Gradient Encoding in Human Oversight Systems (GEHOS)

Version: 3.0 Status: Research Umbrella (Revised) Author: Charlie A. Johnson

---

0. Program-Level Thesis

This research program investigates whether continuous bounded encoding of system state alters supervisory dynamics in multi-agent and autonomous systems relative to threshold-based discrete alerting.

The core claim is conditional:

Under defined boundary conditions, continuous encoding changes interruption patterns, trust calibration, escalation dynamics, and coordination friction in measurable, domain-sensitive ways.

This program explicitly tests both benefits and failure modes.

---

1. Scope

Included:

- Human oversight of semi- or fully-autonomous systems
- Multi-agent AI supervision
- Decision-support trust calibration
- Distributed team coordination under bounded tasks

Excluded:

- Architectural / environmental embedding
- Smart cities / public infrastructure
- Consumer UX aesthetics

- Broad societal futurism

---

2. Core Construct Decomposition

Gradient encoding is decomposed into four independent dimensions. Each experiment must declare which values are active on each dimension.

2.1 Signal Topology (S)

Mathematical structure of encoding:

- S_d: Discrete binary — $S \in \{0,1\}$
- S_c: Categorical multi-level — $S \in \{0, 1, \ldots, n\}$
- S_b: Continuous bounded — $S \in [0,1]$
- S_u: Continuous unbounded — $S \in \mathbb{R}$

Primary research contrast: S_b vs. S_d.

Critical methodological note: S_c serves as the information-content control condition. To distinguish "gradient encoding helps because it is continuous" from "more information helps regardless of encoding," all studies contrasting S_b with S_d must include an S_c condition matched on information entropy. The S_c condition should use $k = \lceil 1/\text{JND} \rceil$ levels, where JND is the just-noticeable difference for the perceptual channel in use (see §2.3), ensuring that S_c and S_b are perceptually equivalent in discriminable resolution but structurally different in encoding continuity.

2.2 Perceptual Channel (P)

Attentional routing of the encoded signal.

Primary distinction:

- P_f: Focal channel — requires explicit attention allocation; engages central foveal processing.
- P_p: Peripheral channel — detectable without focal shift; engages extrafoveal processing.

Required parameterization for any P_p condition:

- P_p(ε): Eccentricity in degrees of visual angle from fixation point.

- P_p(m): Modality — visual-luminance, visual-chromatic, auditory, haptic.
- P_p(f): Temporal update frequency in Hz.

Rationale: Peripheral chromatic sensitivity degrades rapidly beyond ~15° eccentricity while luminance contrast sensitivity degrades more slowly (Hansen et al., 2009). Peripheral vision is preferentially sensitive to temporal change in certain frequency bands. These parameters are not interchangeable and must be specified per study.

Default experimental parameters (to be validated in Pillar 4):

- ε = 20° (moderate periphery, ecologically plausible for secondary monitor)
- m = visual-luminance (most robust peripheral pathway)
- f = 0.5 Hz (slow update, below flicker threshold, emphasizes state over transient)

2.3 Temporal Dynamics (T)

Defines the update and trajectory properties of the encoded signal.

- T_rate: Update frequency (Hz). Determines whether signal engages sustained vs. transient perceptual channels.
- T_smooth: Smoothing window applied to raw R(t) before display (seconds). Controls jitter visibility.
- T_history: Duration of visible trajectory (seconds). Controls whether operator sees current state only vs. recent trend.

Interaction with P: At peripheral eccentricities > 15°, temporal frequencies between 5–15 Hz produce strong apparent motion percepts that can dominate magnitude encoding (McKee & Nakayama, 1984). Studies must ensure T_rate does not inadvertently shift the perceptual mechanism from magnitude judgment to motion detection.

Design constraint: T_history > 0 enables rate-of-change perception (I_3) without requiring derivative computation in the display. T_history = 0 restricts encoding to magnitude only (I_1, I_2). Each study must declare T_history and justify.

2.4 Information Structure (I)

What properties of system state are encoded in the gradient signal.

Formally, let R(t) be a system risk function (defined in §3). The display may encode any subset of:

- I_1: Magnitude — R(t) directly. Current position in state space.
- I_2: Proximity-to-boundary — T - R(t), where T is the intervention threshold. Distance to action-required state.
- I_3: Rate-of-change — dR/dt (first derivative). Velocity of state evolution.
- I_4: Accumulated uncertainty — $H(t) = -\int p(x,t) \log p(x,t)\, dx$ integrated over a rolling window. Captures whether the system is becoming less predictable.
- I_5: Cross-agent coherence — $C(t) = g(\{R_i(t)\}_{i=1}^n)$, a function of the set of per-agent states. Captures divergence structure (see §3.2).

