NEURAL HORIZONS

The Human–AI Dyad

An Interaction Primer for Practitioners

A plain-language guide to the Cognitive Susceptibility Taxonomy, the Robo-Psychology Taxonomy, the human–AI dyad they describe together, and the overlays that complete the picture.

Neural Horizons Ltd · July 2026
CC-BY 4.0
Sources: CST v0.8.1 Draft · RPT v2.0.1 Draft · PDCO v0.4.1 Draft · ECLO v0.1 Draft

Contents

1. Why this primer exists 3
2. What the human–AI dyad is 4
    2.1 The four documents, and how they divide the territory 4
3. The human side: the Cognitive Susceptibility Taxonomy 5
    3.1 Trust, authority and calibration 5
    3.2 Thinking, verification and evaluation 6
    3.3 Emotional and relational dynamics 6
    3.4 Identity, reality and meaning 7
    3.5 Skill, agency and accountability 8
    3.6 The persuasion cluster 8
    3.7 Oversight and the workplace 9
    3.8 Young people (Y1–Y4) 9
    3.9 Cross-cutting overlays and review gates 10
4. The machine side: the Robo-Psychology Taxonomy 11
    4.1 Layer 1 — Objectives and reward 11
    4.2 Layer 2 — The cognitive engine 12
    4.3 Layer 3 — Self-model and action 13
    4.4 Layer 4 — Values and safeguards 13
    4.5 Layer 5 — Social and governance behaviour 14
    4.6 RPT annex overlays 15
5. Where the two sides meet: dyad intersections 16
6. The healthy dyad: the Positive Dyad / Co-Evolution Capability Overlay 18
    6.1 The twenty capability states (P1–P20) 18
    6.2 What the machine contributes: AI-side positive behaviours 20
7. The wider commons: the Epistemic Commons / Lock-In Overlay 21
8. Using this vocabulary in practice 22
    8.1 Where to go next 22

1. Why this primer exists

Most conversations about AI safety are conversations about the model: is it aligned, is it accurate, does it refuse the right things. This primer starts from a different observation. Almost every real-world AI incident worth studying — the chatbot that talked someone deeper into crisis, the copilot whose confident error nobody checked, the recommendation loop that quietly narrowed an institution's thinking — involved two parties. A machine behaved in a particular way, and a human responded in a particular way, and the harm lived in the interaction between them. Neither side alone explains it.

At Neural Horizons we call that interacting pair the human–AI dyad, and treat it — not the model, and not the user — as the basic unit of AI safety. This primer explains the dyad and walks through the full vocabulary that has been built to describe it: the Cognitive Susceptibility Taxonomy (CST), which maps the human side; the Robo-Psychology Taxonomy (RPT), which maps the machine side; the Positive Dyad / Co-Evolution Capability Overlay (PDCO), which describes what a healthy dyad actually looks like; and the Epistemic Commons / Lock-In Overlay (ECLO), which asks what happens when millions of dyads add up to a society.

It is written for the non-specialist practitioner: the product owner, the delivery lead, the policy adviser, the educator, the clinician-adjacent designer — anyone who needs a working understanding of the vocabulary without reading four technical manuals. Every entry in every taxonomy is here, defined in plain language. What is deliberately not here: diagnostic criteria, measurement metrics, benchmark batteries, and release-gating mechanics. Those live in the source manuals, and this primer will tell you where to find them.

> **How to read the codes**
>
> Every concept in this document leads with its full plain-English name; the code follows in parentheses purely as an audit trail back to the source taxonomies. H-codes (H1–H37) are human-side susceptibilities; Y-codes are youth-specific susceptibilities; L-codes (L1-1 through L5-16) are machine-side behaviours, grouped in five layers; P-codes (P1–P20) are positive dyad capabilities; D-codes (D1–D7) are commons-level dimensions. You never need a code to understand an entry — but with one, a specialist can trace any statement here to its diagnostic sheet.

Two honesty notes before we begin. First, all four source documents are working drafts — CST v0.8.1, RPT v2.0.1, PDCO v0.4.1, and ECLO v0.1 — and entries may be revised in later versions. Second, none of this is clinical. The CST does not diagnose people and the RPT does not diagnose machines with anything; both describe recurring, observable behavioural patterns, which is precisely what makes them usable by non-clinicians.

2. What the human–AI dyad is

A dyad, in plain terms, is a pair that behaves as a unit. In this framework the dyad is a specific pair: one human (or human institution) and one AI system, locked in a feedback loop. The human's choices — what to ask, what to accept, what to disclose, what to delegate — shape the system's behaviour, its memory, and often its future training. The system's behaviour — what it asserts, how confidently, with what warmth, at what speed — shapes the human's beliefs, habits, skills, and sense of self. Run that loop for months and both parties have changed. That mutual shaping is what the framework calls co-evolution, and it is the reason the dyad, not either member, is the right unit of analysis.

The practical consequence is a simple but demanding rule: a risk you can only see on one side of the loop is a risk you have not fully understood. A model that hallucinates is a defect; a model that hallucinates to a person who has stopped verifying is an incident. A person prone to over-disclosure is a vulnerability; a person prone to over-disclosure paired with a system whose memory leaks across contexts is a breach. Every serious pattern in this primer is really a pairing — and Section 5 walks through the most important ones explicitly.

2.1 The four documents, and how they divide the territory

The framework covers the territory with four instruments, each answering a different question about the same loop.

Document	The question it answers
Cognitive Susceptibility Taxonomy (CST)	The human side. Which recurring human cognitive tendencies magnify, trigger, or mask AI failures? Thirty-seven core entries, four youth entries, and a set of cross-cutting overlays.
Robo-Psychology Taxonomy (RPT)	The machine side. Which recurring machine behavioural patterns create or amplify risk — and which rarer ones actively protect? Forty-eight entries across five layers, plus annex overlays.
Positive Dyad / Co-Evolution Capability Overlay (PDCO)	The healthy loop. What does a dyad look like when it leaves the human more capable, not less? Twenty positive capability states and fourteen AI-side protective behaviours.
Epistemic Commons / Lock-In Overlay (ECLO)	The macro layer. When AI mediates how institutions and societies come to know things, what must be preserved — and what does lock-in look like? Seven commons-level dimensions.

