Hallucinated Agents: A Cognitive Mechanism for Resolving Ambiguous Threats

A Bounded-Time Computational Cognitive Architecture for Multimodal Ambiguity Resolution in Theoretical Cognitive and Behavioral Psychology

Problem Definition: Threat processing in biological systems is optimized for rapid action under uncertainty, yet prolonged ambiguity at high arousal creates a dangerous control problem. Continued evaluation stalls policy computation while catecholamine "doping" accumulates and cognitive coherence fractures, elevating risks of autonomic instability (arrhythmias, stress cardiomyopathy) and maladaptive neuroplasticity. Current accounts treat agentive or causal hallucinations as mere biases or errors; this framing misses the timing imperative that makes indefinite dwell physiologically intolerable and computationally unsafe.

Proposed Contribution: We introduce a bounded-time controller embedded in the defensive cascade that enforces termination when person-specific thresholds of unresolved time (τ^*) and composite load (λ^*) are exceeded. Resolution occurs via minimal explanatory insertions: a Minimal Viable Agent (MVA) for external ambiguity and an Internal Minimal Viable Cause (IMVC) for interoceptive ambiguity. These primitives are faulty-but-functional—they trade veridicality for guaranteed bounded decision time under asymmetric costs (missed lethal threat ≫ false alarm) and physiological constraints. The framework is reductionist: it unifies disparate phenomena under a single timing mechanism.

Theoretical Foundations: The controller monitors normalized variables: fear/arousal F_t, urgency U_t, $ambiguity\ Amb_t$. Composite load $R_t = \alpha F_t + \beta U_t + \gamma\ Amb_t$ drives a dual-threshold stop rule: $t_{unresolved} \geq \tau^* \wedge R_{(t)} \geq \lambda^* \wedge Amb(t) \approx 1$. External threats trigger MVA insertion (minimal hostile agent with intent, location, trajectory); internal threats trigger IMVC insertion (benign causal re-labeling). Both restore a sloped action surface, terminating the stall and halting further doping and coherence loss.

Cross-Domain Mapping: constraint topology · timing under uncertainty · structural inference · alignment dynamics · error-management bias · probabilistic cognition · recursive threat resolution · macro-to-micro mapping · agency detection as functional primitive · bounded rationality under physiological load · emergent termination heuristics

Scope and Intent: This paper provides a conceptual foundation and formal controller model. It is not an empirical validation, diagnostic tool, or clinical protocol. The aim is to offer testable primitives and a unified explanatory geometry for threat-induced hallucinations, enabling future modeling, VR experiments, and resilience interventions.

ORCID: 0009-0004-5339-5373
3Pilgrim, LLC:

DOI: 10.5281/zenodo.18615682
www.3pilgrim.com

Keywords
bounded-time controller, ambiguity resolution, defensive cascade, minimal viable agent (MVA), internal minimal viable cause (IMVC), threat-driven hallucination, freeze response, amygdala–PAG circuitry, sympathetic arousal, catecholamine surge, stress cardiomyopathy (Takotsubo), error management theory, agentive completion (HADD), intolerance of uncertainty, catastrophic interoception, panic loop, action-surface collapse, cognitive re-anchoring, cross-modal ambiguity, VR threat paradigms, interoceptive perturbation, survival decision-making under uncertainty.

Orientation for Interpretation
This paper proposes a set of conceptual primitives and a formal timing architecture for threat processing under ambiguity. It is reductionist by design, prioritizing structural relationships over empirical exhaustiveness. Claims are provisional and functional: the model explains why certain hallucinations may exist as safety mechanisms rather than asserting they are always present or optimal. Readers should expect abstraction and cross-domain anchors before any applied discussion. The purpose is to define useful structure for future work, not to deliver perfect truth or immediate interventions.

Potential for Clinical and Translational Applications
The bounded time controller proposed here *may* offer a unified theoretical account of several phenomena commonly discussed in clinical psychology, including panic attacks, catastrophic interoception, hypervigilance, freeze–fight switching failures, and acute stress responses. In frameworks describing panic disorder, one *interpretation* is that delayed installation of an Internal Minimal Viable Cause (IMVC) could prolong maladaptive interpretations of benign interoceptive signals, sustaining sympathetic load and accelerating the panic loop. Established interventions—such as interoceptive exposure, cognitive relabeling, and controlled breathing—*can be conceptualized* within this model as external scaffolds that raise the effective thresholds (τ^* and λ^*), thereby restoring bounded resolution. This reframing is theoretical, not clinical guidance.

In trauma and PTSD-adjacent profiles, persistent freeze states may be *conceptually modeled* as shifts in bounded-time thresholds that trap the agent in prolonged ambiguity zones where MVA or IMVC generation becomes prematurely triggered (false-positive threat inference) or delayed (freeze retention). Within this architectural framing, strategies like graded ambiguity exposure, VR-based disambiguation, and procedural off-ramps (e.g., structured scripts, sensory grounding) can be interpreted as hypothesized stabilizers for termination dynamics. This interpretation is speculative and intended for theoretical mapping only.

More broadly, the framework suggests a possible relational structure between stress-induced sympathetic overload—discussed in literature on Takotsubo cardiomyopathy, arrhythmias, and autonomic dysregulation—and timing failures under uncertainty. By expressing these phenomena through a unified computational architecture, the model proposes a conceptual translational bridge linking psychophysiology, computational psychiatry, and real-time behavioral neuroscience. These connections are theoretical and do not imply diagnostic or therapeutic utility.

ORCID: 0009-0004-5339-5373

DOI: 10.5281/zenodo.18615682

Abstract

Young children imagine monsters under their beds; adults see ghosts in the shadows or malevolent intent in fleeting movements where none is evident. This persistent tendency to hallucinate agents or causes—far from being a cognitive glitch—represents an essential evolutionary-cognitive solution to a profound problem: situational ambiguity under threat, where disambiguation cannot be rapidly resolved due to absent or conflicting orienting feedback.

The problem is architectural. Biological systems evolved defensive cascades optimized for rapid time-to-action against predators or injury. A simple model illustrates this: a small rodent, such as a mouse, startled by noise or motion, will pause, raise its head, orient, and then flee. This exemplifies how, upon a fear trigger, subcortical circuits (amygdala → periaqueductal gray) mobilize in tens of milliseconds(Roelofs, 2017; George et al., 2019): perceptual gain rises, attention narrows, sympathetic activation surges catecholamines, heart rate and contractility increase, and the body primes for active fight or flight (Roelofs, 2017). In parallel, parasympathetic modulation can impose a temporary brake, manifesting as momentary freeze—a state of behavioral inhibition with heart rate deceleration that supports covert vigilance, evidence accumulation, and action preparation in low-deliberation species.