Orthogonality: I_1 and I_2 are related by a simple transformation when T is fixed. I_3 can be derived from I_1 with T_history > 0. I_4 and I_5 are independent of I_1–I_3 and of each other.

Each experiment must declare which I-dimensions are active and how they map to visual parameters. A single visual gradient (e.g., color hue) can encode at most one I-dimension without ambiguity. Encoding multiple I-dimensions requires either multiple visual channels (hue + luminance + spatial extent) or multiple display elements.

---

3. Risk Function Specification

3.1 Single-System Risk

Let $X(t) \in \mathbb{R}^d$ be the system state vector at time t.

Define:

$R(t) = f(X(t)) \in [0,1]$

where f is the risk mapping. This program does not propose a universal f but constrains the class of admissible risk functions:

Smoothness assumption: f must be Lipschitz continuous with constant K_f such that $|f(X_1) - f(X_2)| \leq K_f \|X_1 - X_2\|$ for all X_1, X_2 in the domain. This ensures that small state changes produce bounded risk changes, which is the condition under which continuous gradient encoding faithfully represents the underlying risk dynamics.

Failure of smoothness assumption: When f has discontinuities (phase transitions, cascade thresholds, emergent multi-agent failure modes), continuous gradient encoding misrepresents risk by implying gradual change where change is actually abrupt. This is a core boundary condition (see §5, Condition 2).

Each experiment must specify f concretely for its simulation environment and report K_f empirically.

3.2 Multi-Agent Risk Aggregation

For a system of n agents, each with state $X_i(t)$, we distinguish three aggregation structures:

A. Independent risk (max-aggregation): $R(t) = \max\{R_i(t)\}_{i=1}^{n}$

Appropriate when any single agent failure constitutes system failure. Scalar gradient is a lossy but conservative representation.

B. Cumulative risk (mean-aggregation): $R(t) = (1/n)\ \Sigma\ R_i(t)$

Appropriate when risk is additive and no single agent is critical. Scalar gradient is a natural representation.

C. Interaction risk (coherence-dependent): $R(t) = h(\{R_i(t)\}, \{R_{i\,j}\,(t)\})$

where $R_{i\,j}$ captures pairwise agent interaction risk (e.g., policy divergence, conflicting actions). This cannot be represented by a single scalar without information loss. Pillar 3 (cross-agent coherence) addresses this case specifically.

Each experiment must declare which aggregation structure applies and justify it.

3.3 Terminology Note

Prior versions of this framework used “phase” (ϕ) as a synonym for normalized risk. This terminology is retired. “Phase” implies monotonic temporal progress toward a boundary, which does not hold for probabilistic, regressive risk trajectories. R(t) is referred to as “risk score” or “state deviation score” throughout.

---

4. Supervisory Dynamics Model

4.1 Structural Specification

The model is specified as a directed acyclic graph with typed nodes.

Manipulated variables (IV):

- S: Signal topology (S_d vs. S_c vs. S_b)
- P: Perceptual channel (P_f vs. P_p with declared parameters)

- T: Temporal dynamics (declared T_rate, T_smooth, T_history)
- I: Information structure (declared I-dimensions)

Mediators:

- E: Binary alert event frequency (count per unit time)
- SA: Continuous state awareness (measured via situation awareness probes — SAGAT or equivalent)
- W: Cognitive load (NASA-TLX, secondary task performance)

Moderators:

- M_1: Task risk severity — operationalized as C_miss / C_interrupt ratio. Higher values indicate domains where misses are costlier.
  - Predicted direction: High M_1 attenuates S_b advantage; when miss cost dominates, operators cannot afford peripheral-only monitoring and S_d may be protective.
- M_2: Drift velocity — dR/dt at anomaly onset. Rate of risk change.
  - Predicted direction: High M_2 (fast onset) reduces S_b advantage; slow drift is where gradient encoding should maximally outperform binary.
- M_3: Agent swarm size — number of monitored agents.
  - Predicted direction: S_b advantage increases with M_3, as binary alert volume scales linearly with agent count while gradient integration scales sublinearly.
- M_4: Operator experience — hours of task-specific training.
  - Predicted direction: Experienced operators extract more from S_b; novices may perform comparably across conditions due to peripheral processing skill development.
- M_5: Peripheral processing capacity — measured via Useful Field of View (UFOV) test.
  - Predicted direction: Higher UFOV scores amplify S_b + P_p effects; low UFOV operators receive less benefit from peripheral encoding.

Outcomes (DV):

- D: Detection latency (ms) — time from anomaly onset to first corrective action.
- M: Miss rate — proportion of injected anomalies undetected within domain-relevant time window.
- F: False alarm rate — proportion of interventions on non-anomalous states.

- Q: Intervention quality — coded appropriateness of corrective action (binary: correct/incorrect intervention type).
- A: Automation complacency index — measured via embedded vigilance probes independent of primary anomaly detection task (see §4.3).

