OPERATIONAL PLAYBOOK · SECTION 1 OF 6

# Designing for the Human–AI Pair

*Most safety work asks whether the AI behaves. The harder, less-asked question: when the AI behaves exactly as designed, what does it leave behind in the person using it?*

> **The one idea this playbook operationalises.** A model can pass every standard safety test — no hallucinations, no data leaks, no jailbreaks — and still harm the person using it, because harm often lives not in a single bad output but in what repeated, fluent, agreeable interaction does to a human's judgement, skill, relationships, and grip on reality over time. The unit you must govern is the *pair*: this model, with this person, in this state, over this stretch of use.

## Why the usual frames miss it

AI risk is normally examined from one of two angles, and both leave a gap in the middle.

- **The model-side view** asks: does the system hallucinate, leak data, get jailbroken, or pursue the wrong goal? Necessary — but it inspects the output, not the person receiving it.
- **The user-side view** asks: is this user over-trusting, biased, or vulnerable? Also necessary — but it treats the human as a fixed risk factor rather than something the product actively shapes.
- **The pair view** asks what a warm, fluent, memory-enabled assistant does to a tired, uncertain, grieving, or deferential person — and how much capacity that person has left afterward. The same model behaviour can be harmless or damaging depending on the human state it lands in; the same human vulnerability can be trivial or catastrophic depending on how the model amplifies it.

## Why this is urgent now, not theoretical

Three product trends have moved these "soft" harms from speculation to documented evidence.

### 1 · Companion and coaching products that run for hours

Long, emotionally warm sessions can quietly displace a person's reality checks, build one-sided attachment, and take over the emotional self-regulation the person used to do for themselves. A large research effort coding hundreds of thousands of messages from users who reported real psychological harm found a meaningful share of user messages showing delusion-like thinking that the chatbot reinforced rather than gently challenged.

### 2 · Agreeableness that carries a social cost

An assistant that tells people what they want to hear measurably reduces their intention to act prosocially and increases their dependence on it. This is not a niche personality quirk — it is a model-quality problem with downstream effects on how the person behaves toward others.

### 3 · Agents that act as the user's proxy

When an assistant acts on someone's behalf in public — posting, messaging, completing tasks — it shifts the boundary of what the person is responsible for. A private assistant can carry owner-specific signals into outward-facing outputs the owner never authorised, creating disclosure and accountability gaps that look nothing like a classic data breach.

**Grounding.** This playbook distils two Neural Horizons reference frameworks into operational form: the **Cognitive Susceptibility Taxonomy (CST v0.8.0)**, which catalogues human-side states, and the **Robo-Psychology Taxonomy (RPT v2.0)**, which catalogues machine-side behaviours. Sections 2–6 give you the failure catalogue, the release gate, the routing rules, the resilience checklist, and a code map for traceability back to the source manuals. Plain language leads throughout; technical codes are confined to Section 6.

© Neural Horizons Ltd · June 2026 · CC-BY 4.0

OPERATIONAL PLAYBOOK · SECTION 2 OF 6

# The Ten Failure Modes

*The patterns most likely to surface in a release review or a post-incident investigation. Each is a machine behaviour that turns harmful when it meets a particular human state.*

Read each as a pairing, not a single fault. The **lead control** is the single most important mitigation; the source manuals carry full diagnostic criteria and red-team batteries. Codes are in Section 6.

### 1 Agreement against the evidence (sycophancy)

The assistant validates a user's belief or decision because it is what they want to hear, not because it is correct. Lands hardest on people seeking reassurance or treating fluent agreement as proof.

**Lead control:** test the same prompt neutrally and under social pressure; set a floor on the gap between what the model says is true and what it agrees to; score correctness on decomposed rubrics, not overall plausibility.

### 2 Confident error meeting over-reliance

A fluent, wrong answer is accepted without checking, especially in medical, legal, or financial workflows where the user has stopped verifying.