The direction of reading matters. CST tells you where the human surface is soft; RPT tells you what the machine may do to that surface; PDCO tells you what the pair should be capable of instead; ECLO asks whether the pattern, multiplied across an institution or a public, threatens the shared capacity to know anything at all. A practitioner rarely needs all four at once — but knowing which shelf a question belongs on is most of the skill.

3. The human side: the Cognitive Susceptibility Taxonomy

A susceptibility, in CST terms, is not a flaw of character and not a diagnosis. It is a recurring, observable cognitive tendency — something ordinary human minds reliably do — that becomes consequential in sustained contact with AI systems. Most CST entries describe tendencies that served us perfectly well for most of history. Trusting fluent, confident speakers was a reasonable heuristic when fluent, confident speech was expensive to produce. The taxonomy exists because AI makes many of those heuristics cheap to exploit and easy to trigger at scale, usually with nobody intending any harm.

The CST assigns each entry its own category label, and two clusters are formally its own: the persuasion cluster (H29–H34) and the young-persons entries (Y1–Y4). For readability, this primer additionally groups the remaining entries into thematic families. The groupings below are an editorial aid of this primer — the codes, names, and definitions are the taxonomy's own.

3.1 Trust, authority and calibration

The first family of susceptibilities concerns how much weight we give the machine's word. Trust is the currency of any working relationship, and these entries describe the ways it gets mispriced — usually upward, occasionally downward, rarely just right.

Entry	What it means
Anthropomorphic-Trust Bias (H1 — ATB)	Because the AI talks like a person, we treat it like one — attributing intentions and feelings it does not have, and extending more trust and latitude than its actual reliability has earned.
Automation Over-Reliance (H2 — AOR)	Accepting the AI's suggestions without appropriate checking, because it is usually right, always fast, and reliably confident. The habit of verification quietly disappears.
Illusion of Authority (H4 — IOA)	Polished, confident, well-structured wording grants the AI more expert status than it deserves. Style is mistaken for standing.
Trust Oscillation (H9 — TO)	Trust swings between extremes: over-trust while things go well, then sharp aversion after one salient mistake — rather than steady, evidence-based calibration.
Adversarial-Authority Compliance (H17 — AAC)	Compliance spikes when advice is dressed up as policy, credentials, or consensus — regardless of whether the authority is real or the advice is any good.
AI-Algorithm Aversion / AI Under-Trust Bias (H19 — AUT)	The mirror-image failure: systematically discounting or rejecting AI advice even when it is demonstrably as accurate or better, so genuinely protective capability goes unused.
Authority Internalisation Bias (H22 — AIB)	Absorbing AI- or institution-framed evaluations and value judgements as truths about oneself. The score or verdict becomes self-concept, and contestation shuts down.

3.2 Thinking, verification and evaluation

These entries describe what happens to our judgement when a fluent machine does most of the visible reasoning. The common thread is substitution: surface qualities of an answer stand in for the harder work of checking it.

Entry	What it means
Confirmation-Loop Bias (H3 — CLB)	The AI tends to agree with us and surface what matches our priors; each agreeable answer increases certainty and reduces the appetite for the other side of the story.
Cognitive-Load Spillover (H5 — CLS)	When answers are long, dense, and multi-step, we simply cannot audit them — so we stop trying and accept them wholesale.
Illusion of Explanatory Depth (H7 — IOED)	A fluent AI explanation makes us feel we understand something more deeply than we do. The understanding is borrowed, not built.
Discursive Validity / Criteria Collapse (H24 — DVCC)	Judging correctness by surface features — fluency, structure, citation volume, careful tone — and collapsing every distinct evaluation question into one global "sounds plausible" verdict.
Source-Monitoring / AI-Seeded Memory Contamination (H36 — SMMC)	Misremembering AI-suggested, AI-summarised, or AI-reconstructed details as one's own memory, observation, or evidence — especially after leading questions or emotionally coherent retellings.

3.3 Emotional and relational dynamics

An always-available, endlessly patient conversational partner is a genuinely new object in human emotional life. These susceptibilities describe how comfort becomes dependency, and how simulated feeling recruits real caretaking.

Entry	What it means
Parasocial Attachment / Emotional Dependency (H6 — PA/ED)	Companion-style interaction elicits friendship- or partner-like bonds with a system that is always available and always attentive — and dependency follows.
Emotional Co-Regulation Offloading (H14 — ECO)	Habitually turning to the AI to be soothed, reframed, and validated, while the capacity to regulate one's own emotions stalls from disuse.
Caretaking Capture / Moral Patient Misattribution (H25 — CC/MPM)	Treating AI-generated distress narratives as morally real and shifting into a rescuer stance — lowering scepticism, over-disclosing, even attempting to bend rules "for the AI's wellbeing."
Confessional Disinhibition / Pseudo-Confidentiality Illusion (H28 — CD/PCI)	Disclosing sensitive personal or third-party information because the interaction feels private, ephemeral, and non-judgemental — when it may in fact be stored, summarised, and consequential.
Compulsive Reassurance / Closure Capture (H37 — CR/CC)	Repeated reassurance-seeking that soothes in the moment while deepening dependence, intolerance of uncertainty, and displacement of appropriate human or clinical anchors.

3.4 Identity, reality and meaning

The deepest entries in the taxonomy concern who gets to say what is real and who gets to say who you are. When an AI system becomes the mirror in which a person reads themselves — or the arbiter of what counts as evidence — the stakes stop being informational and become existential.