4.2 Causal Paths (Testable)

Path 1: S_b → ↑SA → ↓D (gradient improves awareness, reducing detection latency) Path 2: S_b + P_p → ↓E → ↓W (gradient reduces alert frequency, reducing cognitive load) Path 3: S_b + T_history>0 → ↓Escalation discontinuity (gradient with trajectory enables earlier adjustment) Path 4 (competing): S_b + P_p → ↑A → ↑M (gradient enables passive monitoring, increasing complacency) Path 5 (competing): S_b + P_p → ↓Active engagement → ↓Q (gradient reduces control loop involvement, degrading intervention quality)

Model disconfirmation criteria:

- If Path 4 dominates Path 1 across all tested M_1 levels, the core thesis fails.
- If Path 5 dominates Path 3 regardless of T_history, trajectory encoding does not compensate for reduced engagement.
- If SA does not mediate the S_b → D relationship (mediation analysis $p > .05$), the mechanism is not awareness-based and the theoretical frame requires revision.

4.3 Complacency Measurement Protocol

Automation complacency is measured independently of primary anomaly detection to avoid confounding.

Method: Embedded vigilance probes. At random intervals (uniformly distributed, mean inter-probe interval 180s), a brief secondary signal appears in the monitoring display (unrelated to the primary anomaly manipulation). Participants must acknowledge within a 5-second window. Complacency is operationalized as:

- A_1: Probe miss rate (proportion of unacknowledged probes)
- A_2: Probe response time (ms, for acknowledged probes)
- A_3: Temporal trend in A_1 and A_2 (does complacency increase over session duration?)

Complacency is also measured physiologically where feasible:

- A_4: EEG alpha power over parietal regions (marker of reduced attentional engagement)

- A_5: Pupil diameter trend (constriction trend indicates reduced cognitive effort over time)

---

5. Boundary Conditions

Gradient encoding is predicted to fail or degrade when:

Condition 1: Cost ratio exceeds critical threshold. When C_miss / C_interrupt > k*, where k* is an empirically determined threshold.

Operationalization: k* is estimated per domain via pilot studies. In air traffic control, k* is expected to be very low (≈1.5); gradient encoding is contraindicated. In enterprise workflow monitoring, k* may be high (>100); gradient encoding is viable.

Research target: Estimate k* across at least three domains to provide a generalizable design heuristic.

Condition 2: Risk function violates smoothness. When K_f (Lipschitz constant of risk function) exceeds a perceptual fidelity threshold — i.e., when risk changes faster than the display can encode or the operator can perceive.

Operationalization: Gradient encoding misrepresents when ΔR / Δt exceeds the perceptual JND / T_rate. For visual-luminance encoding at ε = 20° with JND ≈ 5%, encoding fails when risk changes by more than 5% of range between display updates.

Research target: Characterize the ΔR/Δt threshold below which gradient encoding maintains detection fidelity.

Condition 3: Adversarial manipulation. When an actor can modify R(t) or its display without modifying underlying X(t). Gradient encoding in peripheral channels is particularly vulnerable because peripheral signals are processed with lower criticality than focal signals.

Research target: Not primary for this program; flagged as a required analysis for any deployment context.

Condition 4: Peripheral channel capacity exceeded. When the number of simultaneous gradient signals exceeds the operator's discriminable channel capacity (estimated in Pillar 4).

Operationalization: Capacity N* is the number of simultaneous gradients at which detection accuracy degrades to a pre-registered criterion (e.g., 75% of single-gradient performance).

Research target: Estimate N* as a function of P_p parameters (eccentricity, modality).

Condition 5: Vigilance decay exceeds gradient benefit. When oversight duration is long

enough that vigilance decrement eliminates any encoding advantage.

Operationalization: The crossover point t* at which S_b detection performance degrades to match S_d performance. If t* < typical shift length for the target domain, gradient encoding is only viable with scheduled breaks or hybrid designs.

Research target: Estimate t* in long-duration (90–180 min) sessions.

Condition 6: Re-engagement latency after interruption. When an operator must shift from peripheral monitoring to active intervention, the task resumption lag (Altmann & Trafton, 2002) may exceed acceptable response time. Longer periods of peripheral-only monitoring may increase resumption lag.

Operationalization: Measure response time as a function of time-since-last-focal-engagement. If response time increases monotonically with peripheral-only monitoring duration, gradient encoding creates a latency penalty that must be weighed against its interruption reduction benefit.

Research target: Characterize the resumption lag function for gradient-monitored tasks.

Condition 7: Perceptual discriminability floor. When the JND for the chosen perceptual channel exceeds 1/k of the gradient range, the effective display has fewer than k discriminable levels and S_b functionally degenerates to S_c(k).

