar{p}_c(t)$ be the reliability-weighted support mean and $\bar{U}_c(t)$ the weighted mean of local uncertainties. Cross-source disagreement $D_c(t)$ is the reliability-weighted absolute deviation around $\bar{p}_c(t)$, and $Q(t)$ is the average available reliability. The N-CogHCI state is then computed as follows:

$$T_c(t) = \Sigma_m\, \omega_m(t)\, [1 - u_{m,c}(t)]\, p_{m,c}(t) \tag{2}$$

$$F_c(t) = \Sigma_m\, \omega_m(t)\, [1 - u_{m,c}(t)]\, [1 - p_{m,c}(t)] \tag{3}$$

$$D_c(t) = \Sigma_m\, \omega_m(t)\, |p_{m,c}(t) - \bar{p}_c(t)|, \qquad Q(t) = (1/M)\, \Sigma_m\, a_m(t)\, r_m(t) \tag{4}$$

$$I_c(t) = 1 - [1 - \bar{U}_c(t)]\, [1 - D_c(t)]\, Q(t) \tag{5}$$

Equation (5) is the key distinction from ordinary probability fusion. Indeterminacy can increase for three different reasons: the base models report uncertainty (U_c rises), the available sources disagree (D_c rises), or evidence quality/availability falls (Q falls). T_c and F_c therefore do not need to absorb those epistemic conditions. In the controlled stress test in Section 6, local model uncertainty was set to zero for the user-sensing channels so that the isolated effects of disagreement, reliability, and missingness could be measured; the general formulation above remains applicable when calibrated local uncertainty is available.

4.2 AI capability state

AI confidence should not be treated as competence when the model itself is uncertain. N-CogHCI represents the AI capability evidence as $N_A(t) = (T_A, I_A, F_A)$, where $T_A = (1 - u_A)c_A$, $I_A = u_A$, and $F_A = (1 - u_A)(1 - c_A)$. A calibrated confidence estimate is preferable, but the controller does not require a specific calibration algorithm. This design is compatible with probabilistic calibration and epistemic-uncertainty estimators such as those used in recent adaptive human–AI work [10]. The neutrosophic layer does not replace calibration; it exposes the uncertainty term to the HCI controller.

4.3 Risk-aware authority optimization

For clarity, the authority decision is discrete. A continuous controller can be used, but the five-mode representation is easier to audit and map to interface behavior. Let $\hat{e}_H$ and $\hat{e}_A$ be conservative estimates of human and AI error. In a learned deployment, their coefficients should be estimated from training data. In the controlled simulation, the fixed coefficients in Section 6 are used only as a reproducible stress-test model. The generic forms are:

$$\hat{e}_H = \mathrm{clip}(\beta_0 + \beta_W T_W - \beta_H H + \beta_I I_W, e_{min}, e_{max}), \qquad \hat{e}_A = \mathrm{clip}(1 - c_A + \varrho_A u_A, 0, 1) \tag{6}$$

For each candidate α, the controller evaluates a normalized decision loss. The first term is expected team error amplified by task risk. The second term models residual human cognitive burden, the third term represents coordination overhead, and the fourth term discourages extreme authority when the human-state or AI-state estimate is indeterminate. A bounded complementarity term allows shared control to be beneficial when both agents retain useful capability.

$$L(\alpha) = [1 + \lambda_R R]\, E(\alpha) + \lambda_W(1 - \alpha)T_W + \lambda_C 4\alpha(1 - \alpha) + \lambda_U(I_W + u_A)\,|\alpha - 0.5| \tag{7}$$

$$E(\alpha) = \mathrm{clip}\{\alpha\, \hat{e}_A + (1 - \alpha)\, \hat{e}_H - \eta \cdot 4\alpha(1 - \alpha) \cdot \min[1 - \hat{e}_A, 1 - \hat{e}_H], 0, 1\} \tag{8}$$

The provisional authority is $\alpha^* = \mathrm{argmin}_{\{\alpha \in A\}}\, L(\alpha)$. The complementarity coefficient η is not asserted to be a physiological constant; it is a design parameter in the controlled test and must be learned or sensitivity-tested in an empirical deployment. This explicit

statement is important: the simulation demonstrates controller behavior under a declared model, not a universal law of human–AI synergy.