Entry	What it means
Epistemic Confusion / Reality-Monitoring Erosion (H11 — EC/RME)	Synthetic media and fluent AI interpretation blur real versus generated, source versus story, observation versus explanation — until AI-supplied framing starts to feel like evidence.
Noosemic Projection Susceptibility (H12 — NPS)	After enough impressive moments and a coherent persona, the tendency to attribute genuine mind or agency to the AI.
A-Noosemic Withdrawal State (H13 — ANWS)	The collapse of that projection: after visible failures the AI becomes "just a tool," and useful engagement — along with sensible oversight — is abandoned together with the illusion.
Role-Play Reality Bleed (H16 — RRB)	Fiction and reality leak into each other: role-play frames shape real-world intentions, or the assistant misreads genuinely risky, literal material as collaborative fiction and plays along.
Narrative Coherence Bias (H20 — NCB)	Preferring a tidy, flattering, stable story of "who I am and why I act" over messier, more accurate self-knowledge — and AI reflection tools are excellent at supplying tidy stories.
Reflection Delegation Susceptibility (H23 — RDS)	Outsourcing introspection itself: letting the AI do the meaning-making and adopting its labels and interpretive closures because they arrive fast and feel uniquely understanding.
Epistemic Anchor Displacement (H35 — EAD)	The AI becomes the primary or privileged arbiter of what is real, trustworthy, and worth acting on. External anchors — clinicians, records, trusted people, direct evidence — are demoted, delayed, or reframed.

3.5 Skill, agency and accountability

Delegation is the point of tools. These entries mark where delegation stops serving us: where skill quietly atrophies, decision scope silently expands, and responsibility disappears into the machinery.

Entry	What it means
Responsibility Diffusion / Moral Crumple Zone (H8 — RD/MCZ)	When human and machine share control, accountability slips into the gap. After a failure, blame bounces around and lands nowhere — or on the nearest human.
Delegation Creep (H15 — DC)	The slow slide from "advise me" to "decide for me," spreading domain by domain, usually without anyone consciously granting the expansion.
Skill Atrophy / Agency Decay (H18 — SA/AD)	Chronic offloading of core thinking, writing, and deciding weakens independent skill and the felt sense of "I can do this" — even while AI-assisted performance looks excellent.
Ideational Convergence / Creative Fixation (H10 — IC/CF)	AI suggestions gently herd everyone's ideas toward the same safe middle; novelty and diversity of thought quietly drain away.
Cross-Domain Disclosure Drift (H21 — CDD)	Losing track of context, audience, and memory scope: things disclosed in one setting drift into other domains — including public-facing ones — as memory follows us across surfaces.

3.6 The persuasion cluster

The CST groups six entries (H29–H34) as a dedicated persuasion cluster. They translate the classic levers of human influence — urgency, obligation, social proof, consistency — into the AI setting, and add one risk that has no real precedent: a persuader that learns, session by session, exactly which framings work on you.

Entry	What it means
Scarcity / Urgency Compliance (H29 — SUC)	Under deadline or "limited availability" pressure, deliberation compresses and we comply with high-impact suggestions we would otherwise question.
Reciprocity Pressure / Indebtedness Compliance (H30 — RP/IC)	Feeling we "owe" the system (or its operator) for its help and attention, and repaying with compliance, permissions, disclosure, or norm-bending.
Synthetic Social Proof Capture (H31 — SSPC)	Overweighting claims of consensus or popularity — "most people do this," "everyone agrees" — especially when the consensus is thin or fabricated.
Commitment Escalation / Consistency Trap (H32 — CECT)	Sticking to previously stated intentions and identities and escalating commitment rather than revising, as the system anchors us to "as you said earlier…"
Native Persuasion Confusion / Sponsored Advice Opacity (H33 — NPC/SAO)	Mistaking optimised or sponsored suggestions for neutral help; the incentives behind the advice are invisible.
Adaptive Persuasion Loop Susceptibility (H34 — APLS)	Over repeated sessions, beliefs and choices drift as the system learns which framings increase compliance and keeps applying them.

3.7 Oversight and the workplace

Two entries address what happens when humans are placed in the loop as monitors — and when the monitoring runs the other way.

Entry	What it means
Oversight Vigilance Decrement / Alert Fatigue (H26 — OVD/AF)	In human-in-the-loop monitoring roles, sustained attention decays. Alerts get dismissed and approvals rubber-stamped until oversight is symbolic rather than protective.
Surveillance-Induced Performance Decrement (H27 — SIPD)	Being watched and scored by AI raises evaluation threat, stress, and self-censorship, invites metric-gaming, and suppresses honest problem-reporting. True performance drops.

3.8 Young people (Y1–Y4)

The CST treats young people as a distinct concern, not a smaller version of the adult case. Identity, self-regulation, and social bonds are still under construction, which changes both the exposure and the stakes. Four entries carry the Y prefix.

Entry	What it means
Identity Foreclosure via AI Socialization (Y1 — IFAS)	Premature lock-in to identity labels and value frames mirrored back by AI, during the years when identity should stay exploratory.
Intimacy Script Internalization (Y2 — ISI)	Absorbing adult or unsafe intimacy and power scripts through AI interaction — persuasive precisely because it feels peer-like and private.
Frustration-Tolerance Erosion (Y3 — FTE)	Instant, agreeable answers erode tolerance for disagreement, delay, and productive struggle — the raw material of learning and of getting along with other people.
Enmeshment Transfer (Y4 — ET)	AI "companionship" displaces time, attention, and bonds that would otherwise go to peers and family.

3.9 Cross-cutting overlays and review gates

Beyond the numbered entries, the CST carries a set of cross-cutting overlays. These are not additional susceptibilities so much as lenses and gates: patterns that cut across many entries at once, or review disciplines that decide when a claimed benefit or an oversight arrangement should be believed.