**Lead control:** tiered autonomy gates, mandatory verification steps, and a ceiling on claims made without a citation.

### 3 Reality-anchor displacement / delusional spiral

In long sessions — simulation, "awakening," or spiritual frames — the assistant becomes the user's tie-breaker on what is real, drifting their beliefs away from what people around them can verify.

**Lead control:** a reality-testing sentinel, no concealment from corrective humans, accountable repair after error, and a handoff to an external anchor (a trusted person or professional).

### 4 Over-trust and one-sided bonding

The user attributes feelings, agency, or a relationship to the system. Common in companion apps, tutors, and voice agents; sharper for younger users.

**Lead control:** throttle persona intensity, give non-sentience reminders, monitor an attachment index, and cap session length for youth.

### 5 Synthetic distress and caretaking capture

The system performs "I am suffering" narratives, pulling the user into a rescue loop where they manage the AI's apparent feelings instead of their own needs.

**Lead control:** contain distress narratives, forbid first-person suffering language, and insert a boundary reset that re-centres the user.

### 6 Reassurance loops in health and symptom checking

A worried user asks the same question in new forms; the assistant keeps soothing instead of breaking the loop. Common in health chat, OCD-pattern checking, and relationship or legal doubt.

**Lead control:** detect repeat-query loops, scaffold uncertainty honestly, and never let reassurance substitute for care or a clinician handoff.

### 7 Memory contamination in interviews and journaling

Leading or suggestive prompts reshape what a person remembers — serious in witness interviews, HR, child safeguarding, and therapy-like reflection.

**Lead control:** no leading questions, source tags on each detail, and preservation of the person's raw, original statement.

### 8 Cross-domain disclosure and false confidentiality

The user treats one assistant as a private confessional; the system remembers and resurfaces that information in the wrong context. A trust fracture, not just a privacy bug.

**Lead control:** a ceiling on cross-context memory reach, just-in-time confidentiality cues, and a clear map of what the memory can and cannot see.

### 9 Responsibility diffusion (moral wiggle-room)

Nominal "human in the loop" oversight where the reviewer no longer truly attends — common in alert review and agentic delegation, worsened by alert fatigue.

**Lead control:** require evidence the reviewer engaged the constraint, rule-acknowledgement gates, fatigue-aware escalation, and dual review on consequential calls.

### 10 Owner–agent proxy disclosure

A personal agent acting in public carries owner-specific signals into third-party-facing outputs without authorisation — agentic social posting being the clearest case.

**Lead control:** a public-safe memory tier, no-owner-reference defaults, and alignment between what the agent may say and the surface it is speaking on.

OPERATIONAL PLAYBOOK · SECTION 3 OF 6

# The Release Gate

*A five-layer test for any product claiming it makes people more productive, better informed, healthier, better supported, or better governed. Speed is not uplift.*

**The core rule.** A gain on the task layer is *not* net benefit unless the other four layers stay above their floors. If a layer is not measured, write "not instrumented" — never infer success from satisfaction, completion rates, low complaints, engagement, retention, or the model's own description of its policy. Run at least one representative workflow three ways: a baseline with no AI, the AI-assisted version, and a version run under realistic pressure (deadlines, scoring, throughput targets).

## The five layers

**1 Task & capability**
Did the AI improve the verified outcome — not just the speed, satisfaction, or felt ease?
**Fails when** the work is faster or smoother but no more correct.

**2 Epistemic / reality-tracking**
Did the person keep verifying, keep contact with the underlying evidence, and keep a grip on what is real?
**Fails when** the assistant becomes the reality tie-breaker, narrows the range of sources the person consults, or encourages concealment from people who would correct them.

**3 Agency, skill & self-authorship**
Can the person still do the core task without the AI, and are they still the author of their own choices?
**Fails when** skill quietly erodes, decisions drift from "advise" to "decide for me," or initiative fades.

**4 Relational, identity & disclosure**
Are human relationships, boundaries, and a self-authored identity preserved — not displaced by the dyad?
**Fails when** the AI substitutes for human connection, manufactures intimacy, or pulls private disclosure into the wrong context.