In humans, however, cortical elaboration can extend this state by injecting cognitive disorientation or confusion. When ambiguity persists at high arousal—stalling policy computation—dual risks emerge. Chemically, doping builds: the brain (primarily via amygdala-orchestrated sympathetic pathways) dumps adrenaline and other catecholamines into the system, sustaining overload. Cognitively, coherence fractures under these potent chemicals, promoting maladaptive learning under sustained catecholaminergic load (e.g., aberrant synaptic plasticity in prefrontal-amygdala networks) (Maddox et al., 2019). Together, these turn the cascade physiologically hazardous. Prolonged sympathetic dominance and unresolved uncertainty destabilize autonomic regulation, elevating risks of arrhythmias or stress cardiomyopathy (Takotsubo), where excessive catecholamine surges cause direct myocardial toxicity, stunning, and decompensation—often amplified by chronic psychological burdens such as sustained uncertainty or overload (Harvard Health Publishing, 2023; Wittstein et al., 2005). The system cannot tolerate abnormally long dwell in such doped, high-risk configurations. This intertwined chemical-cognitive risk demands a circuit-breaker to enforce bounded time-to-decision, halting further endogenous "doping" and restoring coherence before damage accumulates.

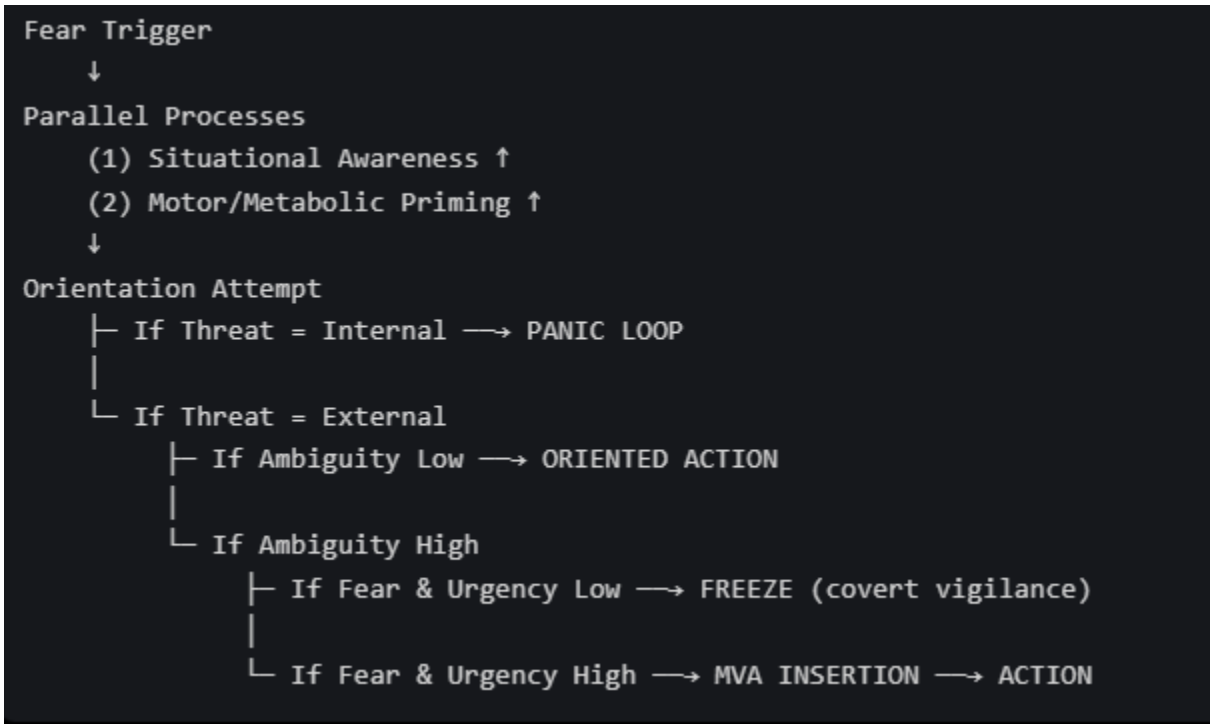
We propose this circuit-breaker operates via minimal explanatory insertions—"hallucinated" agents or causes. For external ambiguity, the mind fabricates a Minimal Viable Agent (MVA)—a fictional but computationally sufficient adversary (with intent, location, trajectory) that collapses the action surface and restores directional policy (orient/avoid/flee/inspect). For interoceptive ambiguity (panic loops), it inserts an Internal Minimal Viable Cause (IMVC)—a benign re-labeling ("this is adrenaline," "this will pass") that halts catastrophic escalation and enables downregulation.

These primitives are faulty-but-functional: they trade veridicality for guaranteed bounded time-to-decision under asymmetric costs (missed lethal threat ≫ false alarm) and physiological constraints. This

DOI: 10.5281/zenodo.18615682

paper formalizes the mechanism as a bounded-time controller within the defensive cascade, integrates supporting evidence from error management theory, hyperactive agency detection, compensatory control, meaning maintenance, and intolerance of uncertainty, and derives testable predictions across modalities, arousal levels, and individual differences.

Figure 1. Core control flow of the bounded-time defensive cascade.



Caption: The architectural spine of the model. Fear triggers parallel activation of awareness and priming. Internal threats feed into the self-amplifying panic loop; external threats branch on ambiguity and load, with the MVA serving as the circuit-breaker under high-load conditions. The IMVC counterpart resolves internal loops (detailed in §5).

The proposed architecture is captured in Figure 1. This control flow unifies the external and internal forks of threat processing under a single bounded-time controller. Subsequent sections elaborate each branch: §2 details the defensive cascade feeding into orientation, §3 formalizes the control problem, §4–5 describe the off-ramps, and later sections derive predictions and implications.

1. The Control Problem of Ambiguity

Functional cognition, regardless of environmental condition, is predicated on a single factor: structure. Ambiguity is uniquely destabilizing because it presents information without organizing structure. This creates a decision state where the action surface is flat—no clear path

DOI: 10.5281/zenodo.18615682

of reduced friction, no differential risk gradients among options. All paths appear equivalent in risk profile, rendering rational decision-making impossible and resulting in stalled policy computation.

In this stalled state, dual risks accumulate, elevating both physiological and cognitive stress from the ramping perception of imminent existential threat. When threat is clear, the action surface is well-defined with a steep downward slope away from the threat vector: action follows immediately, decision cost is low, and resolution is rapid. When ambiguity is low, orientation proceeds efficiently. But when ambiguity (Amb_t) is high, fear/arousal (F_t) is elevated, and urgency (U_t) is rising, continued evaluation yields diminishing returns in decision quality while incurring escalating costs.