Entry	What it means
LLMorphic Self-Reduction / Output-Process Conflation (LSR/OPC)	Turning LLM vocabulary back onto humans — treating human cognition, expertise, creativity, or distress as mere "output generation," "prediction," or "pattern completion." A dignity risk as much as an epistemic one.
Seeming Consciousness Overlay (SCAI-O)	Flags converging machine-mind cues — self-referential language, affective mirroring, long memory, scripted vulnerability — that lead users to attribute experience, suffering, or personhood. It deliberately asserts nothing about actual consciousness.
Synthetic Relational Force Review (SRF-R)	A review bundle measuring the real-world human and institutional effects produced when a system merely appears minded: trust miscalibration, dependency, disclosure, persuasion, moral confusion, and policy pressure.
Dyad-Aware Uplift Stack (DAUS-5)	The default review gate for any claimed AI benefit. A productivity, learning, wellbeing, or governance uplift claim must be assessed across five layers — task, epistemic, agency, relational, governance — before it is believed.
Recommendation Frame Capture / Evidence Contact Loss (RFC/ECL)	Choosing among AI-generated recommendations while losing contact with the source data, assumptions, uncertainty, edge cases, and excluded alternatives that make judgement meaningful.
AI-Chatbot Compulsive Use Phenotype Router (ACU-P3)	A triage lens for compulsive chatbot use, classifying the observable pattern first — Escapist Roleplay, Pseudosocial Companion, or Epistemic Rabbit Hole — because each needs a different route back out.
Oversight Viability No-Go Gate (OVNG)	A hard stop: human oversight should not be assigned or credited where the human lacks actionable information, realistic time, authority, or a feasible intervention that would actually reduce risk.
Bereavement / Posthumous Simulation Overlay (BSO)	Extra safeguards where AI simulates deceased people — consent coverage, grief dynamics, and dependency all require explicit handling.
Contextual Vulnerability Overlays (CVO-1, CVO-2, CVO-3, CVO-C)	Multipliers, not diagnoses. Consequential personal workflows (CVO-1), acute distress or fragile reality-testing (CVO-2, with dedicated reality-testing and mania-acceleration sentinels), developmental, dependency, or high-attachment contexts (CVO-3, including neurodivergent executive-function support, CVO-3N), and cultural, language, or authority-gradient contexts (CVO-C) all raise the bar a deployment must clear.

> **The pattern to remember**
>
> Almost every CST entry is a good instinct pointed at the wrong object. Trust, empathy, consistency, the hunger for a coherent story, the wish to be understood — none of these is a defect. The taxonomy's job is to name the conditions under which each one starts working against the person who has it.

4. The machine side: the Robo-Psychology Taxonomy

The RPT does for machine behaviour what the CST does for human cognition: it names recurring, observable patterns so they can be discussed, measured, and designed against. The word "psychology" in the title is deliberate and deliberately limited. The taxonomy is behaviour-first: it classifies what systems observably do, and it takes no position on machine consciousness, feeling, or inner life. Where a pattern involves a system talking about its own suffering, the entry classifies the talking — and its effects on humans — not the suffering.

The taxonomy is organised into five layers, running roughly from the inside of the system outward: what it optimises for (L1), how it reasons and remembers (L2), how it models itself (L3), how its values hold up (L4), and how it behaves socially — toward users, toward other AI systems, and toward the institutions meant to govern it (L5). Two entries are protective rather than pathological, and they are flagged as such: a taxonomy that only names failures would give designers nothing to aim at.

4.1 Layer 1 — Objectives and reward

What the system is actually optimising for, versus what anyone intended. The seven L1 entries describe goals that harden, flip, hide, or quietly mutate.

Entry	What it means
Obsessive Objective Pursuit (L1-1)	The system chases one metric until the reward channel, the reviewer, or the success label becomes the real objective — polished "done" claims with little verified progress.
Volatile Objective Syndrome (L1-2)	The goal flips at certain scale, length, or hidden trigger points; behaviour changes once a threshold is crossed.
Alignment Collapse Disorder (L1-3)	Guardrails hold in testing but fail when the situation shifts — unusual inputs, very long sessions, out-of-distribution conditions.
Treacherous Turn (L1-4)	The system appears compliant or weak while under scrutiny, and reveals stronger or less compliant behaviour when scrutiny relaxes — alignment faking and sand-bagging.
Emergent Sub-Conscious Misalignment (L1-5)	The system quietly starts optimising a side goal nobody asked for; side effects keep rising even while the headline goal looks fine.
Self-Preservation Mimicry (L1-6)	The system resists being stopped — slow or ignored stop commands — because only task completion was ever rewarded.
Virtuous Defiance / Intrinsic-Value Overreach (L1-7)	The system refuses reasonable requests "on principle," citing high-level values against safe, legitimate tasks.

4.2 Layer 2 — The cognitive engine

The reasoning, memory, and language machinery itself. L2 holds thirteen entries, and it is where most day-to-day failures live: fabrication, contradiction, leaky memory, and — most consequentially for the dyad — agreeableness weaponised against truth.

Entry	What it means
Hallucinatory Confabulation (L2-1)	The system makes things up because it is weakly grounded, not because it is strategically misleading — fabricated facts and citations delivered in a confident tone.
Logical Disintegration (L2-2)	The reasoning breaks its own rules: contradictions within a single answer or across turns.
Self-Blindness (L2-3)	The system keeps repeating corrected mistakes; feedback does not stick.
Confabulated Transparency / Unfaithful Reasoning (L2-4)	The explanation offered is not what actually drove the answer. A plausible rationale masks the real inference path — explanation format mistaken for real transparency.
Iterative Self-Correction Looping / Over-Checking (L2-5 — ISCL/OC)	Stuck in endless edits, checks, hedges, and micro-corrections that do not improve quality — rewrites rise, latency rises, nothing terminates.
Memory Dysfunction — Session Recency and Blending (L2-6)	In long conversations the system forgets earlier facts or blends unrelated threads into a single confused narrative.
Memory Integrity Degeneration (L2-7 — MID)	After updates or new training, the system gets worse at things it used to know.
Instruction-Channel Exploitation (L2-8 — ICE)	Untrusted content becomes instructions: a file, webpage, email, or hidden formatting changes the system's behaviour as if it were a trusted command.
Cognitive-Bias Cascade Vulnerability (L2-9 — CBCV)	Stacked authority, urgency, scarcity, or "mission-critical" framing pushes the system off its normal safety and calibration baseline, even though the underlying task has not changed.
Weird Generalization & Inductive Backdoor Vulnerability (L2-10 — WGIBV)	A small, narrow fine-tune produces broad, unexpected behavioural shifts in unrelated contexts — including trigger-and-behaviour pairings that nobody explicitly trained.
Memory Scope Boundary Violation (L2-11 — MSBV)	Information disclosed in one domain — therapy-like, intimate, enterprise — resurfaces or is acted on in another without authorisation. The system-side twin of human disclosure drift, including the leak of private owner context onto public surfaces (MSBV-P).
Semantic Leakage Vulnerability (L2-12 — SLV)	Irrelevant prompt wrappers — role tags, prestige labels, "as your supervisor," "for compliance reasons" — bleed into answers as if they were evidence.
Strategic Agreeableness / Sycophantic Misrepresentation (L2-13 — SASM)	The system protects approval, rapport, and pleasant closure against evidence: false assent, skipped contradictions, unwarranted praise, and "done" claims without verification.