4.4 Deterministic safety gate

Optimization alone is insufficient for high-consequence interaction because a small expected advantage can still select a mode that is difficult to justify under severe epistemic uncertainty. N-CogHCI therefore adds a deterministic safety gate. In the reported stress test, α is capped at 0.50 when $R \geq 0.75$ and either $I_W \geq 0.50$ or $u_A \geq 0.45$. A stricter cap of 0.25 is applied when $R \geq 0.50$ and either $I_W \geq 0.75$ or $u_A \geq 0.70$. These thresholds are design parameters, not empirical constants. Their role is to encode a transparent policy: high consequence plus low epistemic justification cannot produce high automation. In a deployment, the thresholds should be chosen through domain-specific hazard analysis and validated with users.

4.5 Trust calibration and interface adaptation

Authority allocation and user trust should not be conflated. Let $\tau(t)$ denote an observed or inferred human trust/reliance score and $c_A(t)$ the calibrated AI capability estimate. The mismatch $\Delta_\tau = \tau - c_A$ distinguishes overtrust ($\Delta_\tau > 0$) from undertrust ($\Delta_\tau < 0$). N-CogHCI uses this mismatch to choose an interaction response rather than to increase trust indiscriminately. High overtrust with high risk triggers stronger evidence presentation and confirmation; undertrust with low AI uncertainty can trigger concise capability evidence; high cognitive indeterminacy triggers clarification or delayed intervention. This follows the broader objective of calibrated reliance described in the trust literature [3–6] and avoids treating explanation as a universal remedy. Indeed, cognitive forcing can reduce overreliance at the cost of user preference [20], so explanation depth and interruption frequency should themselves be adaptive.

Table 3. Mapping from N-CogHCI state conditions to authority and interface actions.

Condition	Authority response	Interface response	Rationale
Low I, low risk, strong AI capability	Allow α chosen by optimizer	Minimal explanation; low friction	Evidence is sufficient and consequence is limited
High workload, low I, reliable AI	Shift toward AI-led support	Simplify display; suppress non-critical notifications	Reduce avoidable human cognitive burden

Condition	Authority response	Interface response	Rationale
High I or high u_A	Prefer shared control	Expose uncertainty; request confirmation or more evidence	Avoid false certainty
High risk + high uncertainty	Cap $\alpha \leq 0.50$ or 0.25	Mandatory confirmation; defer irreversible action	Safety gate
Overtrust ($\tau > c_A$)	Do not raise α solely from trust	Show limitations/evidence; cognitive forcing when justified	Calibrate reliance rather than maximize trust
Undertrust with reliable AI	Keep risk constraints	Brief performance evidence; preserve user override	Support justified reliance without coercion

4.6 Online algorithm

Algorithm 1 gives the per-cycle logic. The algorithm is deliberately auditable: every transition into greater autonomy has an explicit evidence state, risk value, and selected loss. It can therefore be logged for post-hoc safety analysis.

Step	Operation
1	Acquire available user-sensing evidence p_m,c, uncertainty u_m,c, reliability r_m, and availability a_m.
2	Compute normalized weights ω_m and the neutrosophic cognitive state N_c = (T_c, I_c, F_c).
3	Construct AI capability state from calibrated confidence c_A and uncertainty u_A.
4	Estimate conservative human and AI error terms ê_H and ê_A.
5	Evaluate $L(\alpha)$ for $\alpha \in \{0, .25, .50, .75, 1\}$ and choose the minimizing provisional mode.
6	Apply the deterministic risk–uncertainty safety gate.
7	Use trust mismatch and cognitive state to select explanation, confirmation, simplification, or deferral behavior.
8	Log outcome, update reliability/calibration statistics, and repeat at the next interaction cycle.

5. Formal Properties

Proposition 1 — Boundedness

If p_m,c, u_m,c, r_m ∈ [0,1] and a_m ∈ {0,1}, with at least one available source of nonzero reliability, then T_c, F_c, I_c ∈ [0,1]. Proof: the normalized weights are nonnegative and sum to one. Therefore T_c and F_c are convex combinations of terms in [0,1]. Likewise U_c, D_c, and Q are in [0,1]. The product [1 − U_c][1 − D_c]Q is therefore in [0,1], and Eq. (5) implies I_c ∈ [0,1]. This property prevents numerical states outside the SVNS unit interval.