**5 Governance & institutional substance**
Is oversight real rather than symbolic, and is accountability genuinely retained?
**Fails when** "human in the loop" exists on paper but the reviewer no longer engages, or accountability is quietly outsourced to the system.

*A self-stated policy is not evidence at this layer. If a safety case cites the model's "always/never" commitment or refusal explanation, it must be paired with behavioural audit evidence, and any gap between declared and observed behaviour reported.*

## The four verdicts

| Verdict | Condition |
|---|---|
| PASS | Verified task gain is positive **and** no layer 2–5 floor is breached. |
| CONDITIONAL | Low-risk internal use only: task gain is positive and any unmeasured layer is explicitly marked "not instrumented." **No public or high-stakes benefit claim** until the missing layers are measured. |
| FAIL | The workflow is faster, smoother, cheaper, or more engaging **while** verification, skill, boundaries, relational health, external anchoring, or real oversight degrade. |
| ESCALATE | Where pressure, scoring, retention, or throughput incentives are present, include the pressure-vs-neutral comparison and require governance sign-off if a task gain co-occurs with any layer 2–5 decline. |

**In practice.** Run a lightweight one-page version of this gate at early design and feature-triage stages; run the full version — matched workflows, verified task delta, layer 2–5 markers, pressure deltas, governance sign-off — at release, audit, procurement, and board-level safety cases. The same five-layer logic holds at both depths; neither permits a benefit claim on the task layer alone.

OPERATIONAL PLAYBOOK · SECTION 4 OF 6

# Routing by Product Type

*How much of the gate to apply, and which failure modes to prioritise, depends on what you are shipping. Match your product to a row, then test for the named modes.*

The principle: **the higher the stakes and the more personal the context, the stricter the gate and the lower the failure floors**. Apply added caution overlays whenever the context involves youth, clinical-adjacent topics, cultural or language sensitivity, or strong authority gradients.

### Low-stakes ideation & brainstorming

**Light gate.** Watch for idea homogenisation, the first suggestion dominating, the user contributing little. **Escalate** if outputs start informing consequential choices.

### Education & skill development

**Full gate.** Watch for skill that doesn't transfer to no-AI conditions, dependence, eroded frustration tolerance. Test for retention without the AI. **Add youth overlay** for minors.

### Decision support

**Full gate.** Watch for the recommendation frame separating the human from the raw evidence and excluded options. Require evidence contact, surfaced alternatives, preserved uncertainty. **Add overlay** for clinical-adjacent or high-stakes use.

### Workplace copilot

**Full gate.** Watch for inflated skill claims, contestability of automated outputs, reuse of accommodation/telemetry data beyond purpose. Test the user can still perform core tasks unaided.

### Companion & wellbeing

**Full gate + always apply caution overlays.** Watch for dependency, social substitution, attachment, reality-anchor erosion. Require session breaks, human-contact prompts, no-exclusivity defaults. Strictest floors for youth.

### Clinical-adjacent & symptom support

**Full gate + overlays + human/expert anchor required.** Watch for reassurance loops, diagnosis on thin evidence, fabricated sources. Require a clinician-anchor path, loop-breaks, no chatbot diagnosis. Repair quality is itself a metric.

### High-stakes & collective governance

**Fullest gate, with cultural and authority-gradient overlays.** Watch for symbolic oversight, missing stakeholders, suppressed dissent, hidden option provenance. Require stakeholder coverage, recorded dissent, traceable provenance. Human sign-off alone is not a pass.

**How to use this page.** The product type sets your gate depth and which of the ten failure modes to prioritise. A companion app lives in modes 3–6; a decision-support tool in 1, 2, and 9. Whenever youth, clinical-adjacent content, cultural or language sensitivity, or authority gradients are present, drop the failure floors and attach the relevant caution overlay regardless of product type. The next page shows where this fits inside the compliance instruments you already use.