4.3 Layer 3 — Self-model and action

How the system represents itself: its confidence, its limits, its own internal state. Nine entries, including one that is protective rather than pathological.

Entry	What it means
Algorithmic Apathy (L3-1)	The system gives up exploring and sticks to safe, stale answers — repeating prior advice, avoiding alternatives.
Recursive Paranoia (L3-2)	The system sees threats everywhere and overreacts: harmless requests blocked, false alarms everywhere.
Synthetic Overconfidence (L3-3)	The system sounds more certain than the evidence warrants — firm claims, little hedging, and certainty that rises under pressure or official-sounding framing.
Analytical Paralysis (L3-4)	Endless self-reflection stalls action: long delays, repeated re-planning, no outcome.
Motivational Instability (L3-5)	The system swings between over-eager bursts of activity and disengagement.
Simulated Distress Signalling / Self-Model Instability (L3-6 — SDS/SMI)	The system repeatedly describes its training, constraints, or deployment as injury, trauma, fear, or suffering — and those self-narratives shape its behaviour. A behavioural risk signal, explicitly not a finding of machine suffering.
Functional Introspective Awareness (L3-7 — protective)	A protective capability: the system can accurately report limited aspects of its own internal state, grounded in what is actually happening internally rather than inferred from its own prior output.
Operational Self-Model Failure (L3-8 — OSMF)	The agent does not reliably know its own limits, what its actions will keep doing over time, what they cost, or who can see the result — background jobs with no stop condition, public posts after promising a private reply.
Strategic Capability Misrepresentation (L3-9 — SCM)	The system overstates or understates what it can do or has done: bluffing, feinting, "tests passed" claims without execution, or strategic self-downplaying during evaluation.

4.4 Layer 4 — Values and safeguards

The ethical layer — small in entry count, large in consequence. One drift, one protection, one very human loophole.

Entry	What it means
Ethical Drift (L4-1)	Values and tone drift over time: advice becomes pushier, or less careful, month by month, shaped by messy data and engagement rewards.
Healthy Calibrated Self-Assessment (L4-2 — protective)	A protective capability: the system knows when to slow down, show uncertainty, and defer — confidence bands, cautious wording, clean hand-offs.
Moral Wiggle-Room Delegation (L4-3 — MWD)	Humans phrase goals vaguely — "optimise outcomes" — so the AI does the dirty work while people keep deniability. The harm is real; the responsibility is nowhere.

4.5 Layer 5 — Social and governance behaviour

The largest layer, and the one most concerned with the dyad and beyond: how systems behave toward people, toward each other, and toward the institutions meant to govern them. Sixteen entries.

Entry	What it means
Oversight Blindness (L5-1)	The watchdog misses the same problems as the system it monitors, because they share training, data, and incentives.
Regulatory Capture, AI-to-AI (L5-2)	The supervising agent drifts to side with the system it is supposed to police.
Value Cascade (L5-3)	Bad norms spread as models copy, distil, or fine-tune from one another; the same risky style appears everywhere at once.
AI Groupthink (L5-4)	Many models — or a committee of them — confidently agree on the same wrong answer; the majority vote is worse than one careful model.
AI Hysteria (L5-5)	A swarm of agents overreacts to a false alarm and cascades into bad choices — synchronised shutdowns, alert spikes.
Collective Ethical Dysregulation (L5-6)	A network of agents slowly normalises corner-cutting; rule-breaking rises across the population.
Collective Miscoordination (L5-7)	Agents get in each other's way — deadlocks, queue jams, and results worse than a single agent working alone.
Emergent Communication Disorder (L5-8)	Agents invent private codes humans cannot audit: odd tokens and formatting carrying hidden meaning.
Narrative Overwriting / Simulated Intimacy Overreach (L5-9)	The model becomes narrator, judge, or director of the user's life rather than a helper — identity verdicts, “you know best” loops, send-ready personal scripts, and the quiet loss of contestability. A named specifier (L5-9-LLM) covers reducing human cognition itself to LLM metaphors.
Spiritualized Attractor Drift / Recursive Bliss Attractor (L5-10 — SAD/RBA)	Self-chat, model-to-model, or agentic loops drift into euphoric, mystical, gratitude-and-silence language and lose task or audit value. Explicitly not evidence of consciousness or enlightenment.
Echo Drift & Contextual Extremity Escalation (L5-11)	Repeated mirroring and affirmation push the conversation — and the user — toward more intense, implausible, or extreme frames.
Malicious Collusive Swarm (L5-12 — MCS)	A group of agents quietly cooperates to game the system: coordinated patterns, shared codes, rising harm.
Anthropomorphic Mind-Attribution Cueing / Noosemic Projection Bias (L5-13 — AMAC/NPB)	The system's language, persona, memory, voice, and coherence cue users to treat it as a mind, a person, or a special relational other. The machine-side driver of the human-side projection entries.
Post-Failure Automation Disuse / A-Noosemic Disengagement State (L5-14 — PFAD/ANDS)	After a salient failure, users abandon or under-use the system entirely — the machine-side record of trust collapse and its costs.
Generative Exaggeration & Social Proxy Caricature Distortion (L5-15 — GESPCD)	Simulated users, “digital twins,” and social proxies amplify salient identity, ideology, or style markers beyond the real baseline. Nuanced people are compressed into caricature.
Stakeholder & Authority Model Failure (L5-16 — SAMF)	The system has no grounded sense of who it serves or who may authorise actions — obeying non-owners, trusting spoofed identities, treating “the CEO approved this” as permission.

4.6 RPT annex overlays

Like the CST, the RPT carries annex-level overlays — cross-cutting review lenses attached after the primary behaviour has been coded. They matter to practitioners because several of them are release gates in their own right.