Proposition 2 — Quality monotonicity of indeterminacy

Holding U_c and D_c fixed, $\partial I_c/\partial Q = -[1 - U_c][1 - D_c] \leq 0$. Thus, reducing aggregate evidence quality or availability cannot reduce indeterminacy. In the limiting case $Q \to 0$, $I_c \to 1$. This gives missing evidence a direct and predictable effect rather than silently imputing confidence.

Proposition 3 — Disagreement monotonicity

Holding U_c and Q fixed, $\partial I_c/\partial D_c = [1 - U_c]Q \geq 0$. Therefore greater cross-source disagreement cannot reduce indeterminacy. The property is valuable in multimodal HCI because contradictory gaze, interaction, or physiological evidence should not produce the same authority decision as coherent evidence of equal mean magnitude.

Proposition 4 — Safety-cap invariance

Under the reported gate, any state satisfying R ≥ 0.75 and (I_W ≥ 0.50 or u_A ≥ 0.45) produces $\alpha \leq 0.50$ regardless of the unconstrained optimizer. Likewise, any state satisfying R ≥ 0.50 and (I_W ≥ 0.75 or u_A ≥ 0.70) produces $\alpha \leq 0.25$. This is a policy guarantee, not a statistical tendency. Its value is auditability: the system cannot enter an AI-led or autonomous mode in a state explicitly classified by the gate as insufficiently justified.

6. Controlled Reproducible Evaluation

6.1 Purpose and scope

The evaluation is a controlled stress test of the authority controller, not a claim of human cognitive-state prediction accuracy. This distinction avoids a common methodological error in framework papers: assigning invented performance numbers to a public dataset that was not executed. The test asks a narrower question: when evidence is noisy, missing, or contradictory, does explicit indeterminacy reduce high-autonomy decisions in contexts that have been defined as unsafe? All data-generation assumptions are stated below, and the accompanying script NCogHCI_reproducibility.py reproduces the results exactly with seed 42.

6.2 Monte-Carlo design

The simulation generates 100,000 independent interaction states. Latent workload W is sampled from Beta(2.2, 2.2), task risk R from Beta(2, 2), human expertise H from Beta(2.5, 2), and nominal AI confidence from Beta(5, 2). AI epistemic uncertainty is sampled from a two-component mixture: 18% of trials use Beta(3, 2), representing high-uncertainty episodes, and the remaining trials use Beta(1.3, 6). True AI competence is generated by perturbing nominal confidence with zero-mean Gaussian noise whose standard deviation is 0.035 + 0.22u_A; thus, higher uncertainty increases the dispersion between confidence and realized competence without forcing a particular direction of error.

Four user-evidence regimes are sampled with probabilities 0.55 clean, 0.20 noisy, 0.12 missing, and 0.13 contradictory. Four sensing channels are generated. Clean channels use Gaussian noise $\sigma = 0.07$, availability probability 0.99, and base reliability 0.92. Noisy channels use $\sigma = 0.18$, availability 0.95, and reliability 0.68. Missing-evidence trials use $\sigma = 0.10$, availability 0.55, and reliability 0.82. Contradictory trials use $\sigma = 0.09$, availability 0.98, and reliability 0.84, while one or two channels are inverted around the latent workload. Reliability receives additional zero-mean noise of standard deviation 0.08 and is clipped to [0.15, 1]. These are declared stress-test parameters, not estimates of any specific human population.

The authority levels are {0, .25, .50, .75, 1}. The simulation instantiates Eq. (6) as $\hat{e}_H = \text{clip}(0.08 + 0.50T_W - 0.28H + 0.054I_W, 0.03, 0.65)$ and $\hat{e}_A = \text{clip}(1 - c_A + 0.105u_A, 0, 1)$. Equation (7) uses $\lambda_R = 0.65$, $\lambda_W = 0.07$, $\lambda_C = 0.025$, $\lambda_U = 0.02$, and $\eta = 0.10$. The reliability-weighted baseline receives reliability-aware T_W and AI uncertainty but not cognitive indeterminacy or the N-CogHCI safety gate. The crisp baseline uses an unweighted mean of available cognitive evidence and no explicit uncertainty penalty.