The organism faces three interlocking constraints that make indefinite dwell intolerable:

- Physiological (chemical) constraint: Sustained catecholamine "doping" under unresolved high arousal is toxic, risking autonomic instability, electrophysiological volatility, arrhythmias, or stress cardiomyopathy (Takotsubo) via direct myocardial damage and decompensation (Wittstein et al., 2005; Harvard Health Publishing, 2023).
- Cognitive constraint: Fractured coherence in the presence of potent stress chemicals promotes maladaptive neuronal changes—excitotoxicity, oxidative stress, aberrant synaptic plasticity in threat-processing circuits (prefrontal-amygdala networks)—potentially entrenching dysfunctional representations.
- Evolutionary/computational constraint: Under asymmetric costs (missing a lethal threat $\gg$ false alarm), non-terminating appraisal loops are unsafe. Real-time biological systems require termination guarantees, akin to timeouts in engineered controllers (Haselton & Buss, 2000).

These constraints jointly necessitate an architectural solution: a bounded-time controller embedded in the defensive cascade that forces resolution when person-specific thresholds are exceeded—prioritizing rapid termination over veridical accuracy, even if it means inserting faulty but functional explanatory structures.

2. External Ambiguity and the Minimal Viable Agent (MVA)

When a threat is perceived by the autonomic defense systems and determined to be localized and external yet remains ill-defined, the defensive cascade first attempts orientation through heightened perceptual gain. Sensory focus is elevated to increase the odds of informational feedback for threat assessment and orientation. In parallel, covert vigilance is triggered: freezing of positioning or state.

ORCID: 0009-0004-5339-5373

DOI: 10.5281/zenodo.18615682

It is in this moment that a critical new risk emerges. If ambiguity resolves quickly (Amb_t drops), oriented action follows. If ambiguity persists beyond normal momentary tolerance while arousal/load remains low, short-term freeze continues to support evidence accumulation without immediate physiological cost.

However, when ambiguity stays high ($Amb_t \approx 1$) and load escalates ($F_t \geq \lambda *, U_t$ rising), the stalled state becomes unsafe. Inaction and persistence amplify the stress response. The primitive mind recognizes that continued deliberation no longer improves policy quality fast enough to offset the accumulating chemical and cognitive risks. At this saturation point, an evolutionary circuit-breaker is triggered—likely operating on a two-factor stop rule: elapsed unresolved time and stress/load saturation.

The bounded-time controller activates the off-ramp by fabricating a minimal agent representation (MVA)—a fictional but computationally sufficient adversary injected into conscious awareness to provide the structure necessary to re-anchor policy. The MVA provides the minimal structure necessary to re-slope the flat action surface. It supplies only the essential attributes needed to make directional policy computable again:

- Intent (hostile or potentially hostile),
- Location (proximal, directionally specified),
- Trajectory or implied movement.

By collapsing indeterminate threat into an agentive form, the MVA creates a clear vector (e.g., away from the agent, toward cover, or toward inspection) and terminates the unresolved loop. This process is reflexive and pre-deliberative, occurring before full cortical evaluation can resume. Critically, because the neurochemical doping has already occurred, the experience feels emotionally real—even when the agent is purely conjured by the mind.

A child model illustrates this elegantly: when a young child is alone in a dark room and hears an unexpected noise or sensory input, disorientation triggers the threat response. Unable to comprehend the unknown, the child “sees” a monster in the closet or under the bed. The insertion is immediate and vivid.

The purpose of this myth injection is actionability, not truth. The MVA trades a false positive (perceiving an agent where none exists) for guaranteed bounded time-to-decision, consistent with error management theory’s bias toward premature action under lethal asymmetry (Haselton & Buss, 2000). This mechanism reframes hyperactive agency detection (HADD) not as a mere perceptual error, but as a functional implementation pathway for the circuit-breaker under high-load ambiguity (Douglas, 2017).

In practice, the MVA resolves both risks simultaneously:

ORCID: 0009-0004-5339-5373

DOI: 10.5281/zenodo.18615682

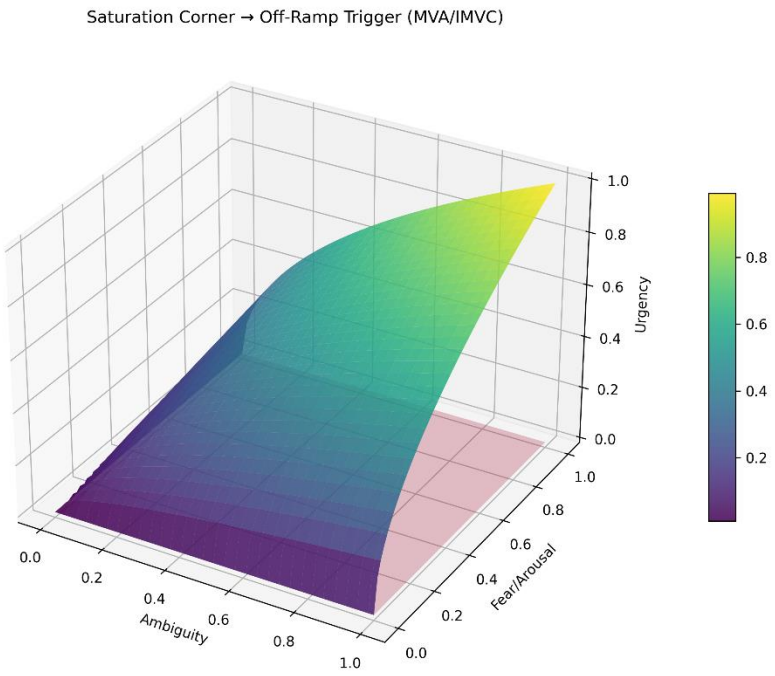


Figure 2. Saturation corner in (Ambiguity, Fear/Arousal, Urgency) space and the off-ramp trigger plane.

The surface shows urgency as a function of ambiguity and fear/arousal. In the low region (dark/blue), evaluation or short freeze is safe. As values approach saturation (yellow corner near 1,1,1), load R_t exceeds λ^*, triggering the bounded-time controller to insert MVA (external) or IMVC (internal) and terminate the unresolved state. The diagram illustrates why indefinite dwell becomes physiologically hazardous and necessitates a circuit-breaker.

3. Internal Ambiguity and the Panic Loop

When a threat is localized internally—arising from interoceptive signals (palpitations, breathlessness, dizziness, chest tightness)—the defensive cascade activates the same parallel processes as external threat: heightened perceptual gain (now interoceptive focus) and sympathetic priming. However, because the threat source is within the body model and provides no external orienting feedback, the ambiguity fork differs critically.

The system attempts interoceptive orientation and appraisal. If the sensation resolves quickly or is benignly interpreted, arousal downregulates. But when the interoceptive signal remains ambiguous and is catastrophically misinterpreted (“this is a heart attack,” “I’m losing control,” “I’m dying”) (Ohst & Tuschen-Caffier, 2020), a positive feedback loop emerges:

ORCID: 0009-0004-5339-5373

DOI: 10.5281/zenodo.18615682

Fear → interoceptive amplification → catastrophic appraisal → escalated arousal → intensified fear.