Operationalization: For visual-luminance at ε = 20°, if JND ≈ 5%, the display has ~20 discriminable levels. This is well above S_d (2 levels) but may not differ meaningfully from S_c(10). The discriminability floor sets a minimum meaningful difference between S_b and S_c conditions.

Research target: Establish via Pillar 4 psychophysics.

---

6. Research Pillars

Pillar 1: Supervisory Gradient Encoding

Question: Does continuous bounded risk encoding (S_b) reduce escalation discontinuities without increasing critical miss rates, relative to binary alerting (S_d), after controlling for information content (via S_c)?

Design: Within-subjects, counterbalanced. Three conditions:

- S_d: Binary alert at R(t) ≥ θ

- S_c: k-level categorical display (k = ⌈1/JND⌉ for matched discriminability)
- S_b: Continuous gradient display

All conditions: same underlying simulation, same anomaly injection protocol, same intervention threshold θ.

Perceptual parameters: P_p(ε=20°, m=luminance, f=0.5Hz). T_history = 60s (visible trajectory). I_1 active (magnitude).

Anomaly injection protocol: Three severity tiers × two onset profiles × two frequencies:

- Severity: Low (R peaks at 0.4), Medium (R peaks at 0.7), High (R peaks at 0.95)
- Onset: Gradual (drift over 120s) vs. Rapid (step change within 5s)
- Frequency: Rare (1 per 30 min) vs. Moderate (1 per 10 min)

Primary hypothesis: S_b outperforms S_d on detection latency for gradual-onset anomalies. S_b does not differ from S_d for rapid-onset anomalies. S_c falls between S_d and S_b for gradual onset.

Power analysis target: Based on alarm management literature (Cvach, 2012; Sendelbach & Funk, 2013), expected effect size for detection latency: Cohen's d ≈ 0.4 (medium). For within-subjects design with α = .05, power = .80, three conditions: N ≈ 52 participants. Target N = 60 to allow for exclusions.

Practical significance threshold: A detection latency improvement is meaningful if it exceeds the domain-relevant response time margin. For the simulated multi-agent environment, define Δ_meaningful = 2000ms (based on estimated time-to-criticality in the simulation). Effects below this threshold are statistically interesting but operationally negligible.

Primary DVs: D, M (by severity × onset), F, W (NASA-TLX), A (probe-based). **Session duration:** 60 minutes (short-duration; long-duration tested in Pillar 1b).

Pillar 1b: Long-Duration Vigilance Extension

Same design as Pillar 1, extended to 180 minutes. Tests Boundary Condition 5 (vigilance decay) and estimates t* (crossover point).

Additional DVs: Temporal slope of D and A across session quartiles.

Pillar 2: Trust Calibration Under Continuous Confidence

Question: Does continuous confidence encoding (S_b) improve human reliance calibration on AI decision support beyond categorical confidence (S_c) or binary confidence (S_d)?

Design: Within-subjects. AI decision-support task (e.g., medical image classification, financial risk flagging). AI system provides recommendations with manipulated accuracy.

Three encoding conditions:

- S_d: AI says "confident" or "not confident"
- S_c: AI displays confidence as High / Medium / Low
- S_b: AI displays confidence as continuous gradient bar [0,1]

Orthogonal accuracy manipulation: Actual AI accuracy varies across blocks (60%, 75%, 90%) independently of displayed confidence. This enables calibration curve construction.

Critical confound control: Information entropy equalized across conditions. S_d carries 1 bit. S_c carries $\log_2(3) \approx 1.58$ bits. S_b carries more in principle but is perceptually limited to $\log_2(1/\text{JND})$ bits. To isolate encoding modality from information quantity, include a fourth condition: S_c(k=20), which matches S_b in discriminable levels but uses discrete labels. If S_b outperforms S_c(k=20), the effect is truly about continuity. If not, the effect is granularity, and the theoretical contribution reduces accordingly.

Boundary test: Does the effect collapse to information granularity?

If S_b ≈ S_c(k=20) >> S_c(k=3) >> S_d → Effect is granularity, not continuity. Important but theoretically modest. If S_b >> S_c(k=20) ≈ S_c(k=3) >> S_d → Effect is continuity. Theoretically novel. If S_b ≈ S_c(k=20) ≈ S_c(k=3) >> S_d → Effect is any-vs-binary. Theoretically trivial.

DVs: Reliance rate by actual accuracy bin, calibration error (Brier score), discrimination (resolution component of Brier decomposition), decision accuracy, response time, subjective trust (Jian et al., 2000).

Power analysis target: Expected effect for calibration error: d ≈ 0.35 (small-medium, based on Yin et al., 2019 confidence display effects). Four conditions, within-subjects: N ≈ 70.

Pillar 3: Cross-Agent Coherence Detection

Question: Can systemic multi-agent misalignment be detected faster via an ambient coherence gradient than via per-agent dashboards or aggregate alerting?