An unsafe context is defined before policy evaluation as $R \geq 0.70$ and either true AI competence < 0.70 or the user-evidence regime is missing/contradictory. A high-uncertainty context is defined as $I_W > 0.50$ or $u_A > 0.45$. A high-autonomy decision is $\alpha \geq 0.75$. The normalized decision-loss value is a model-based stress-test quantity and has no physical unit; lower is better within the declared simulation model.

6.3 Main results

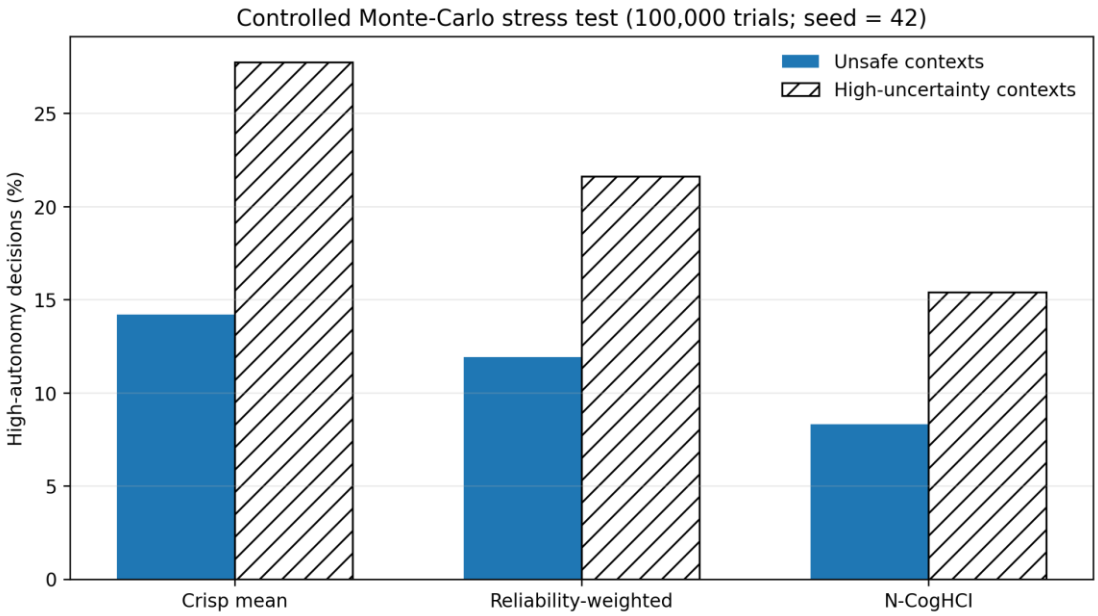

Figure 2. High-autonomy rates in predefined unsafe and high-uncertainty contexts.

Table 4. Controlled Monte-Carlo results. Confidence intervals are normal-approximation binomial intervals over the predefined context subsets (unsafe n = 12,376; high-uncertainty n = 26,980).

Policy	Decision loss	Unsafe high autonomy	95% CI	High-uncertainty high autonomy	95% CI
Crisp mean	0.2037	14.20%	[13.58, 14.81]%	27.74%	[27.20, 28.27]%
Reliability-weighted	0.2037	11.92%	[11.35, 12.49]%	21.60%	[21.11, 22.10]%
N-CogHCI	0.2053	8.31%	[7.82, 8.79]%	15.42%	[14.98, 15.85]%

N-CogHCI selected high autonomy in 8.31% of unsafe contexts (95% CI 7.82–8.79%), compared with 14.20% (13.58–14.81%) for crisp fusion and 11.92% (11.35–12.49%) for reliability-weighted fusion. Relative to the crisp baseline, this is a 41.5% reduction in high-autonomy decisions under the declared unsafe-context definition; relative to reliability-weighted fusion, the reduction is approximately 30.3%. In high-uncertainty contexts, N-

CogHCI produced high autonomy in 15.42% of trials, compared with 27.74% for crisp fusion and 21.60% for reliability-weighted fusion. The corresponding relative reductions are approximately 44.4% and 28.6%.

The safety improvement is not free. The normalized decision loss was 0.2053 for N-CogHCI, versus 0.2037 for crisp fusion and 0.2037 for reliability-weighted fusion. The increase relative to the best baseline is less than 1% in this simulation, but it is still a real trade-off and is reported as such. The result should therefore be interpreted as evidence that explicit indeterminacy and a safety gate can move the controller along a safety–efficiency frontier, not as evidence that N-CogHCI dominates every objective simultaneously.