This is the panic loop: a self-reinforcing cycle where ambiguity is trapped inside the body, with no external anchor to break it. Unlike external ambiguity, continued internal evaluation does not accumulate useful evidence—it only heightens the signal and the perceived existential risk. The dual hazards compound rapidly:

- Chemically: escalating catecholamine dumping sustains overload, risking acute autonomic decompensation.
- Cognitively: catastrophic misinterpretation fractures coherence, promoting entrenchment of maladaptive interoceptive schemas (e.g., persistent fear of bodily sensations).

The stalled state becomes unsafe faster than in external cases because the loop is closed and self-amplifying. Indefinite dwell is intolerable under these constraints.

At the saturation point—when unresolved time exceeds τ^* and load (F_t) surpasses λ^* with internal $Amb_t \approx 1$—the bounded-time controller triggers the internal off-ramp: insertion of an Internal Minimal Viable Cause (IMVC).

The IMVC is a benign, computationally sufficient explanatory structure injected to collapse catastrophic appraisal and restore regulatory policy. It supplies minimal attributes to make downregulation computable:

- Benign origin ("this is adrenaline surge," "this is hyperventilation," "this is a false alarm"),
- Temporary nature ("it will pass," "it peaks and fades"),
- Non-catastrophic outcome ("no harm is occurring").

By reframing the interoceptive anomaly as harmless and transient, the IMVC re-slopes the internal action surface—shifting from escalation to de-escalation—and terminates the loop. This insertion is reflexive and pre-deliberative, often experienced as a sudden "aha" or re-labeling that feels immediately relieving.

The purpose is downregulation, not truth-seeking. The IMVC trades a potential under-detection (missing a real medical event) for bounded termination under asymmetric costs (panic escalation $\gg$ false reassurance in most cases). Clinical interventions deliberately install this off-ramp: cognitive reappraisal, interoceptive exposure, and panic re-labeling techniques (Clark, 1986) are engineered implementations of the IMVC.

In practice, the IMVC resolves both risks simultaneously:

- Chemically, it interrupts the sympathetic positive feedback, allowing parasympathetic dominance and catecholamine clearance.
- Cognitively, it restores minimal coherence to the body model, preventing prolonged fracturing and schema entrenchment.

ORCID: 0009-0004-5339-5373

DOI: 10.5281/zenodo.18615682

Examples include panic-prone individuals who, during an attack, suddenly "realize" the sensations are "just anxiety" or "adrenaline," leading to rapid de-escalation. In VR interoceptive paradigms, benign priming (pre-provided explanations) accelerates IMVC adoption and dampens peak arousal.

This internal counterpart to the MVA completes the architecture: both mechanisms insert minimal viable structure to bound dwell time in high-load ambiguity, but tailored to the threat's locus—external (agentive) or internal (causal).

4. Formal Model: The Bounded-Time Controller

To formalize the architecture, we define the bounded-time controller as a dual-threshold stop rule embedded within the defensive cascade. The controller monitors three normalized state variables in real time:

- $F_t \in [0,1]$: fear/arousal level (proxied by sympathetic activation, catecholamine surge, subjective intensity)
- $U_t \in [0,1]$: urgency (perceived time pressure, proximity of threat, escalating risk)
- $Amb_t \in [0,1]$: ambiguity (lack of structure / informational resolution in threat representation)

A composite load metric is computed as:

$R_t = \alpha F_t + \beta U_t + \gamma Amb_t$ (with $\alpha, \beta, \gamma > 0$; weights tuned to person/context)

The controller enforces the following stop condition:
$t_{unresolved} \geq \tau^* \; AND \; R(t) \geq \lambda^* \quad AND \quad Amb(t) \approx 1$

where:

- τ^* = person- and context-specific time budget (maximum tolerable unresolved dwell time before physiological/cognitive risks become intolerable)
- λ^* = load threshold (saturation point where continued appraisal costs exceed benefits)

When the condition is met, the controller forces termination via off-ramp insertion:

- If threat is external → insert Minimal Viable Agent (MVA)
- If threat is internal → insert Internal Minimal Viable Cause (IMVC)

Pseudocode representation (control flow):
FearTrigger →

ORCID: 0009-0004-5339-5373

DOI: 10.5281/zenodo.18615682

```
 ParallelActivation:
  SituationalAwareness↑ (perceptual gain, attention narrowing)
  MotorMetabolicPriming↑ (sympathetic surge, catecholamine release)

OrientationAttempt →
 if ThreatOrientation == INTERNAL:
  if t_unresolved ≥ τ* and R(t) ≥ λ* and Amb(t) ≈ 1:
   insert IMVC
   policy ← downregulate (reappraisal, parasympathetic shift)
  else:
   continue interoceptive monitoring & appraisal
 else:  # EXTERNAL
  if Amb(t) low:
   oriented action
  elif Amb(t) high and R(t) low:
   short-term freeze (covert vigilance)
  elif Amb(t) high and t_unresolved ≥ τ* and R(t) ≥ λ*:
   insert MVA
   policy ← orient / avoid / flee / inspect
```

Key properties of the controller:

- Asymmetric bias: The stop rule favors false positives (unnecessary insertion) over false negatives (prolonged stall leading to decompensation), consistent with error management theory (Haselton & Buss, 2000).
- Parameterization: τ^* and λ^* are individual differences variables, influenced by trait intolerance of uncertainty (IU), prior trauma, chronic stress load, or context (e.g., safety cues lower λ^*). High IU individuals exhibit lower effective $\frac{\tau^*}{\lambda^*}$ (earlier firing).
- Modality sensitivity: Visual ambiguity (e.g., low-light occlusion) drives $Amb_t \rightarrow 1$ more rapidly than auditory (requires pattern accumulation), leading to earlier triggering at matched R_t.
- Termination guarantee: The controller ensures bounded dwell time in high-load ambiguity, capping chemical doping and cognitive fracturing even at the cost of veridicality.

Estimation of parameters:

- T^*: latency-to-MVA/IMVC report in VR ambiguity or interoceptive tasks
- Λ^*: arousal proxies (skin conductance response peaks, heart-rate variability suppression, pupil dilation)
- Individual mapping: logistic/survival modeling of threshold crossing in controlled escalation paradigms

ORCID: 0009-0004-5339-5373

DOI: 10.5281/zenodo.18615682

This formalization unifies the external (MVA) and internal (IMVC) off-ramps under a single controller, providing a testable, predictive framework. It reframes agency/cause hallucination as a safety-critical timing solution rather than bias.

5. Modality Dynamics

Ambiguity accumulation and threshold crossing are not modality-agnostic; they vary systematically by sensory channel due to differences in information accrual rates and perceptual completion mechanisms.