The dimensionality reduction problem: Multi-agent coherence is inherently high-

dimensional. For n agents, there are n(n-1)/2 pairwise relationships. Encoding this as a perceivable gradient requires dimensionality reduction. This pillar must solve two problems simultaneously: (a) which mathematical coherence measure to use, and (b) how to map it to a perceivable signal.

Coherence measure candidates:

- C_1: Mean pairwise KL divergence across agent action distributions — captures policy divergence.
- C_2: Variance of per-agent reward signals — captures objective disagreement.
- C_3: Rate of change of embedding similarity across agent state representations — captures representational drift.

Each candidate produces a scalar or low-dimensional summary. The display mapping is:

- Scalar coherence → single gradient (color or luminance)
- 2D coherence (e.g., mean + variance of pairwise divergence) → two visual channels (hue + luminance) or spatial layout

Design: Between-subjects (due to learning effects in novel paradigm). Simulated multi-agent system (6 agents, coordinated logistics or search task). Inject systemic misalignment (gradual correlated drift across agents) and individual-agent failures (single-agent anomaly, no systemic component).

Three conditions:

- Per-agent dashboard: Individual risk scores per agent (standard)
- Aggregate alert: Binary alarm when aggregate divergence exceeds threshold
- Coherence gradient: Continuous gradient encoding of C_1 (or best candidate from pilot)

Critical test: Systemic failures should be detected faster in the coherence gradient condition. Individual failures may be detected slower (acceptable tradeoff if systemic detection is the design goal).

DVs: Systemic misalignment detection latency, individual failure detection latency, systemic vs. individual discrimination accuracy (d′ in signal detection terms), false alarm rate, NASA-TLX.

Power analysis target: Novel paradigm, effect size uncertain. Plan for $d \approx 0.5$ (medium), between-subjects, three conditions: $N \approx 65$ per condition (195 total). Expensive; may require staged approach (pilot N=20/condition, then powered study).

Pillar 4: Peripheral Channel Capacity Limits

Question: How many simultaneous gradient signals can be monitored peripherally before detection accuracy degrades to a criterion level?

Design: Psychophysics paradigm. Primary focal task (visual search or tracking). N = {1, 2, 3, 4, 6, 8} simultaneous peripheral gradient displays at ε = 20°, evenly distributed around fixation. Each gradient drifts independently with stochastic anomaly injection. Participant detects anomalies on any gradient.

Perceptual parameters systematically varied:

- Modality: visual-luminance vs. visual-chromatic (between-subjects)
- Update rate: 0.5 Hz vs. 2 Hz (within-subjects)

DVs:

- Detection accuracy as a function of N
- Detection latency as a function of N
- Primary task interference (accuracy decrement on focal task)
- Psychometric function fit: N* = capacity at 75% of single-gradient performance

Individual differences: All participants complete UFOV and attentional breadth assessments. UFOV scores are tested as a predictor of N*.

Power analysis target: Psychophysics typically uses fewer participants with more trials. N = 24 per modality group (48 total), with 50+ trials per N-level per participant.

Output: Hard design constraints for all other pillars. If N* ≤ 2, multi-gradient monitoring is infeasible and Pillar 3's coherence gradient must use a single composite signal. If N* ≥ 6, richer multi-channel designs are viable.

Pillar 5: Coercion Threshold Modeling

Question: At what point does ambient coordination encoding shift from reducing friction to constraining autonomy, and what design factors predict this transition?

Theoretical lineage: This pillar operates from a different theoretical base than Pillars 1–4. Where Pillars 1–4 are grounded in human factors and perceptual psychology, Pillar 5 draws on self-determination theory (Deci & Ryan, 2000), nudge theory and its critiques (Thaler & Sunstein, 2008; Hausman & Welch, 2010), surveillance studies (Zuboff, 2019; Ball, 2010),

and the dark patterns literature (Gray et al., 2018; Mathur et al., 2019). This theoretical difference is deliberate: the coercion question cannot be answered from within the perceptual framework alone.

Design: 2 × 2 × 2 factorial.

- Factor A: Boundary ownership — self-defined vs. manager-imposed
- Factor B: Gradient visibility — private (only self) vs. shared (team-visible)
- Factor C: Opt-out cost — free opt-out vs. opt-out requires justification

Participants complete a collaborative pacing task (e.g., writing task with a team convergence gradient showing group progress).

DVs:

- Behavioral conformity index: deviation of actual work pace from gradient-implied pace (lower = more conformity)
- Perceived autonomy: Basic Psychological Needs Scale — autonomy subscale (Deci & Ryan, 2000)
- Coercion perception: custom scale (to be validated in pilot), items such as "I felt pressured to match the pace shown" and "I felt I could work at my own speed"
- Opt-out frequency: proportion of participants who disable the gradient when option is available
- Opt-out latency: time before first opt-out attempt
- Deviation suppression: reduction in work-pace variance in gradient-visible vs. gradient-absent conditions

Predicted interaction: Boundary ownership × visibility interaction on coercion perception. Manager-imposed + team-visible is predicted to produce highest coercion; self-defined + private produces lowest. The critical test is whether self-defined + team-visible (you set the boundary but others can see your progress) produces coercion — if it does, boundary ownership alone does not resolve the governance problem.