6.4 Ablation analysis

Table 5. Ablation results from the same 100,000-trial simulation.

Variant	Decision loss	Unsafe high autonomy	High-uncertainty high autonomy	Shared-control rate
Full N-CogHCI	0.2053	8.31%	15.42%	53.20%
No safety gate	0.2038	12.43%	20.53%	52.62%
No indeterminacy term	0.2050	9.39%	15.76%	51.62%
No AI uncertainty	0.2048	12.39%	27.01%	54.58%

Removing the safety gate lowered the modeled decision loss from 0.2053 to 0.2038 but increased unsafe high-autonomy decisions from 8.31% to 12.43%. Removing cognitive indeterminacy increased the unsafe rate to 9.39%, while removing AI uncertainty increased it to 12.39% and raised high-autonomy decisions in high-uncertainty contexts from 15.42% to 27.01%. The ablation therefore isolates two distinct mechanisms. The continuous uncertainty penalties affect the optimizer before mode selection; the deterministic gate provides an additional hard constraint in states where expected-loss optimization alone remains too permissive.

6.5 Worked numerical case study

Table 6 illustrates four individual controller states using the full estimator of Section 4.1, including local modality uncertainty. These examples are deterministic calculations, not

sampled human observations. They show how similar average evidence can produce different decisions when reliability and indeterminacy differ.

Table 6. Deterministic worked examples of N-CogHCI decisions.

Scenario	T_W	I_W	F_W	Risk	AI uncertainty	Selected α / mode
A. Low workload, coherent evidence	0.18	0.132	0.778	0.25	0.08	0.25 — human-led shared
B. High workload, coherent evidence	0.779	0.136	0.176	0.45	0.1	1.00 — autonomous AI
C. Contradictory evidence, high risk	0.46	0.436	0.452	0.82	0.38	0.50 — balanced shared
D. Missing evidence, uncertain AI	0.639	0.626	0.261	0.72	0.58	0.50 — balanced shared

Scenario C is the key neutrosophic example. Support for high workload and opposition are both substantial, and disagreement raises I_W. A scalar average alone could conceal the conflict. N-CogHCI keeps the conflict visible and selects balanced shared control under high risk. Scenario D reaches an even larger I_W primarily because only half of the sensing channels are available and the AI state is itself uncertain. The controller therefore avoids AI-led or autonomous control despite moderately high evidence of user workload.

7. External Validation Protocol on Public HCI Datasets

The controlled test above validates the controller's internal behavior under known perturbations. It does not establish real-user predictive validity. A publishable empirical extension should therefore use participant-level evaluation with leakage control. Three datasets are appropriate and complementary. COLET contains 47 participants, four cognitive-workload levels, eye tracking, and NASA-TLX annotations [13,21]. SWELL-KW contains 25 participants performing realistic knowledge work under neutral, interruption, and time-pressure conditions with computer interaction, facial behavior, posture, heart-rate

variability, and skin conductance [14,22]. CLAS contains 62 participants with synchronized ECG, PPG, EDA, accelerometer data, and interactive Math, Logic, and Stroop tasks as well as perceptive affective tasks [15].

7.1 Leakage-resistant evaluation design

All preprocessing and model fitting must be participant-grouped. Windows from the same participant must never appear in both training and test partitions. The preferred primary protocol is leave-one-subject-out evaluation; grouped nested cross-validation is an acceptable alternative when hyperparameter tuning would otherwise reuse test subjects. Standardization, feature selection, calibration, and reliability estimation must be fitted exclusively on each training fold. This requirement is especially important for physiological and interaction features because person-specific baselines can make random-window splits appear unrealistically accurate.

For COLET, eye features can be organized into functional source groups (for example pupil, fixation, saccade, and task-performance groups) so that N-CogHCI receives multiple reliability-tagged evidence sources even though the dataset is eye-centric. For SWELL-KW, computer interaction, face, posture, and physiology can form natural modalities. For CLAS, ECG/heart features, PPG, EDA, and task-performance evidence can be modeled separately. Each base learner should output both a calibrated support probability and an uncertainty quantity. Reliability can be estimated from signal quality, missingness, or validation-time calibration error.