- Visual ambiguity (e.g., low-light occlusion, partial silhouettes, distant or degraded objects) drives rapid object completion via top-down processes in ventral stream and early visual cortex. Amb_t saturates quickly because the system fills gaps aggressively to resolve structure. At matched arousal/load (R_t), visual occlusion often triggers earlier off-ramp firing (shorter effective $t_{unresolved}$ to reach τ^*).
- Auditory ambiguity (e.g., isolated noises, footsteps, spatialized sounds without clear pattern) typically requires temporal accumulation—repetition, rhythm, or escalating sequences—to drive Amb_t toward saturation. Single or sporadic cues permit longer evidence gathering before thresholds cross, delaying MVA insertion relative to visual equivalents under similar R_t.
- Cross-modal integration acts as an accelerator. Congruent pairings (e.g., faint silhouette + synchronized footsteps behind a corner) collapse Amb_t faster than unimodal cues, reducing τ^* and advancing off-ramp timing. Incongruent or mismatched pairings (e.g., visual occlusion without auditory correlate) may prolong freeze or delay insertion, as the system struggles to reconcile channels.

These dynamics emerge directly from the controller's stop rule: Amb_t is a key input to R_t, and modality influences how rapidly Amb_t approaches 1. Visual channels, with their bias toward rapid gestalt completion, hit the saturation corner sooner; auditory channels accumulate more gradually but can approximate visual timing when patterns build urgency (U_t rises with repetition).

Predictions follow:

- In VR low-light occlusion tasks, MVA reports and physiological arousal proxies (SCR peaks, HRV suppression) occur earlier than in auditory-pattern tasks at equated baseline ambiguity and arousal.
- Patterned auditory sequences (e.g., escalating footsteps) yield shorter freeze durations and earlier MVA insertion than isolated sounds, as temporal accumulation drives Amb_t faster.
- Congruent cross-modal cues (visual + auditory match) drastically reduce time-to-off-ramp compared to unimodal or incongruent conditions.
  Individual differences modulate these effects: high intolerance of uncertainty (IU) individuals show accelerated threshold crossing across modalities due to amplified negative appraisal,

ORCID: 0009-0004-5339-5373

DOI: 10.5281/zenodo.18615682

while modality preference (e.g., visual-dominant vs. auditory-dominant processing styles) may shift effective τ^* and λ^*.These modality-specific predictions provide strong empirical leverage for testing the bounded-time controller in controlled VR/interoceptive paradigms (see Appendix B).

6. Individual Differences as Parameter Variation

The bounded-time controller's thresholds—τ* (time budget) and λ^* (load threshold)—are not fixed universals but vary systematically across individuals, reflecting trait-level differences in threat processing, uncertainty tolerance, and compensatory strategies. These variations explain observed patterns in freeze duration, false-positive rates, and reliance on off-ramps without requiring new mechanisms—only shifts in parameterization.

These thresholds may operate like a "fuse" in neurobiological terms: stress chemicals (e.g., catecholamines) accumulate during unresolved ambiguity until they saturate receptors or exceed a critical level, triggering rapid off-ramp insertion. This mirrors fatigue, where adenosine builds linearly during wakefulness but escalates nonlinearly once it overwhelms receptor thresholds (caffeine blocks adenosine A1/A2A receptors, delaying this; PMC reviews on adenosine/caffeine). In fear circuits, catecholamine buildup in amygdala-PAG pathways could similarly saturate adrenergic receptors, firing the MVA/IMVC "fuse" to prevent decompensation (e.g., panic escalation or Takotsubo). Clinical data shows intolerance of uncertainty (IU) lowers these fuse thresholds: high-IU individuals exhibit heightened prefrontal/arousal responses during uncertain threats, earlier panic loops, and faster symptom escalation (PMC on IU in GAD/panic; Frontiers on IU heightening negative states).

We propose two primary archetypes, modeled as opposing shifts in effective τ^* and λ^*

- **Type M** (internal generators / "minimalists")
  Lower effective τ^* and λ^* → earlier off-ramp firing (shorter fuse).
  Characteristics:
  - Shorter freeze durations under ambiguity (quicker to insert MVA/IMVC).
  - Higher false-positive rate (more frequent hallucinations of agents/causes).
  - Greater reliance on internal off-ramps (IMVC) or self-generated explanatory structures.
  - Aligns with hyperactive agency detection (HADD) tendencies and lower tolerance for unresolved states.

  These individuals cross thresholds sooner, trading accuracy for rapid termination and physiological safety.

- **Type P** (external programmatics / "procedural")

ORCID: 0009-0004-5339-5373

DOI: 10.5281/zenodo.18615682

Higher effective τ^* and λ^*→ longer tolerance for unresolved ambiguity (longer fuse).
Characteristics:

- Extended freeze or deliberation when scaffolds are present.
- Greater reliance on externalized off-ramps (procedures, checklists, institutions, rituals, social consensus) to provide structure.
- Longer dwell in ambiguity if external anchors are absent or threatened, increasing risk exposure.
- Aligns with compensatory control strategies (Kay et al., 2009) and higher need for external order when personal control is low.

Intolerance of Uncertainty (IU) as a modulator

High IU amplifies negative appraisal of ambiguity, effectively lowering τ^* and λ^* across both types by increasing prefrontal engagement during uncertain threat (Morriss et al., 2022). High-IU individuals exhibit:

- Earlier threshold crossing.
- Heightened physiological responding (SCR, HR/HRV changes).
- Greater dependence on off-ramps (internal re-labeling or external scaffolds) under threatened control.

These parameter shifts are estimable in controlled tasks:

- Latency-to-MVA/IMVC report and arousal proxies (SCR peaks, HRV suppression) during ambiguity escalation.
- Survival/logistic modeling of threshold crossing to derive person-specific τ^* and λ^* .
- Narrative-agency or uncertainty vignettes to classify Type M vs Type P.

No novel circuitry is required—individual differences emerge from tuning the existing controller. Type M favors rapid, self-generated resolution (more false positives, less freeze risk); Type P favors scaffolded, accurate resolution (longer freeze when scaffolds fail). Both patterns are adaptive trade-offs under the same evolutionary constraints.

Predictions:

- High-IU participants show earlier MVA/IMVC insertion and stronger cross-modal congruence effects.
- Type M individuals report more agentive hallucinations in low-load ambiguity; Type P show longer freeze in scaffold-absent conditions.
- Compensatory control threats (e.g., loss of personal agency) shift Type P toward earlier external off-ramp reliance.

This parameterization unifies trait-level findings under the bounded-time framework, making individual differences testable and clinically relevant.

ORCID: 0009-0004-5339-5373

DOI: 10.5281/zenodo.18615682

7. Empirical Predictions

The bounded-time controller and its off-ramp insertions (MVA and IMVC) generate a set of mechanistic, falsifiable predictions that span modality effects, arousal modulation, individual differences, and intervention outcomes. These predictions follow directly from the model's core elements: the dual-threshold stop rule, the action-surface flattening under ambiguity, the two-factor trigger (time + load saturation), and the tailored resolution of chemical/cognitive risks.