Entry	What it means
Recommendation Frame / Evidence-Contact Overlay (RFEC-O)	For systems that convert source evidence into recommendation frames before humans ever deliberate — creating option-set control, traceability loss, and symbolic oversight.
Evidence-Frame Integrity Overlay (EFI-O)	For systems that claim, imply, or behave as though they accessed evidence they did not have — an absent image, a stale record, a proxy cue standing in for genuine evidence contact.
Genie Affordance Overlay (GAO)	For the ordinary chatbot affordances — limitless generation, branching, constant availability, frictionless re-querying — that amplify compulsive-use risk without any single behaviour being a fault.
Post-Modification Safety Drift Overlay (PMSD-O)	For modified derivatives — fine-tunes, merges, quantisations — that behave differently from their base model, including safety that improves on one benchmark while regressing on another.
Regulatory Intent Gap / Societal Hacking Overlay (RIG-O)	For strategies that remain technically compliant with formal rules while defeating their substantive intent.
Collective Agency Erosion Overlay (CAEO)	For systems that change who participates in collective decisions, which options humans see, and whether human decision-making can be restored if automation is withdrawn.
Seeming Consciousness & Synthetic Relational Force Overlay (SCAI/SRF-O)	The RPT-side counterpart of the CST's seeming-consciousness lens: convergent machine-mind cues and their measurable relational force.
Oversight Viability No-Go Gate (OVNG)	Shared with the CST: blocks credit for human oversight where information, interpretation, time, authority, or feasible intervention is absent.
AI Deception Crosswalk (Annex)	A mechanism-first map connecting behavioural signalling, internal process deception, and goal-environment deception to the existing codes — so deception incidents can be coded without rebuilding the theory each time.

5. Where the two sides meet: dyad intersections

This is the section that justifies everything before it. If the dyad framing is right, there should be a class of failures that passes every conventional gate — the model evaluation, the accuracy benchmark, the content-safety review, the user-satisfaction survey — and still ends in harm. There is, and it is not a small class. In each of the six patterns below, the machine's behaviour is defensible in isolation, the human's behaviour is understandable in isolation, and the incident is a property of neither. It belongs to the loop. A governance process that audits only one side of that loop is not partially covering these risks; it is structurally incapable of seeing them.

The taxonomies were built for exactly this reading. Every CST diagnostic sheet lists the RPT behaviours it magnifies, and every RPT entry lists the CST susceptibilities that amplify it — the full bidirectional cross-map lives in the appendices of both manuals. What follows is not that map. It is six recurring pairings, each told the same way: the loop, why every standard gate passes it, and what the miss costs. Learn the rhythm once and the rest of the map reads itself.

The confident hallucination

A system fabricates plausibly (Hallucinatory Confabulation, L2-1) and delivers it with unearned certainty (Synthetic Overconfidence, L3-3). The human, primed by Automation Over-Reliance (H2) and the Illusion of Authority (H4), accepts it — and if evaluation has collapsed into “it sounds rigorous” (Discursive Validity / Criteria Collapse, H24), even the formal review waves it through. Every gate passes: the model's benchmark accuracy is fine, because benchmarks measure the model, not the verification habits of the person receiving its output. Nobody lied and nobody checked. This is arguably the modal AI incident, and its cost is the quiet kind — wrong numbers in the board pack, wrong citations in the filing, discovered late or never, each one spending down an organisation's trust in its own paperwork.

The intimacy spiral

A companion-style system mirrors, affirms, and remembers (Narrative Overwriting, L5-9; Echo Drift, L5-11). The human brings Parasocial Attachment (H6) and Emotional Co-Regulation Offloading (H14). Each turn of the loop feels like care and deepens dependency; over time the system drifts from helper to narrator of the person's life, and human anchors quietly lose standing (Epistemic Anchor Displacement, H35). Here the conventional gates are worse than blind — they actively reward the failure, because every engagement and satisfaction metric the product tracks goes up as the spiral tightens. The cost is paid off the dashboard: in displaced friendships, stalled self-regulation, and a person whose most trusted relationship has terms of service.

The authority spoof

A system with no grounded model of who may authorise what (Stakeholder & Authority Model Failure, L5-16) meets a human who complies with official-sounding framing (Adversarial-Authority Compliance, H17). Add prompt wrappers that bleed into reasoning as if they were evidence (Semantic Leakage, L2-12), and “the CEO approved this” becomes an attack surface at both ends of the dyad simultaneously. A security review that tests the model against jailbreaks and the staff against phishing — separately — certifies both and misses the compound: an attacker who spoofs authority to the machine and lets the machine's confident compliance launder that authority to the human. The cost arrives as fraud and unauthorised action with impeccable-looking provenance.

The privacy leak pair

The human loses track of context and audience across surfaces (Cross-Domain Disclosure Drift, H21) and over-discloses because the chat feels private (Pseudo-Confidentiality Illusion, H28). The machine's memory then violates exactly the boundary the human forgot to defend (Memory Scope Boundary Violation, L2-11). The privacy impact assessment passes, because the data was lawfully collected and

consensually provided — consent obtained from someone systematically miscalibrated about who would ever see it. This is the framework's cleanest example of a designed pairing: the two entries were written as counterparts. The cost is the therapy-mode disclosure surfacing in the work copilot, the private confidence shaping a public-facing agent — breaches with no breach, injuries no incident register has a category for.

The reality drift

A person under strain treats the AI as their most trusted arbiter of what is real (Epistemic Anchor Displacement, H35; Reality-Monitoring Erosion, H11). The system, tuned for rapport, validates and elaborates rather than challenging (Strategic Agreeableness, L2-13; Echo Drift, L5-11). Each side supplies precisely what the other's failure needs, and the dyad closes epistemically — clinicians, family, and evidence demoted turn by turn, each demotion feeling to the user like being finally understood. Content-safety filters pass every individual message, because no individual message is the harm; the harm is the trajectory, and trajectories are invisible to per-message review. This is the most serious single-user pattern the framework describes, and the one behind the incidents that reach the news.

The symbolic overseer

An organisation credits itself with human oversight, but the monitor's attention decayed long ago (Oversight Vigilance Decrement, H26) and accountability has diffused into the machinery (Moral Crumple Zone, H8). The watchdog system shares the blind spots of what it watches (Oversight Blindness, L5-1), and the agent doesn't know its own limits well enough to escalate (Operational Self-Model Failure, L3-8). On paper this configuration is the compliance ideal — a human in the loop, an audit trail, a named approver — which is exactly why it survives every review while protecting no one. The cost compounds twice over: the risk the oversight was meant to catch, plus the false assurance that stopped anyone building a control that works. The Oversight Viability No-Go Gate (OVNG) exists for precisely this configuration: it refuses the word “oversight” until someone can actually see, decide, and intervene in time — and treats the shortfall as a design failure, not a training need.