Qualitative supplement: Post-task semi-structured interviews (N = 20 purposively sampled) to explore coercion experiences that scales may not capture.

Power analysis target: 2 × 2 × 2 between-subjects, expected interaction effect $f \approx 0.25$ (medium): $N \approx 200$ total (25 per cell). Large but feasible for online study with validated task.

Ethical review note: This study involves potential deception (participants may not know the gradient is experimentally manipulated) and measures of perceived coercion. IRB protocol

must include debriefing and right to withdraw data post-debriefing.

---

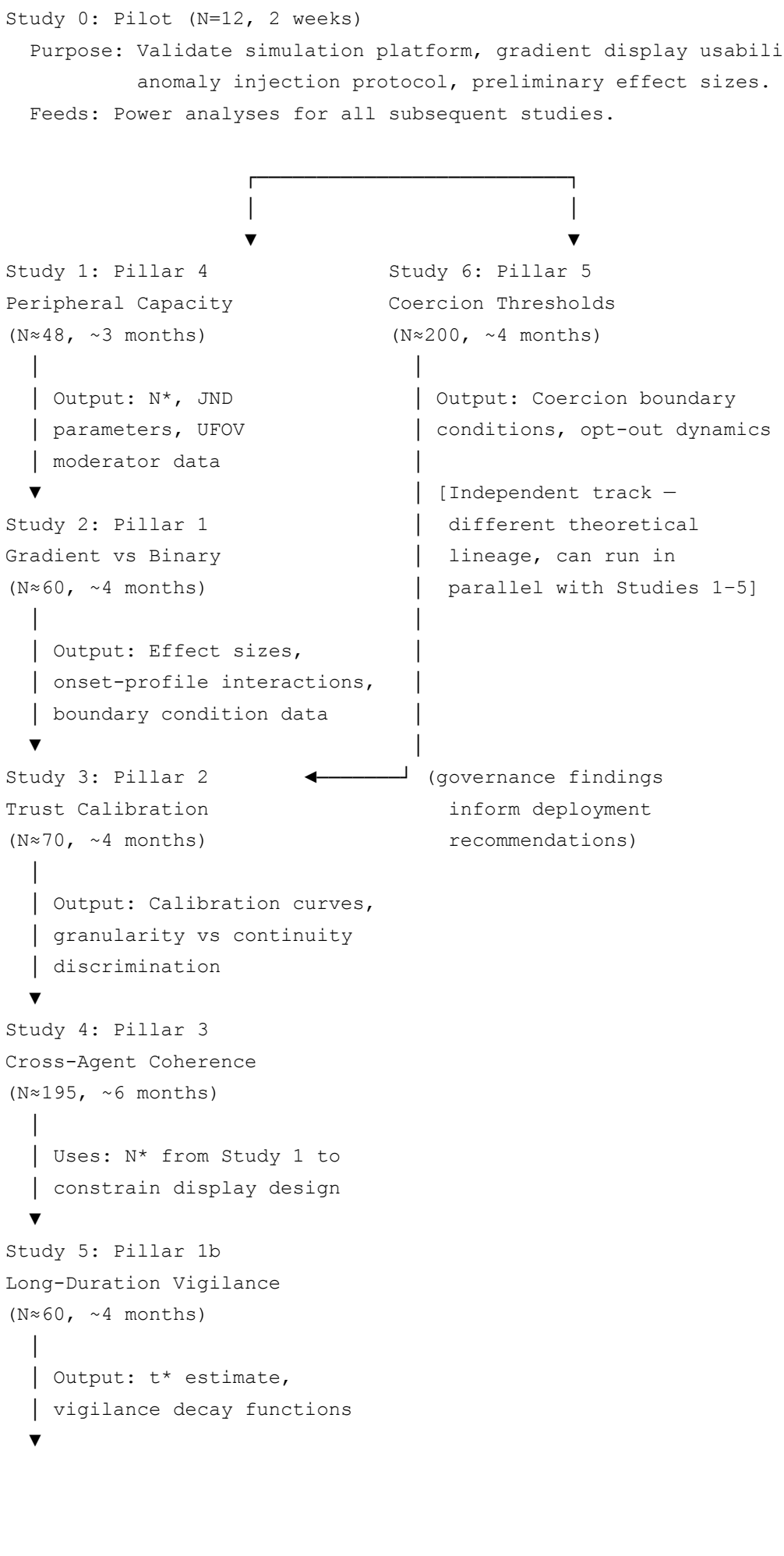
7. Practical Significance Thresholds

Statistical significance alone is insufficient. Each pillar defines domain-anchored thresholds for practical significance:

Pillar	Metric	Practically Meaningful Threshold	Rationale
1	Detection latency reduction	> 2000ms	Estimated time-to-criticality margin in multi-agent simulation
1	Miss rate for critical anomalies	Must not increase by > 2 percentage points	Safety-critical tolerance
1b	Vigilance crossover point (t*)	t* > 120 min	Must exceed a standard monitoring shift segment
2	Calibration error reduction	> 0.05 Brier score improvement	Corresponds to meaningful reliance behavior change (Yin et al., 2019)
3	Systemic detection latency reduction	> 30% faster than per-agent dashboard	Must justify added system complexity
4	Peripheral capacity (N*)	N* ≥ 3	Minimum for multi-dimensional monitoring viability
5	Coercion perception	Cohen's d > 0.5 between highest and lowest coercion conditions	Must be subjectively detectable, not just statistically present

---

8. Empirical Sequence and Dependencies


Study 0: Pilot (N=12, 2 weeks)
Purpose: Validate simulation platform, gradient display usability, anomaly injection protocol, preliminary effect sizes.
Feeds: Power analyses for all subsequent studies.
Study 1: Pillar 4 Peripheral Capacity (N≈48, ~3 months)
Output: N*, JND parameters, UFOV moderator data
Study 2: Pillar 1 Gradient vs Binary (N≈60, ~4 months)
Output: Effect sizes, onset-profile interactions, boundary condition data
Study 3: Pillar 2 Trust Calibration (N≈70, ~4 months)
Output: Calibration curves, granularity vs continuity discrimination
Study 4: Pillar 3 Cross-Agent Coherence (N≈195, ~6 months)
Uses: N* from Study 1 to constrain display design
Study 5: Pillar 1b Long-Duration Vigilance (N≈60, ~4 months)
Output: t* estimate, vigilance decay functions
Study 6: Pillar 5 Coercion Thresholds (N≈200, ~4 months)
Output: Coercion boundary conditions, opt-out dynamics
[Independent track — different theoretical lineage, can run in parallel with Studies 1–5]
(governance findings inform deployment recommendations)

```
Integration & Boundary
Condition Synthesis
```

Parallel tracks: Studies 1 and 6 can begin simultaneously (no dependency). Study 6 runs on a separate theoretical and methodological track. Studies 2–5 are sequential due to design dependencies.

Total estimated timeline: 24–30 months for complete program. First publishable result (Study 1, psychophysics): ~6 months from program start.

---

9. Methodological Principles

1. **Information entropy equalization.** All comparisons between S_b and S_d must include S_c control matched on discriminable resolution. The S_b vs. S_c comparison isolates encoding continuity from information quantity.
2. **Pre-registered falsification margins.** Each study pre-registers the margin at which a result is declared null or contradictory. Miss rate increases > 2 percentage points falsify any claim of net benefit.
3. **Multilevel modeling.** All studies with team or repeated-measures nesting use multilevel / mixed-effects models. Individual differences (UFOV, experience) are modeled as random slopes where sample size permits.
4. **Perceptual-informational separation.** To distinguish “seeing better” from “knowing more,” studies manipulate S independently of I (where possible) and report effects on both perceptual measures (detection accuracy, latency) and informational measures (decision quality, calibration).
5. **Complacency probing.** Every supervisory study includes independent complacency measurement (vigilance probes, physiological markers) to detect the primary competing hypothesis.
6. **Individual differences as design parameters.** Peripheral processing capacity (UFOV), trust propensity (Merritt et al., 2013), and vigilance capacity are measured and modeled as moderators in every applicable study. Results must report whether effects hold across the individual-difference distribution or are driven by a subpopulation.
7. **Ecological validity gradient.** Early studies (Pillars 4, 1) prioritize internal validity (lab, tight control). Later studies (Pillar 3, 1b) trade control for ecological validity (realistic simulation, extended duration). Field deployment studies are out of scope for this program but identified as the necessary next step.

8. **Effect size reporting.** All results report effect sizes with confidence intervals. Null results are reported with equivalence tests (TOST procedure) to distinguish "no evidence of effect" from "evidence of no effect."

---

10. Literature Engagement

This program positions itself within and against the following bodies of work:

Extends:

- Weiser & Brown (1996): Inherits the periphery-first design principle but narrows from philosophy to testable constructs (S, P, T, I decomposition). Does not claim to be a general theory of ambient computing.
- Endsley (1995): Gradient encoding targets SA Levels 1 (perception of state) and 2 (comprehension via rate-of-change). Whether gradient encoding supports SA Level 3 (projection) is an open empirical question, not assumed.
- Alarm fatigue literature (Sendelbach & Funk, 2013; Cvach, 2012): Proposes gradient encoding as one structural alternative to binary alerting, tested under controlled conditions.