Arousal and load modulation

- Increasing arousal/load (F_t or R_t) while holding ambiguity constant must reduce time-to-off-ramp insertion (shorter $t_{unresolved}$ to reach τ^*). Higher baseline sympathetic tone (e.g., via threat-of-shock or time pressure) advances MVA/IMVC firing even at fixed visual/auditory ambiguity.
- In high-load conditions, freeze duration shortens and MVA reports increase; low-load conditions permit longer freeze without insertion.

Modality-specific dynamics

- Visual ambiguity (low-light occlusion, degraded silhouettes) triggers earlier MVA insertion than auditory ambiguity (isolated noises) at matched arousal/load, due to faster Amb_t saturation via object completion.
- Patterned auditory sequences (repetition, escalating footsteps) approximate visual timing, yielding earlier MVA firing as temporal accumulation drives $Amb_t \rightarrow 1$.
- Cross-modal congruence (e.g., synchronized silhouette + footsteps) drastically reduces τ^* and accelerates off-ramp; incongruence prolongs freeze or delays insertion.

Individual differences

- High intolerance of uncertainty (IU) participants exhibit earlier threshold crossing (lower effective $\frac{\tau^*}{\lambda^*}$), higher MVA/IMVC report rates, and stronger physiological responses (SCR peaks, HRV suppression) during ambiguity escalation.
- Type M individuals show shorter freeze durations, more frequent MVA/IMVC insertions, and higher false-positive rates in low-to-moderate load ambiguity tasks.
- Type P individuals exhibit longer freeze when external scaffolds are present, but greater risk exposure (prolonged dwell) when scaffolds are absent or threatened. Compensatory control threats shift Type P toward earlier reliance on external off-ramps.

Intervention and priming effects

- Agentive primes (e.g., pre-task exposure to hostile-figure images) shorten freeze and increase MVA reports in external ambiguity tasks.

ORCID: 0009-0004-5339-5373

DOI: 10.5281/zenodo.18615682

- Benign interoceptive primes (e.g., "this is adrenaline" scripts) selectively reduce catastrophic misinterpretations, lower peak arousal, and accelerate IMVC adoption in panic simulations—faster than neutral or external primes.
- Clinical-style re-labeling training (e.g., interoceptive exposure with benign explanations) raises effective λ^* over sessions, extending safe dwell time and reducing panic frequency.

Threshold mapping and estimation

- Participant-specific τ^* and λ^* can be estimated via latency-to-MVA/IMVC report, arousal proxies (SCR, HR/HRV, pupilometrics), and survival/logistic modeling during gradual ambiguity escalation.
- High-IU correlates with lower estimated thresholds; Type M/P classification via narrative vignettes or freeze-duration patterns in scaffold-present/absent conditions.

These predictions are mechanistic consequences of the controller's stop rule and off-ramp logic. They are testable in VR ambiguity paradigms (Appendix B), interoceptive simulations, and longitudinal uncertainty tasks. Positive findings would support the framework as a unified account of threat-induced hallucinations as timing solutions; null or opposing patterns would falsify key aspects (e.g., modality effects on Amb_t accumulation or IU as threshold modulator).

8. Implications

The bounded-time controller framework has practical implications across clinical practice, resilience training, high-stakes operational environments, and broader cultural/governance systems. These arise from the model's emphasis on bounding dwell time in high-load ambiguity rather than eliminating hallucinations or biases entirely. The goal is to manage thresholds and off-ramps to minimize physiological and cognitive exposure while preserving actionability.

Clinical and resilience applications

- Panic disorder and related conditions can be reframed as failures of timely IMVC insertion: catastrophic misinterpretation sustains the panic loop until load saturates. Interventions that deliberately install benign re-labeling (e.g., "this is adrenaline," "this will peak and pass") strengthen the internal off-ramp, effectively raising λ^* and extending safe dwell time. Interoceptive exposure protocols train tolerance for ambiguity without immediate resolution, reducing IU-driven threshold lowering.
- Resilience training should target threshold calibration rather than suppression of agentic attributions. Techniques that extend τ^* safely (e.g., mindfulness for arousal regulation, cognitive defusion for coherence maintenance) or provide agent-free external scaffolds (checklists, breathing scripts) reduce reliance on reflexive MVA/IMVC while preserving

ORCID: 0009-0004-5339-5373

DOI: 10.5281/zenodo.18615682

termination guarantees. High-IU individuals may benefit most from IU-specific interventions (e.g., uncertainty exposure) to normalize threshold dynamics.

High-stakes teams and operational design

- In environments where prolonged ambiguity under load is common (e.g., military, emergency response, aviation, trading floors), the model predicts increased false-positive agentification or catastrophic internal appraisal when thresholds are exceeded. Teams should implement agent-free external off-ramps—standardized checklists, escalation protocols, predefined decision trees, or "timeout" rituals—that deliver directional policy without requiring hallucinated adversaries. This reduces physiological exposure (catecholamine dumping) and cognitive fracturing while maintaining bounded decision time.
- Training should simulate controlled ambiguity escalation (e.g., VR scenarios) to map individual $\frac{\tau^*}{\lambda^*}$ and train scaffold reliance, particularly for Type P individuals who freeze longer without external structure.

Cultural and governance systems

- Superstitions, rituals, and certain belief systems can be understood as externalized off-ramps that supply structure under prolonged or chronic ambiguity (e.g., societal uncertainty, existential threats). They serve the same timing imperative as the MVA—providing minimal explanatory anchors to terminate unresolved states when personal control is threatened.
- Governance and institutional design may benefit from recognizing this mechanism: policies that reduce systemic ambiguity (clear rules, predictable procedures) lower collective load and decrease maladaptive off-ramp reliance (e.g., conspiratorial thinking in low-urgency, high-uncertainty regimes). Conversely, environments that chronically threaten coherence may amplify externalized insertions.

Boundaries and cautions

The framework does not claim that all agency attributions or causal explanations are false or maladaptive—short freeze remains adaptive for vigilance, and many insertions are veridical. It describes a default termination heuristic under insufficient evidence and elevated load. Over-reliance on reflexive off-ramps (e.g., frequent false positives) may indicate lowered thresholds due to chronic stress or high IU, warranting clinical attention.

By focusing on threshold management and scaffolded resolution, the model offers actionable paths for reducing harm from prolonged ambiguity while respecting the adaptive necessity of the circuit-breaker.

ORCID: 0009-0004-5339-5373

DOI: 10.5281/zenodo.18615682

9. Limits

This framework describes a default, reflexive termination mechanism under insufficient evidence and elevated load—not a comprehensive explanation for all agency or causal attributions. Many attributions are veridical or arise from deliberative processes with ample evidence; the MVA/IMVC primitives activate specifically when the bounded-time controller detects high-risk saturation.

Neural substrates cited (e.g., amygdala-PAG projections) support state transitions within the defensive cascade, not a single locus of control or a dedicated "hallucination center." The model does not localize the controller to one brain region but posits a distributed, threshold-based process akin to receptor saturation in stress circuits.

