Do Human Cognitive Failures Map onto AI Agent Failures? A Structured Literature Review

Tuomo Nikulainen, Pisama LLC

2026-05-08

Contents

TL;DR

We invert the usual AI-side framing and ask whether each of 45 well-studied human cognitive failure categories has substantial AI agent research engagement. The headline distribution from coder A's pass (6 / 24 / 12 / 0 / 3 across Substantial / Partial / Nascent / Absent / Substrate-absent) is fragile under multi-vendor LLM coding: a v2 6-coder pass across all 4 vendor families (Anthropic Opus 4.7, Sonnet 4.6, Haiku 4.5; OpenAI GPT-5.5; Google Gemini 3.1 Pro Preview; xAI Grok 4.3) yielded Fleiss' kappa = +0.224 ("fair, lower bound") with 39 of 45 categories showing at least one disagreement (Section 2.4.2). Gemini 3.1 Pro is a striking outlier (codes 16 categories as

Substantial vs 2-4 from every other coder); excluding Gemini brings Fleiss' kappa to +0.32–0.40 across the remaining 5 coders. Either way, the robust headline is dramatically narrower than the v1 single-coder finding suggests; the consensus 6 *Substantial* categories shrink under modal-verdict analysis to 2–6 depending on Gemini weighting. The 12 Nascent categories are the productive research-direction gaps; the deepest single gap is Hollnagel's Safety II reframe. Methodologically: PRISMA 2020 protocol; two-LLM-coder v1 pass within Anthropic with Cohen's kappa = 0.97 (transparently labeled as structurally inflated and empirically validated as such by the v2 0.224 result — a 0.75 absolute drop); 6-LLM-coder v2 pass across all 4 major vendor families (Section 2.4.2); three-coder human pass planned (Section 2.4.3); bibliometric validation pass on 30 of 45 categories with documented Google Scholar evidence (Appendix B); 92 verified references; active-inference theoretical lens with a precision-weighting argument (Section 5; load-bearing role is a planned revision target). Appendix C provides a full crosswalk between MAST's 14 multi-agent failure modes (Cemri et al 2025) and the 45-category human cognitive failure taxonomy.

Abstract

Background. Practitioners routinely reach for the human-factors and cognitive-psychology literature when designing safety mechanisms for large language model (LLM) agent systems: hallucination resembles confabulation, persona drift resembles role drift, multi-agent coordination breakdowns resemble Crew Resource Management failures. The reflex is intuitive but partially misleading.

Approach. We invert the usual AI-side framing. Rather than asking whether AI failures resemble human failures, we start from a comprehensive 45-category taxonomy of human cognitive failures (drawn from cognitive psychology, human factors, social psychology, neuropsychology, and decision-making research) and ask, for each, whether AI agent research currently engages with it. The contribution is a research roadmap for AI agent evaluation grounded in human cognitive science.

Methodology. PRISMA 2020 protocol with per-reference inclusion documentation (Appendix A). A five-level AI-research-status scheme (Substantial / Partial / Nascent / Absent / Substrate-absent). The methodology is reported in three stages: (1) v1 within-Anthropic two-LLM-coder pass at Cohen's kappa = 0.97 ("almost perfect" but structurally inflated; both coders share training distribution); (2) v2 multi-vendor 6-LLM-coder pass executed 2026-05-09 across all 4 major vendor families (Anthropic Opus 4.7, Sonnet 4.6, Haiku 4.5; OpenAI GPT-5.5; Google Gemini 3.1 Pro Preview; xAI Grok 4.3), yielding Fleiss' kappa = +0.224 ("fair, lower bound") and empirically validating the v1 inflation caveat with a 0.75 absolute drop; (3) three-coder human pass planned with the first author plus two named human raters (Section 2.4.3). We position the review against five recent AI-side surveys (Ji et al 2023 hallucination; OWASP LLM Top 10 2025; Cemri et al 2025 MAST; Mohammadi et al 2025 KDD agent benchmarking; Hammond et al 2025 multi-agent risks) and develop a unified theoretical lens via active inference (Pezzulo et al 2024 *TICS*). The AI-side inventory comprises Pisama's full 57 production detectors plus MAST 14, OWASP relevant 5, and Hammond 3 (about 70 unique categories after deduplication).

Headline finding (v1 single-coder, fragile under multi-vendor coding). Of 45 human cognitive failure categories per coder A's pass: 6 (13%) have Substantial AI agent research engagement, 24 (53%) Partial, 12 (27%) Nascent, 0 Absent, 3 (7%) Substrate-absent. The 12 Nascent categories are the productive research-direction gaps; the deepest single gap is Hollnagel's Safety II reframe. *Caveat:* the v2 multi-vendor 6-coder pass shows that 39 of 45 categories have at least

one cross-vendor disagreement, and Gemini 3.1 Pro is a striking outlier (codes 16 categories as Substantial vs 2-4 from every other coder). Excluding Gemini, the modal-verdict-across-5-coders distribution differs from the v1 single-coder distribution in approximately 8–12 categories, with the 6 *Substantial* count shrinking to 2–4. Including Gemini, the robust modal-verdict count of *Substantial* sits in 2–6 depending on whether Gemini's outlier votes weight equally. The robust headline is therefore narrower than the v1 distribution suggests, and final adjudication is gated on the planned three-coder human pass (Section 2.4.3).