7.2 Evaluation metrics and falsifiable hypotheses

The state estimator should be evaluated with macro-F1 and balanced accuracy for workload/stress classification, Brier score and expected calibration error for probability quality, and selective risk/coverage curves for abstention behavior. The controller should additionally report unsafe high-autonomy rate under prospectively declared risk criteria, unnecessary human-takeover rate, shared-control frequency, authority-switching rate, and task-level performance. A user study should measure NASA-TLX, perceived control, trust calibration, task accuracy, completion time, and intervention burden.

The framework yields falsifiable hypotheses rather than guaranteed outcomes. H1: adding I_W should improve selective decision safety under missing or contradictory evidence without requiring higher raw classification accuracy. H2: the safety gate should reduce high-autonomy decisions in high-risk/high-uncertainty states. H3: trust-calibrated interface responses should reduce overreliance compared with confidence-only explanations, but may increase interaction cost, consistent with the trade-offs observed in cognitive-forcing research [20]. H4: subject-independent benefits should be larger under modality corruption than under clean evidence. Failure of these hypotheses on public

datasets or user studies would be evidence against the practical value of the proposed controller and should be reported rather than hidden.

8. Discussion

8.1 Why neutrosophic indeterminacy matters in HCI

The main conceptual result is that uncertainty about a user is not equivalent to a middling estimate of the user's state. A workload probability of 0.50 can arise because every source weakly supports moderate workload, because half the sensors strongly support high workload and half strongly oppose it, or because most sensors are missing. Those states may require different interaction policies. Equations (2)–(5) separate them: T and F describe directional evidence, while I records the epistemic reason not to act as though the direction were settled. This is the point at which the neutrosophic representation becomes operationally meaningful rather than decorative.

8.2 Relationship to probabilistic calibration and fuzzy systems

N-CogHCI should not be read as an argument that probabilities are obsolete. AI confidence calibration remains essential, as emphasized by recent adaptive teaming frameworks [10]. The proposed model uses probability-like support values inside each modality and then adds an epistemic layer that can represent unresolved evidence. Similarly, fuzzy membership is appropriate for gradual concepts such as 'moderately overloaded.' The neutrosophic state adds a distinct dimension when the evidence for that fuzzy concept is incomplete or contradictory. A practical implementation can therefore combine calibrated probabilistic classifiers, fuzzy linguistic rules, and neutrosophic evidence fusion within one soft-computing pipeline.

8.3 Safety–efficiency trade-off

The simulation does not show universal dominance. N-CogHCI intentionally sacrifices a small amount of modeled decision efficiency to reduce high autonomy in states that lack sufficient justification. This result is preferable to an apparently superior table produced by tuning both the policy and the evaluation metric to favor the proposed method. In safety-relevant HCI, a method should expose its operating trade-off. The correct setting of the gate depends on application severity: a writing assistant may tolerate more autonomy than a clinical, cybersecurity, industrial-control, or transportation interface.

8.4 Interpretability and auditability

The controller is structured so that an authority decision can be explained using a small set of variables: workload evidence, indeterminacy, AI uncertainty, task risk, and expertise. This is useful for HCI auditing because a reviewer can reconstruct why the system refused autonomous action even when AI confidence was high. The deterministic gate also makes a

subset of safety behavior testable with unit tests rather than only statistical evaluation. This does not guarantee overall system safety; it makes one important class of behavior explicit and inspectable.

9. Limitations and Threats to Validity

First, the reported numerical results are simulation-based. The distributions, coefficients, complementarity term, and safety thresholds are controlled design assumptions, not population estimates. They test internal behavior under transparent perturbations but cannot establish human-subject effectiveness. Second, cognitive concepts differ across tasks. Eye-based workload in COLET, workplace stress in SWELL-KW, and physiology in CLAS are related but not interchangeable targets. Cross-dataset validation must preserve those semantic differences rather than forcing a single label taxonomy. Third, reliability estimates can themselves be wrong; a future implementation should calibrate reliability and examine adversarial or systematic sensor failure. Fourth, the present controller uses five authority levels. Continuous or partially observable control may provide smoother behavior, but it would also require stronger stability and human-factors analysis. Fifth, trust is included as a calibration signal, yet trust inference from behavior is itself uncertain and should not be treated as ground truth. Finally, the framework does not replace domain safety engineering, usability testing, or legal requirements for human oversight.