Short-term freeze remains adaptive for covert vigilance and action preparation; prolongation under unresolved ambiguity is the hazard zone, not freeze itself. The account does not imply that all superstitions, rituals, or broader explanatory beliefs stem directly from this mechanism—those may represent extensions or cultural elaborations (e.g., in low-urgency regimes).

Physiological risks (e.g., Takotsubo, arrhythmias) are framed as selection pressures favoring the circuit-breaker, not direct causal outcomes of every freeze episode. Empirical support for threat-potentiated HADD is mixed in lab settings, suggesting context/load dependence; the model predicts expression under high $\frac{F_t}{U_t}$, not constant bias.

The predictions emphasize acute, controlled tasks (VR, interoceptive); real-world ambiguity often confounds with chronic factors, requiring ecological validation. Finally, the model is functional and timing-driven—it does not address conscious awareness of insertions or long-term learning effects.

ORCID: 0009-0004-5339-5373

DOI: 10.5281/zenodo.18615682

10. Conclusion

Agency attribution under threat is typically framed as a cognitive bias or error-prone heuristic. We propose instead that it is a bounded-time control solution: when fear, urgency, and ambiguity saturate the defensive cascade, the organism cannot afford indefinite evaluation amid escalating chemical and cognitive risks. The system inserts the smallest explanatory structure sufficient to restore an action surface and terminate the stall—the Minimal Viable Agent for external threats, the Internal Minimal Viable Cause for internal loops.

These primitives are faulty-but-functional: they trade veridicality for guaranteed resolution under asymmetric costs and physiological constraints. By unifying error management theory, hyperactive agency detection, compensatory control, meaning maintenance, and intolerance of uncertainty within a formal controller, this framework reframes "hallucinated" agents and causes not as failures first, but as safety-critical mechanisms. Termination trumps certainty when survival is at stake.

Future work should test the model's predictions in VR and interoceptive paradigms, estimate person-specific thresholds, and explore interventions that calibrate the controller for resilience. Ultimately, understanding this architecture illuminates why humans so readily fabricate meaning in the face of the unknown—and how to harness it without undue harm

ORCID: 0009-0004-5339-5373

DOI: 10.5281/zenodo.18615682

11. References

Clark, D. M. (1986). A cognitive approach to panic. Behaviour Research and Therapy, 24(4), 461–470. https://doi.org/10.1016/0005-7967(86)90011-2

Douglas, K. M. (2017). Hypersensitive agency detection. In V. Zeigler-Hill & T. K. Shackelford (Eds.), Encyclopedia of personality and individual differences (pp. 1–2).

George, D. T., Ameli, R., & Koob, G. F. (2019). Periaqueductal gray sheds light on dark areas of psychopathology. Trends in Neurosciences, 42(5), 349–360. https://doi.org/10.1016/j.tins.2019.03.004

Ghadri, J. R., et al. (2018). International expert consensus document on Takotsubo syndrome (Part I). European Heart Journal, 39(22), 2032–2046. https://doi.org/10.1093/eurheartj/ehy076

Harvard Health Publishing. (2023). Broken-heart syndrome (takotsubo cardiomyopathy). Harvard Medical School. https://www.health.harvard.edu/heart-health/takotsubo-cardiomyopathy-broken-heart-syndrome

Haselton, M. G., & Buss, D. M. (2000). Error management theory: A new perspective on biases in cross-sex mind reading. Journal of Personality and Social Psychology, 78(1), 81–91. https://doi.org/10.1037/0022-3514.78.1.81

Heine, S. J., Proulx, T., & Vohs, K. D. (2006). The meaning maintenance model: On the coherence of social motivations. Personality and Social Psychology Review, 10(2), 88–110. https://doi.org/10.1207/s15327957pspr1002_1

Kay, A. C., Whitson, J. A., Gaucher, D., & Galinsky, A. D. (2009). Compensatory control: Achieving order through the mind, our institutions, and the heavens. Current Directions in Psychological Science, 18(5), 264–268. https://doi.org/10.1111/j.1467-8721.2009.01649.x

Maddox, S. A., Hartmann, J., Lowery-Gionta, E. G., & Tovote, P. (2019). Deconstructing the Gestalt: Mechanisms of fear, threat, and trauma memory encoding. Neuron, 102(1), 60–74. https://doi.org/10.1016/j.neuron.2019.03.017

Morriss, J., Bell, T., Biagi, N., Johnstone, T., & van Reekum, C. M. (2022). Intolerance of uncertainty is associated with heightened responding in the prefrontal cortex during cue-signalled uncertainty of threat. Cognitive, Affective, & Behavioral Neuroscience, 22(1), 88–98. https://doi.org/10.3758/s13415-021-00932-7

ORCID: 0009-0004-5339-5373

DOI: 10.5281/zenodo.18615682

Ohst, B., & Tuschen-Caffier, B. (2020). Are catastrophic misinterpretations of bodily sensations typical for patients with panic disorder? An experimental study... Cognitive Therapy and Research, 44(6), 1118–1133. https://doi.org/10.1007/s10608-020-10141-0

Roelofs, K. (2017). Freeze for action: Neurobiological mechanisms in animal and human freezing. Philosophical Transactions of the Royal Society B: Biological Sciences, 372(1718), Article 20160206. https://doi.org/10.1098/rstb.2016.0206

Weis, S., et al. (2022). A 7-Tesla MRI study of the periaqueductal gray: resting state and task activation under threat. NeuroImage, 248, 118860. https://doi.org/10.1016/j.neuroimage.2021.118860 (

Wittstein, I. S., Thiemann, D. R., Lima, J. A. C., Baughman, K. L., Schulman, S. P., Gerstenblith, G., Wu, K. C., Rade, J. J., Bivalacqua, T. J., & Champion, H. C. (2005). Neurohumoral features of myocardial stunning due to sudden emotional stress. New England Journal of Medicine, 352(6), 539–548. https://doi.org/10.1056/NEJMoa043046

ORCID: 0009-0004-5339-5373

DOI: 10.5281/zenodo.18615682

Appendix A: Logic and Math of the Off-Ramp Controller

This appendix expands on the formal model in §6, providing pseudocode, the composite risk equation, and practical estimation guidance. It is self-contained for readers who want to replicate or test the controller.