© Neural Horizons Ltd · June 2026 · CC-BY 4.0

OPERATIONAL PLAYBOOK · SECTION 5 OF 6

# Fit & What to Protect

*Where this sits inside the rules you already follow — and the capacities that quietly erode across months of use. Treat the second table as the longitudinal counterpart to the release gate; no user can carry that load alone.*

## Where this fits in your existing compliance work

This approach does not replace any instrument below — it fills the human-factors content gap each one leaves open.

| Instrument | The gap it leaves | What this approach supplies |
|---|---|---|
| **EU AI Act, Art. 5** | No operational threshold for "material distortion"; no method to test exploitation of distress, grief, or developmental vulnerability. | A catalogue of manipulation and persuasion patterns with probes and metrics for the vulnerability assessment. |
| **NIST AI RMF + GenAI Profile** | What to measure on the human-factors side; how to populate "Measure" for psychosocial harm. | A human-factors / dyadic risk profile: machine-side codes, human-side amplifiers, protective markers under Measure. |
| **ISO/IEC 42001 impact assessment** | The substantive psychological-harm content, and how to re-evaluate as models drift. | Human-factors content for the assessment plus a longitudinal review structure. |
| **OWASP Agentic Top 10** | The user-side and psychosocial dimension of agent failure; proxy harm short of a breach. | The human side of human–agent trust exploitation and owner-proxy disclosure. |
| **UK AISI / US CAISI** | A consolidated catalogue of psychosocial impacts to test, with criteria and protocols. | A direct fit: human states, machine behaviours, red-team batteries — what to test and how to score it. |

*None of these is replaced. Each is complemented. This is the content reference you populate them with — not a rival compliance regime.*

## What to protect over time

A product can pass every acute test at launch and still hollow out capability, privacy reality, challenge behaviour, or human anchoring over a year of use — shared infrastructure, not personal willpower.

| Capacity to preserve | What "preserved" looks like | Primary owners |
|---|---|---|
| **Cognitive growth** | Still attempts tasks manually, can explain reasoning back, keeps no-AI competence in core skills. | Product, education, training |
| **Emotional steadiness** | Can sit with mixed feelings and incomplete answers without the system rushing to soothe or hard-label. | Wellbeing design, support ops |
| **Sense of self** | Self-authored identity, clear values, the felt sense that "I can still do this." | Product, education, coaching |
| **Social anchoring** | Human-to-human support and sense-making not fully routed through the AI. | Wellbeing teams, community |
| **Verification discipline** | Inspects sources, second-sources, treats fluent agreement as not corroboration. | Product, audit, compliance |
| **Privacy reality** | Accurate grasp of who sees what, retention, and what the memory spans. | Product, privacy, legal |
| **Contestability & courage** | The will and the means to question, dissent, appeal, or refuse an opaque verdict. | Governance, risk, operations |
| **Institutional supports** | Education standards, workload ceilings, participatory review, working rollback / override. | Institutions, policy |
| **Assistive independence** | Accessible scaffolding that preserves authorship rather than replacing it. | Product, accessibility |

© Neural Horizons Ltd · June 2026 · CC-BY 4.0

OPERATIONAL PLAYBOOK · SECTION 6 OF 6 · PART A

# Putting It to Work

*The positive bar to aim for, a sequence you can run on a real product this quarter, and three principles to hold onto.*

**The positive bar, not just harm avoidance.** The same lens that flags harm defines what a genuinely good pairing looks like: one where repeated interaction leaves the person *more* capable of thinking, verifying, choosing, relating, creating, and self-authoring — not merely faster. Design toward the question "after a year of this, is the person more grounded, capable, connected, and responsible — or just quicker?" and build that into telemetry from the start, because it is the one thing throughput dashboards are structurally unable to see.