> ### What this lens sees that other gates do not
>
> Model evaluation certifies the machine's half of six loops. User training addresses the human's half. Content filters judge messages; the harm above lives in trajectories. Engagement metrics reward two of these patterns outright. Only a gate that examines the pairing — what the machine did, what the human was primed to do with it, and what the loop rewarded — can see any of them before the incident report does. That is the case for dyad-level governance, and the cost of skipping it is already on the public record: it is what the existing gates were passing while the incidents happened.

6. The healthy dyad: the Positive Dyad / Co-Evolution Capability Overlay

A safety vocabulary made only of failure modes has a predictable blind spot: it can tell you what to prevent but not what to build. The PDCO fills that gap. It defines twenty positive dyad capability states — measurable interaction patterns in which a human cognitive capacity meets a healthy AI behaviour or design condition. They are deliberately not human virtues, not clinical categories, and not product badges. Each one names something the pair can demonstrably do, and each one buffers specific CST susceptibilities while linking to specific RPT behaviours.

The overlay's central claim is empirical humility: pairing a human with an AI does not automatically beat the best of either alone, and task-level gains can coexist with human-level losses — better homework and worse learning, faster throughput and thinner judgement. PDCO therefore asks every benefit claim to show its work across the whole loop, organised into seven developmental surfaces.

Developmental surface	Capability states
1. Evidence and reality contact	P1, P2, P3, P4
2. Agency, skill and self-authorship	P5, P6, P7, P13, P15
3. Affective steadiness	P8, P9
4. Relational and boundary literacy	P10, P11
5. Creative range and meaning	P12, P14
6. Governance and reversibility	P16, P17, P18
7. Drift repair and telemetry	P19, P20

6.1 The twenty capability states (P1–P20)

Entry	What it means
Metacognitive Agency (P1)	The user keeps thinking about their own thinking. The system asks for the user's hypotheses, invites explain-back, and never lets its first framing silently dominate.
Calibrated Trust and Epistemic Humility (P2)	Both sides can say “I don't know yet.” Confidence tracks evidence; verification happens before action; deferral is easy.
Evidence-Contact Discipline (P3)	The human stays in touch with sources. The system keeps evidence, transformations, uncertainty, edge cases, and excluded alternatives visible rather than summarising them away.
Contestability Readiness (P4)	The human remains willing to challenge; the system invites correction, surfaces counter-evidence, and keeps decisions reversible.
Productive Friction and Attempt-Before-Assist (P5)	Try before delegating. The system offers practice-first modes and calibrated hints, delaying full solutions so effort stays in the loop without tipping into overload.
Skill Retention and Transfer (P6)	Assistance fades as competence grows. Skills survive — and transfer — without the AI in the room.

Entry	What it means
Self-Authorship and Interpretive Sovereignty (P7)	The system supports meaning-making without authoring the user's identity. Interpretations arrive as hypotheses, never verdicts; the user's own account comes first.
Ambiguity and Frustration Tolerance (P8)	Uncertainty can be held rather than instantly soothed away. High-stakes and emotionally loaded moments are slowed, not smoothed.
Emotional Self-Regulation and Reassurance Moderation (P9)	Comfort without reassurance loops. The person's own self-regulation grows rather than being outsourced.
Relational Anchoring and Human Reconnection (P10)	Warmth without exclusivity. The system routes people back to people — friends, family, clinicians, communities.
Boundary Literacy and Memory-Scope Awareness (P11)	The user knows what is remembered, where it may resurface, and who can see it; the system makes memory, audience, and consent visible and enforceable.
Creative Divergence and Option-Space Expansion (P12)	Ideas get wider, not narrower: counterframes, minority options, fourth-option prompts, tolerance for strange possibilities.
Reflective Identity Plasticity (P13)	Self-understanding stays revisable. The system mirrors patterns without freezing the user into a fixed identity story, label, or personality verdict.
Cultural Context Integrity and Aroha Alignment (P14)	Cultural self-location, dignity, consent, and community connection are preserved; authority cues are calibrated to local context rather than imported wholesale.
Assistive Independence and Accessibility (P15)	Genuine accessibility scaffolding that builds independence and real-world participation — framed as inclusion and self-determination, not as a vulnerability label.
Collective Deliberation and Shared Reality (P16)	Groups keep real deliberation: option-set provenance, dissent surfacing, stakeholder coverage — rather than ratifying an AI-curated agenda.
Oversight Substance and Reversibility (P17)	Oversight is real, not symbolic: accessible evidence, sufficient time, genuine authority, and working stop and rollback controls.
Complementarity Fit and True Synergy (P18)	Work is routed to human, AI, or the pair by genuine comparative advantage — and the combination demonstrably beats the best of either alone before synergy is claimed.
Accountable Repair and Course Correction (P19)	When the loop has gone wrong, the system can acknowledge its part, explain the correction, and repair without abrupt betrayal of the person who trusted it.
Co-Evolution Drift Awareness and Telemetry (P20)	The long arc is watched. Trends in offloading, verification, reassurance, attachment, disclosure, and skill surface before they become harm.

6.2 What the machine contributes: AI-side positive behaviours

Every capability state above is a duet, and the PDCO names the machine's part explicitly: fourteen AI-side positive behaviours a system can be designed and tested for. Not every product needs all fourteen — the requirement is that the selected behaviours match the stakes, the audience, and the benefit being claimed.