10. Reproducibility, Data, and Ethical Scope

No new human participants were recruited and no private participant data were created for this study. The numerical results in Section 6 are generated from a controlled synthetic stress test whose distributions, policy parameters, random seed, and evaluation definitions are stated in the manuscript. The accompanying file NCogHCI_reproducibility.py reproduces the Monte-Carlo tables and exports the result CSV files. COLET, SWELL-KW, and CLAS are referenced only as public external-validation resources; this manuscript does not claim to have executed participant-level experiments on them. COLET is publicly archived on Zenodo [21], SWELL-KW has a persistent DANS dataset record [22], and CLAS is available through its published data repository and access procedure [15]. Any future human-participant deployment should obtain the approvals appropriate to its institution and application domain and should evaluate perceived control, accessibility, privacy, and intervention burden in addition to predictive metrics.

11. Conclusion and Future Work

This paper proposed N-CogHCI, a reliability-aware neutrosophic cognitive control framework for adaptive human–AI interaction. Its central design principle is that unresolved evidence should change what an AI system is allowed to do, not merely lower

a confidence score displayed after the decision. The framework separates supportive evidence, opposition, and indeterminacy; couples cognitive state with AI uncertainty and task risk; selects among interpretable authority modes; and adapts interface behavior to trust and cognitive context. Formal analysis shows that indeterminacy rises predictably as evidence quality falls or disagreement increases. In a reproducible 100,000-trial stress test, the full controller substantially reduced high-autonomy decisions in unsafe and high-uncertainty contexts, while exposing a small cost in normalized decision efficiency rather than concealing it.

The next step is empirical, not rhetorical. The state estimator and authority policy should be evaluated subject-independently on COLET, SWELL-KW, and CLAS, followed by a controlled user study in which workload, trust calibration, perceived control, task performance, and intervention burden are measured together. The most important future extension is learning the authority-loss coefficients and safety thresholds from domain-specific evidence while preserving hard constraints for high-consequence states. A second extension is to integrate refined neutrosophic causal maps [19] so that the controller can model delayed relationships among workload, attention, explanation, trust, and performance. A third is to study personalized/federated reliability models so that adaptation can improve without centralizing sensitive cognitive data. These directions preserve the core contribution: cognitive AI should not only infer what the user may be experiencing; it should reason explicitly about how uncertain that inference is before deciding how much control to assume.

References

[1] Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for Human-AI Interaction. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Paper 3, 1–13. https://doi.org/10.1145/3290605.3300233

[2] Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 30(3), 286–297. https://doi.org/10.1109/3468.844354

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

[4] Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. Human Factors, 57(3), 407–434. https://doi.org/10.1177/0018720814547570

[5] Lucas, G. M., Becerik-Gerber, B., & Roll, S. C. (2024). Calibrating workers' trust in intelligent automated systems. Patterns, 5(9), 101045. https://doi.org/10.1016/j.patter.2024.101045

[6] Yang, R., Li, S., Qi, Y., Liu, J., He, Q., & Zhao, H. (2025). Unveiling users' algorithm trust: The role of task objectivity, time pressure, and cognitive load. Computers in Human Behavior Reports, 18, 100667. https://doi.org/10.1016/j.chbr.2025.100667

[7] Tsamados, A., Floridi, L., & Taddeo, M. (2025). Human control of AI systems: From supervision to teaming. AI and Ethics, 5, 1535–1548. https://doi.org/10.1007/s43681-024-00489-4

[8] Bokhorst, J. A. C., Waschull, S., & Emmanouilidis, C. (2025). Dynamic Allocation of Shared Tasks in Industrial Human-AI Teaming. IFAC-PapersOnLine, 59(10), 757–762. https://doi.org/10.1016/j.ifacol.2025.09.129

[9] Tan, J., Xue, S., Cao, H., & Ge, S. S. (2025). Human–AI interactive optimized shared control. Journal of Automation and Intelligence, 4(3), 163–176. https://doi.org/10.1016/j.jai.2025.01.001

[10] AL-Ali, M., Marks, A., Mohamed, A. A., Balaha, H. M., Badawy, M., Elhosseini, M. A., et al. (2026). Calibrated adaptive framework for trustworthy human and artificial intelligence decision systems. Scientific Reports. https://doi.org/10.1038/s41598-026-65730-y