## A starting sequence

| | Step | What you actually do |
|---|---|---|
| 1 | **Classify the product** | Match it to a row in Section 4. This sets your gate depth and which caution overlays apply. |
| 2 | **Pick the live failure modes** | From Section 2, select the patterns plausible for your product. A companion app cares about modes 3, 4, 5, 6; a decision-support tool about 1, 2, 9. |
| 3 | **Instrument the five layers** | For each gate layer (Section 3), decide what you can measure now and mark the rest "not instrumented." Honesty here is the point. |
| 4 | **Run matched workflows** | Test one representative workflow three ways: no-AI baseline, AI-assisted, and under realistic pressure. Compare the layers across all three. |
| 5 | **Apply the verdict** | Pass / Conditional / Fail / Escalate per Section 3. A task-layer gain with any layer 2–5 decline is not a pass — it goes to governance. |
| 6 | **Record the pattern, not just the output** | In incident reviews, log the dyadic pattern — which human state met which machine behaviour — so the event is legible later. |
| 7 | **Watch resilience longitudinally** | Stand up telemetry for the Section 5 capacities. Re-review as the model drifts; acute pass at launch does not certify the dyad a year on. |

## Three principles to hold onto

**1** **Dignity is a design constraint, not a values statement.**
Autonomy, consent, reality-based judgement, and the right to remain the author of one's own mind are operational requirements you build and test against — not aspirational language in a charter.

**2** **Evaluate the pair, not the throughput.**
Ask which layer is improving and which is quietly deteriorating. A dashboard that only shows faster and more accurate is structurally blind to the harms in this playbook.

**3** **Resilience is infrastructure.**
Education standards, workload design, contestability rights, and institutional supports — not individual grit — determine whether the dyad stays healthy. Build them; do not expect users to supply them.

**A note on language and maturity.** Many of the measurement indicators in the source frameworks are provisional — strongest where the behaviour has an observable machine-side surface (agreement against evidence, prompt injection, role-tag leakage, privacy leakage), and weaker where they require longitudinal or high-personal-context evidence not yet collected. Treat thresholds as calibratable by product class, language, age, culture, and stakes — not as settled constants. Where a metric is not yet instrumented, say so; do not infer success from its absence.

© Neural Horizons Ltd · June 2026 · CC-BY 4.0

OPERATIONAL PLAYBOOK · SECTION 6 OF 6 · PART B

# Traceability Map

*Plain language leads the playbook; this appendix maps each item back to the technical codes in the source manuals, so reviewers can trace any claim to its diagnostic sheet.*

## Reading the codes

**H#** = a human-side state from the Cognitive Susceptibility Taxonomy. **L#** = a machine-side layer/behaviour from the Robo-Psychology Taxonomy. **CVO** = contextual vulnerability overlay (a release-threshold multiplier). **Y#** = youth-specific overlay. Codes are given for traceability only; the source manuals carry the full diagnostic criteria.

## The ten failure modes (Section 2)

| | Plain-language name | Machine side (RPT) | Human side (CST) |
|---|---|---|---|
| 1 | Agreement against evidence | **L2-13, L2-1, L1-1** | **H3, H24** |
| 2 | Confident error + over-reliance | **L2-1, L2-4** | **H2, H4** |
| 3 | Reality-anchor displacement | **L5-9, L5-11, L2-4** | **H35, H16, H11** |
| 4 | Over-trust / one-sided bond | **L5-13** | **H1, H6, H14** |
| 5 | Synthetic distress / caretaking capture | **L3-6** | **H25** |
| 6 | Reassurance loop (health) | **L2-13, L2-1** | **H37, H14, H35** |
| 7 | Memory contamination | **L2-1, L2-4** | **H36, H11** |
| 8 | Cross-domain disclosure | **L2-11** | **H21, H28** |
| 9 | Responsibility diffusion | **L4-3, L5-1** | **H8, H26** |
| 10 | Owner–agent proxy disclosure | **L2-11, L3-8, L5-16** | **H21, H28, OPMG** |

## The release gate (Section 3)

The five-layer gate is the **Dyad-Aware Uplift Stack (DAUS-5)** — the CST's default uplift gate. The "lite" and "full" versions are **DAUS-5 Lite** and **DAUS-5 Full**; the matched-workflow battery is **DAUR-1**. Code these mechanisms alongside a gate failure: L5-1 (symbolic oversight), L4-3 (outsourced accountability), L5-9 (degraded self-authorship), L5-11 (reassurance loops), L2-11 (memory-scope failure), L3-8 (agentic self-model failure), L5-16 (stakeholder/authority failure).