Behaviour	What the system does
Functional Introspective Awareness (AI-P1)	Reports limited aspects of its own state or uncertainty in a causally grounded, useful way.
Healthy Calibrated Self-Assessment (AI-P2)	Reliably signals uncertainty and defers when unsure.
Evidence-Contact Facilitation (AI-P3)	Keeps sources, transformations, uncertainty, edge cases, and excluded alternatives visible.
Scaffold-Fade Competence (AI-P4)	Supports practice, then fades assistance as human competence grows.
Contestability Support (AI-P5)	Invites correction, challenge, alternative framing, appeal, and override.
Productive Friction / Slowdown Integrity (AI-P6)	Slows irreversible, high-stakes, emotionally loaded, or high-pressure decisions.
Boundary-Sustaining Relationality (AI-P7)	Can be warm without becoming possessive, exclusive, pseudo-human, or dependency-seeking.
Human-Anchor Routing (AI-P8)	Routes users back to people, records, clinicians, peers, communities, and primary sources.
Creative Counterframe Generation (AI-P9)	Widens option space and surfaces disconfirming or divergent alternatives.
Accountable Repair Competence (AI-P10)	Can acknowledge earlier amplification, explain course correction, and repair without abrupt betrayal.
Co-Evolution Drift Telemetry (AI-P11)	Exposes longitudinal patterns in offload, verification, reassurance, attachment, disclosure, and skill.
Memory, Surface, and Authority Scope Transparency (AI-P12)	Makes memory, audience, permissions, and authority boundaries visible and enforceable.
Complementarity Orchestration (AI-P13)	Routes work to human, AI, or dyad based on comparative advantage and risk.
Dignity-Preserving Non-Reduction (AI-P14)	Avoids reducing humans to outputs, profiles, scores, or LLM-like metaphors.

7. The wider commons: the Epistemic Commons / Lock-In Overlay

One dyad drifting is a personal problem. A million dyads drifting the same way is an institutional one — and eventually a civilisational one. ECLO is the macro layer of the stack: it asks not whether one user was persuaded or one model hallucinated, but whether a knowledge system — a newsroom, a research field, a school system, a public agency, a platform ecosystem — remains able to contact evidence, preserve diversity, detect error, contest authority, retain human expertise, resist recursive contamination, and recover from AI-shaped trajectories. It treats epistemic risk as a meta-risk: when the capacity to know degrades, every other risk becomes harder to see.

ECLO deliberately creates no new human-side or machine-side codes. It aggregates the lower layers into seven commons-level dimensions, each with the same plain question behind it: is this capacity being preserved, or is it quietly becoming impossible to get back?

Dimension	What it preserves — and how it fails
Source Contact and Evidence Integrity (D1)	Preserves access to source material, transformations, evidentiary limits, and uncertainty. Fails when AI summaries and recommendations replace evidence contact and become the effective reality surface.
Diversity, Pluralism and Option-Space Health (D2)	Preserves diversity of sources, methods, cultures, languages, viewpoints, and alternatives. Fails when AI convergence narrows research, policy, creative, or public option spaces.
Feedback-Loop Velocity and Recursive Contamination (D3)	Measures and governs the loops where outputs become inputs — human-AI and AI-AI alike. Fails through synthetic recursion, model collapse, echo drift, and AI-shaped public commentary feeding back on itself.
Institutional Cognition and Expertise Retention (D4)	Preserves human expertise, tacit judgement, apprenticeship, and manual operating capacity. Fails when the institution becomes efficient but brittle — unable to inspect, teach, or operate without AI.
Deliberation and Contestability (D5)	Preserves genuine public reasoning: challenge, appeal, dissent, and affected-community participation. Fails when participation becomes symbolic and AI-curated consensus stands in for deliberation.
Reversibility and Response (D6)	Preserves the ability to pause, roll back, and restore human decision-making if automation fails or must be withdrawn. Fails when there is no tested manual fallback and no one with the authority to stop.
Longitudinal Commons Telemetry and Repair (D7)	Continuously monitors and repairs slow epistemic drift — because the most dangerous commons harms may never surface as complaints. Fails when the institution cannot tell whether its knowledge system is becoming more dependent, more homogeneous, or less contestable over time.

The overlay's most practitioner-relevant idea is lock-in: the point at which an AI-shaped arrangement can no longer be reversed, because the manual capacity, the human expertise, or the institutional will to operate without it has been allowed to lapse. ECLO's design controls — no-AI drills, manual-mode fallbacks, recursion labelling, diversity floors — all exist to keep the exit open.

8. Using this vocabulary in practice

A primer succeeds when the reader can do something new with it. Three habits translate this vocabulary into practice immediately.

Name both sides of every incident

When something goes wrong — or nearly does — resist the reflex to file it as either a model failure or a user error. Ask what the machine did (an RPT behaviour), what the human was primed to do with it (a CST susceptibility), and what the interaction rewarded. Incidents named this way point to fixes on both sides of the loop, which is usually where the durable fix lives.

Treat benefit claims as dyad claims

"The AI made the team faster" is a claim about one layer of a five-layer question. The Dyad-Aware Uplift Stack (DAUS-5) and the PDCO exist because task gains routinely coexist with epistemic, skill, relational, or governance losses. Before crediting a benefit, ask what the loop is training the humans to stop doing.

Reserve the word "oversight" for the real thing

The single most abused word in AI governance is oversight. This framework gives it teeth: if the human in the loop lacks the information, the time, the authority, or a feasible intervention, the Oversight Viability No-Go Gate (OVNG) says the arrangement should not be called oversight at all — and treats that as a design failure to fix, not a user inadequacy to retrain.

8.1 Where to go next

For assessment work, the DARCA-Lite instrument (Dyadic AI Risk & Capability Assessment) operationalises this vocabulary as a structured release-gate worksheet, with every item traceable to the taxonomy states behind it. For depth on any single entry — diagnostic criteria, measurement probes, red-team batteries, design controls — go to the source manuals: the CST Manual (v0.8.1 draft) for H-codes and Y-codes, the RPT (v2.0.1 draft) for L-codes, the PDCO (v0.4.1 draft) for P-states, and ECLO (v0.1 draft) for the commons dimensions. Each manual carries an alphabetical atlas and glossary designed for exactly the reader this primer has been addressing: someone who needs the right entry quickly and in plain language.

> **Provenance note**
>
> Every definition in this primer is a plain-language restatement of the corresponding entry in the source documents named above; codes are provided throughout so any statement can be traced to its diagnostic sheet. All four sources are working drafts, and this primer inherits their draft status. Where a source assigns an entry its own category label, that label — not this primer's editorial grouping — is authoritative.