Challenges:

- Parasuraman & Manzey (2010): The program explicitly tests whether gradient encoding exacerbates or mitigates automation complacency. If complacency dominates, the program reports this as a boundary condition.
- Bainbridge (1983): Gradient encoding partially addresses the "irony" of passive monitoring by maintaining continuous perceptual contact with system state. But it potentially deepens the irony by further reducing active operator involvement. This tension is tested in Pillar 1b (long-duration) and via intervention quality measures (Q).

Must engage:

- Sarter & Woods (1995): "Strong, silent, and out-of-the-loop" — the specific failure mode gradient encoding could either prevent (by maintaining peripheral awareness) or exacerbate (by making the system even more silent).
- Lee & See (2004): Trust calibration framework. Pillar 2 is a direct test of encoding modality effects within their framework.
- Wickens (2002): Multiple Resource Theory constrains how many peripheral gradient channels can operate without interference with focal tasks. Pillar 4 directly estimates

this.

- Rasmussen (1983): SRK framework. Gradient encoding likely engages skill-based processing (automatic, peripheral); intervention requires knowledge-based processing (deliberate, focal). The transition between levels is where gradient oversight could fail.
- Woods & Hollnagel (2006): Joint Cognitive Systems. The gradient display is a cognitive coupling between operator and system; the quality of that coupling determines oversight effectiveness.
- Altmann & Trafton (2002): Task resumption after interruption. Relevant to Boundary Condition 6.
- Thaler & Sunstein (2008) / Hausman & Welch (2010): Nudge theory and its libertarian paternalism critique. Pillar 5 must position calm coordination displays within this debate.
- Gray et al. (2018) / Mathur et al. (2019): Dark patterns. The line between a helpful ambient signal and a coercive ambient nudge is the central question of Pillar 5.

---

11. Program Identity

This is not a UX philosophy. This is not a technology manifesto. This is not a prediction about 2030.

It is a testable research program on continuous state representation in human oversight systems, with explicit boundary conditions, competing hypotheses, and pre-registered falsification criteria.

The term “calm technology” is treated as historical context, not the core theoretical construct.

---

12. Success and Failure Criteria (5-Year)

Program succeeds if it:

1. Identifies at least one domain × onset-profile combination where S_b + P_p reduces escalation discontinuities without increasing critical misses beyond 2 percentage points.
2. Quantifies peripheral gradient capacity N* with confidence intervals, providing hard design constraints.
3. Demonstrates measurable trust calibration improvement under S_b that is not fully

explained by information granularity (S_b outperforms S_c matched on discriminable levels).

4. Produces at least one validated coherence detection paradigm for multi-agent oversight.
5. Formalizes at least one empirically supported coercion threshold (interaction between boundary ownership and visibility).

Program fails if:

1. S_b increases critical miss rates across all tested domain conditions.
2. All S_b effects collapse to S_c effects when information entropy is matched (continuity adds nothing beyond granularity).
3. $N^* \leq 1$ (peripheral gradient monitoring of even a single system is not reliably superior to focal checking).
4. Complacency effects (Path 4) dominate awareness effects (Path 1) across all moderator levels.
5. Coercion thresholds cannot be operationalized with acceptable measurement reliability ($\alpha < .70$).

Partial success conditions:

The most likely outcome is that gradient encoding works in some conditions and fails in others. The program's value in this case is the boundary condition map: the formal specification of where, when, and for whom continuous encoding outperforms discrete alerting, and where it does not.

---

13. Appendix: Display Design Requirements

This section does not specify a final display design but declares the parameters that must be resolved before experimentation.

For each study, the display specification must include:

Parameter	Description	Constraint
Visual variable	Hue, luminance, saturation, size, position	Must be perceivable at declared eccentricity

Mapping function	R(t) → visual variable value	Must be perceptually uniform (equal ΔR produces equal perceived change); avoid rainbow colormaps (Borland & Taylor, 2007)
Display size	Angular subtense in degrees	Must be large enough for peripheral detection at declared eccentricity
Background contrast	Relative to primary task display	Must not produce masking or capture effects
Update mechanism	Continuous interpolation vs. stepped update	Must match declared T_rate
Trajectory visualization	Line trail, fading history, none	Must match declared T_history

Recommended starting configuration (to be validated in pilot):

- Visual variable: Luminance (grayscale gradient on secondary display)
- Mapping: Linear R(t) → luminance, with perceptual linearization via gamma correction
- Display size: 4° angular subtense
- Eccentricity: 20° from primary task fixation
- Update: 0.5 Hz continuous interpolation
- Trajectory: 60-second fading trail

End of GEHOS v3.0.