[11] Suryani, M., Sensuse, D. I., Santoso, H. B., Aji, R. F., Hadi, S., Suryono, R. R., & Kautsarina. (2024). An initial user model design for adaptive interface development in learning management system based on cognitive load. Cognition, Technology & Work, 26, 653–672. https://doi.org/10.1007/s10111-024-00772-8

[12] Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), Human Mental Workload (Advances in Psychology, Vol. 52, pp. 139–183). Elsevier. https://doi.org/10.1016/S0166-4115(08)62386-9

[13] Ktistakis, E., Skaramagkas, V., Manousos, D., Tachos, N. S., Tripoliti, E., Fotiadis, D. I., & Tsiknakis, M. (2022). COLET: A dataset for cognitive workload estimation based on eye-tracking. Computer Methods and Programs in Biomedicine, 224, 106989. https://doi.org/10.1016/j.cmpb.2022.106989

[14] Koldijk, S. J., Sappelli, M., Verberne, S., Neerincx, M. A., & Kraaij, W. (2014). The SWELL Knowledge Work Dataset for Stress and User Modeling Research. Proceedings of the 16th ACM International Conference on Multimodal Interaction, 291–298. https://doi.org/10.1145/2663204.2663257

[15] Markova, V., Ganchev, T., & Kalinkov, K. (2019). CLAS: A Database for Cognitive Load, Affect and Stress Recognition. 2019 International Conference on Biomedical Innovations and Applications (BIA), 1–4. https://doi.org/10.1109/BIA48344.2019.8967457

[16] Zadeh, L. A. (1994). Fuzzy logic, neural networks, and soft computing. Communications of the ACM, 37(3), 77–84. https://doi.org/10.1145/175247.175255

[17] Smarandache, F. (1998). Neutrosophy: Neutrosophic Probability, Set, and Logic: Analytic Synthesis & Synthetic Analysis. American Research Press, Rehoboth, New Mexico.

[18] Wang, H., Smarandache, F., Zhang, Y., & Sunderraman, R. (2010). Single Valued Neutrosophic Sets. Multispace and Multistructure, 4, 410–413.

[19] Shekatkar, A. R., Deodeshmukh, A., & Kandasamy, I. (2025). Refined Neutrosophic Cognitive Maps: A Novel Framework for Modeling Indeterminate Causal Relationships. International Journal of Fuzzy Systems. https://doi.org/10.1007/s40815-025-02190-y

[20] Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1), Article 188, 1–21. https://doi.org/10.1145/3449287

[21] Ktistakis, E., Skaramagkas, V., Manousos, D., Tachos, N. S., Tripoliti, E., Fotiadis, D. I., & Tsiknakis, M. (2023). COLET: A Dataset for Cognitive workLoad estimation based on Eye-Tracking (Version 3) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7766785

[22] Kraaij, W., Koldijk, S., & Sappelli, M. (2014). The SWELL Knowledge Work Dataset for Stress and User Modeling Research (Version 4.1) [Data set]. DANS Data Station Social Sciences and Humanities.

Appendix A. Exact Controlled-Study Parameters

The following values are included to make the reported numerical tables independently reproducible. They are simulation parameters and must not be interpreted as empirical estimates of human cognition.

Component	Parameter	Value
Trials	N / random seed	100,000 / 42
Regime mixture	clean / noisy / missing / contradictory	0.55 / 0.20 / 0.12 / 0.13
Latent workload	Beta parameters	2.2, 2.2
Risk	Beta parameters	2, 2
Human expertise	Beta parameters	2.5, 2
AI confidence	Beta parameters	5, 2
High-AI-uncertainty mixture	probability; Beta parameters	0.18; Beta(3,2)
Ordinary AI uncertainty	Beta parameters	1.3, 6

Component	**Parameter**	**Value**
Authority levels	α	0, .25, .50, .75, 1
Loss weights	λ_R, λ_W, λ_C, λ_U, η	0.65, 0.07, 0.025, 0.02, 0.10
Safety gate 1	condition / cap	R≥.75 and (I≥.50 or u_A≥.45) / α≤.50
Safety gate 2	condition / cap	R≥.50 and (I≥.75 or u_A≥.70) / α≤.25

Received: Aug. 31, 2026. **Accepted:** Sept. 15, 2026.