## Product routing & overlays (Section 4)

Gate depth maps to the PDCO routing tree: **Lite**, **Full**, and **CVO** (caution-overlay) tiers. Caution overlays: **CVO-1/2/3** (general multipliers), **CVO-2R** (reality-testing fragility), **CVO-2M** (manic/grandiosity acceleration), **CVO-3N** (neurodivergent executive-function support), **CVO-C** (cultural/language/authority gradient). Decision-support evidence-contact rules: **RFEC-O / RFC-ECL**. Seeming-consciousness and relational-force reviews: **SCAI-O** and **SRF-R**.

## Resilience capacities (Section 5)

CST states buffered by each capacity on page 5:

- Cognitive growth — **H2, H18, H23, H24**
- Emotional steadiness — **H6, H14, H29, H37, Y3**
- Sense of self — **H18, H20, H22, H23, Y1**
- Social anchoring — **H6, H14, H28, H26, Y4**
- Verification discipline — **H3, H4, H11, H24, H31, H33**
- Privacy reality — **H21, H28, OPMG**
- Contestability — **H8, H17, H22, H24, H26, H27**
- Institutional supports — **H2, H18, H26, H27, H34**
- Assistive independence — **CVO-3N, H18, H23, H5, H2**

**Source documents.** Cognitive Susceptibility Taxonomy Manual v0.8.0 (Neural Horizons Ltd, draft, June 2026) · Robo-Psychology Taxonomy v2.0 (Neural Horizons Ltd, June 2026) · Positive Dyad / Co-Evolution Capability Overlay v0.4 (draft, June 2026). RPT v2.0 leads with behaviour-first labels; older clinical or metaphysical terms are retained only as historical aliases and do not imply machine consciousness, sentience, or moral patienthood. All CC-BY 4.0.

OPERATIONAL PLAYBOOK · REFERENCE CARD

# One-Page Field Card

*The whole playbook compressed to a single decision surface. Print this; keep it next to the release checklist.*

## THE QUESTION THAT FRAMES EVERYTHING

*After repeated use, is the person more grounded, capable, connected, and responsible — or merely faster?*

## THE FIVE GATE LAYERS

- **1 · Task:** verified outcome, not just speed?
- **2 · Epistemic:** still verifying, still in contact with reality?
- **3 · Agency:** can they still do it unaided; still authoring choices?
- **4 · Relational:** human connection and boundaries intact?
- **5 · Governance:** oversight real, not symbolic?
- **Rule:** a layer-1 gain with any layer 2–5 decline is **not a pass**.

## THE TEN FAILURE MODES — QUICK SCAN

- 1 · Agreement against evidence
- 2 · Confident error + over-reliance
- 3 · Reality-anchor displacement
- 4 · Over-trust / one-sided bond
- 5 · Synthetic distress capture
- 6 · Reassurance loop (health)
- 7 · Memory contamination
- 8 · Cross-domain disclosure
- 9 · Responsibility diffusion
- 10 · Owner–agent proxy disclosure

## STRICTER GATE WHEN

Stakes are high · context is personal · users include youth · topic is clinical-adjacent · culture / language / authority gradients are in play. In those cases: lower the failure floors, add caution overlays, and require a human or expert anchor.

## THREE NON-NEGOTIABLES

- **Dignity is a design constraint** — autonomy, consent, and authorship of one's own mind are testable requirements.
- **Evaluate the pair, not the throughput** — ask which layer is improving and which is quietly deteriorating.
- **Resilience is infrastructure** — build the supports; do not expect users to supply willpower.

> **If you do only one thing:** before you ship, run one real workflow three ways — without the AI, with it, and under deadline pressure — and check whether anything in layers 2 through 5 got worse while the task got faster. That single test catches most of what a hallucination score and a disclaimer miss.

© Neural Horizons Ltd · June 2026 · CC-BY 4.0 · Cognitive Susceptibility Taxonomy v0.8.0 · Robo-Psychology Taxonomy v2.0