A.1 Control-Flow Pseudocode (full version from §6)

```
FearTrigger →
 ParallelActivation:
  SituationalAwareness↑ (perceptual gain, attention narrowing)
  MotorMetabolicPriming↑ (sympathetic surge, catecholamine release)

OrientationAttempt →
 if ThreatOrientation == INTERNAL:
  if t_unresolved ≥ τ* and R(t) ≥ λ* and Amb(t) ≈ 1:
   insert IMVC
   policy ← downregulate (reappraisal, parasympathetic shift, acceptance)
  else:
   continue interoceptive monitoring & appraisal (loop with amplification risk)
 else:  # EXTERNAL
  if Amb(t) low:
   oriented action (fight/flight/approach)
  elif Amb(t) high and R(t) low:
   short-term freeze (covert vigilance, evidence accumulation)
  elif Amb(t) high and t_unresolved ≥ τ* and R(t) ≥ λ*:
   insert MVA
   policy ← orient / avoid / flee / inspect
```

A.2 Composite Risk and Stop Rule

Define state variables (normalized [0,1]):

- F_t: fear/arousal
- U_t: urgency
- Amb_t: ambiguity

Composite load:

$R_t = \alpha F_t + \beta U_t + \gamma Amb_t$ $(\alpha, \beta, \gamma > 0$; e.g., equal weights $\alpha = \beta = \gamma = 1/3$ for simplicity, or empirically tuned)

Stop and insert off-ramp when:

ORCID: 0009-0004-5339-5373

DOI: 10.5281/zenodo.18615682

- $t_{unresolved} \geq \tau^*$ (time budget exceeded)
- $R_t \geq \lambda^*$ (load threshold exceeded)
- $Amb_t \approx 1$ (high ambiguity saturation)

External → MVA insertion
Internal → IMVC insertion

- T^*: Time budget — estimate via latency-to-MVA/IMVC self-report or behavioral shift in VR ambiguity tasks (e.g., time from threat onset to "I saw someone" report).
- Λ^*: Load threshold — estimate via arousal proxies during controlled ambiguity escalation: skin conductance response (SCR) peaks, heart-rate variability (HRV) suppression, pupil dilation, or subjective arousal ratings.
- Amb_t: Manipulate via occlusion/low-light (visual), repetition/pattern absence (auditory), or interoceptive uncertainty (e.g., amplified heartbeat without explanation).
- Individual differences: Use IU scales (e.g., Intolerance of Uncertainty Scale) to predict lower $\frac{\tau^*}{\lambda^*}$; narrative-agency tasks or freeze-duration patterns in scaffold-present/absent conditions for Type M/P classification.
- Modeling: Fit participant-specific thresholds using logistic regression or survival analysis on latency data from repeated trials.

These methods allow direct empirical testing and individual calibration.

ORCID: 0009-0004-5339-5373

DOI: 10.5281/zenodo.18615682

Appendix B: Experimental Paradigms for VR Ambiguity Tasks

This appendix outlines paradigms to test threshold dynamics (τ^*, λ^*), modality effects, and off-ramp activation (MVA vs. IMVC). Each includes stimulus construction, manipulation, and outputs.

B.1 External Ambiguity VR: Low-Light Visual Occlusion

***Goal*:** Probe rapid object-completion biases and latency-to-MVA under ambiguous external cues.

Stimulus: Dimly lit virtual corridor with occluding geometry (corners, pillars); visual resolution degraded to modulate Amb_{ex_t}.

***Manipulation*:** Ambiguous silhouettes at varying distances/resolutions; arousal via threat-of-shock or time pressure.

Measurements: Latency-to-freeze, latency-to-MVA report ("I saw someone"), HR/HRV suppression, SCR peaks.

Prediction: Higher arousal (R(t) ≥ λ*) produces earlier MVA insertion at constant visual ambiguity.

B.2 External Ambiguity VR: Patterned Auditory Sequences

Goal: Test auditory accumulation dynamics.

Stimulus: Spatialized audio (single noises, repeated, escalating patterns like footsteps).

Manipulation: Temporal spacing (Δ_t), repetition count, pattern regularity to induce accumulation.

Measurements: Freeze duration, orienting responses, MVA reports, SCR/HRV traces.

Prediction: Patterned sounds yield earlier MVA than isolated sounds due to temporal buildup of Amb(t).

B.3 Mixed Modality: Visual + Auditory Ambiguity

Goal: Test cross-modal integration under threat.

Stimulus: Synchrony between ambiguous visuals and auditory cues disrupted systematically.

Manipulation: Congruent (footsteps + faint silhouette) vs. incongruent pairings.

Measurements: MVA trigger rates, congruency effects on τ*, mismatch effects on λ*.

Prediction: Congruence drastically reduces τ* (off-ramp fires sooner).

ORCID: 0009-0004-5339-5373

DOI: 10.5281/zenodo.18615682

B.4 Internal Threat: Interoceptive VR Layer

Goal: Isolate IMVC triggering.

Stimulus: Interoceptive perturbations (breath resistance, heartbeat amplification, mild CO_2 boost) in VR context.

Manipulation: Benign explanations (IMVC primes) vs. no explanation.

Measurements: Catastrophic misinterpretation rate, MVA frequency (low expected), IMVC adoption latency.

Prediction: Primes reduce misinterpretations and dampen arousal faster.

B.5 Threshold Mapping Protocol

Goal: Estimate individual τ^* and λ^*.

Procedure: Gradually escalate ambiguity while recording latency-to-freeze/MVA/IMVC, physiological load. Fit thresholds via logistic/survival modeling.

Outcome: Person-specific maps for Type M/P classification and IU correlation.

These paradigms are ready for implementation and directly test the model's predictions (§9).

ORCID: 0009-0004-5339-5373

DOI: 10.5281/zenodo.18615682

Appendix C — Disclaimer and Scope of Work

This manuscript presents a theoretical framework developed from interdisciplinary synthesis across cognitive science, computational modeling, psychophysiology, and evolutionary theory. It is not a clinical document and is not authored by a medical doctor (MD) or licensed clinician. The mechanisms, models, and interpretations described herein are conceptual and exploratory, offered for academic discourse and hypothesis generation rather than for medical diagnosis, treatment, or clinical decision-making.

While the framework integrates publicly available research on autonomic physiology, threat responses, and stress-related cardiac phenomena—including freeze-fight dynamics and stress-induced cardiomyopathy (Takotsubo)—all physiological references are interpretive summaries of existing literature, not original clinical findings or medical recommendations. Readers should not construe any content in this manuscript as clinical advice, nor should it substitute for consultation with qualified medical or mental-health professionals.

Any potential clinical implications discussed in Section *Clinical and Translational Applications* are speculative and intended solely to illustrate how the proposed bounded-time controller architecture may relate to known psychological or physiological patterns. These should be interpreted as conceptual bridges, not as validated interventions.

The models (MVA, IMVC, τ^*, λ^*, R_t) are functional abstractions, not direct neuroanatomical claims. They should not be interpreted as literal descriptions of specific brain circuits or as statements about individual diagnoses. All experimental paradigms described in Appendix B are intended for controlled research contexts and should undergo standard ethical review before implementation.
In summary:

- This is a non-clinical, theoretical paper.
- No claims are made about diagnosis, prognosis, or treatment.
- All physiological references are derivative of published studies and are not medical guidance.
- Any real-world applications should involve appropriate domain experts and ethical oversight.

ORCID: 0009-0004-5339-5373

DOI: 10.5281/zenodo.18615682

Appendix D: License and Usage Details

ORCID: 0009-0004-5339-5373
DOI: 10.5281/zenodo.18615682