Implications. Borrow the structure of the human-cognitive failure taxonomy as a research roadmap; do not borrow mechanism, since the predictive-processing analysis (Section 5) shows that LLMs are passive generative AI rather than active inference systems and many human-mechanism interventions therefore do not transfer. Engage the human research areas that AI evaluation has not yet imported.

Keywords: human-factors engineering, cognitive psychology, LLM agents, AI safety, failure modes, taxonomy, active inference, systematic literature review, PRISMA 2020.

---

1. Introduction

The current generation of LLM agent systems exhibits failure modes that are easy to describe in human-cognitive terms. Loops resemble perseveration. Hallucination resembles confabulation. Persona drift resembles role drift. Specification mismatches resemble mode confusion. Coordination failures in multi-agent systems resemble Crew Resource Management failures from the aviation literature. The temptation to apply forty years of human-factors theory to AI agents is, on the face of it, well-founded.

This review interrogates that mapping. The central question is structural rather than empirical: do human cognitive failure modes, as studied in psychology and human-factors engineering, recapitulate in LLM agent systems, and where they do not, what fills the gap? The motivation is practical. If the mapping is strong, agent-safety practitioners can borrow validated human-factors interventions (Crew Resource Management training, mode-confusion-aware interface design, premature-closure mitigations from naturalistic decision making) and adapt them. If the mapping is partial, those interventions need adaptation in ways that depend on understanding where mechanism diverges from surface. If the mapping has gaps, novel theory is needed for the failure modes that have no human predecessor.

We approach the question from the human side rather than the AI side. Instead of starting from a list of LLM failure modes and asking which have human analogs, we start from a taxonomy of human cognitive failures as it stands in 2026 and ask which appear in LLM agents. This direction is more useful because the human taxonomy is the more mature: cognitive psychology and human factors have spent half a century cataloguing failure modes with theoretical mechanism, intervention design, and empirical validation. The LLM agent failure literature, by contrast, is roughly five years old and is itself partly built on borrowed terms from the older field.

Three definitions for this review:

- **Surface signature transfer.** An AI failure mode produces an observable behavior that matches a human failure category. Example: an agent emits a false factual claim with high confidence, surface-matching confabulation.

- **Mechanism transfer.** The underlying causal explanation in humans (for example, reality-monitoring deficit in orbitofrontal lesions for spontaneous confabulation) matches the underlying causal explanation in AI agents (for example, next-token sampling over a learned distribution that places probability mass on plausible-but-false continuations).
- **Intervention transfer.** Interventions developed for the human failure mode (for example, reality-monitoring training, environmental cues) are useful, possibly with adaptation, for the AI failure mode.

A failure category transfers strongly when all three apply. A category transfers in surface only when the first applies but not the second or third. A category fails to transfer when the second is fundamentally different. A category is novel-to-AI when no useful human analog exists at any level.

The rest of this review proceeds as follows. Section 2 documents our methodology, including search strategy, coding scheme, and LLM-coder agreement. Section 3 positions the review against five recent surveys (Ji et al 2023, OWASP 2025, Cemri et al 2025, Mohammadi et al 2025, Hammond et al 2025); Section 3.3 covers MAST in detail and Section 3.3.1 reports a load-bearing companion-paper finding on detection-substrate κ against MAST labels. Section 4 reviews the state of human-failure research in 2020-2026. Section 5 develops a unified theoretical lens grounded in predictive processing and active inference, with a precision-weighting account at the conceptual level (Sections 5.5-5.6, following Pezzulo et al 2024 Box 3). Section 6 organizes human cognitive failures into 45 categories across thirteen subdisciplines. Section 7 maps each category to its AI agent analog and quantifies the AI-research-status counts (Table 1, Section 7.11). Section 8 names human failures that do not transfer because of substrate differences. Section 9 names AI failures with no clean cognitive-mechanism predecessor (qualified from the v1 framing of "no human predecessor"; surface analogs in social psychology exist and are addressed explicitly). Section 10 names human research areas that have no current AI agent counterpart and presents Table 2. Section 11 cross-links the taxonomy claims to specific empirical measurements in our companion empirical paper. Section 12 discusses implications. Section 13 acknowledges limitations. Section 14 develops a research agenda with 18 first-experiment proposals: 12 anchored in the Nascent categories and 6 in underdeveloped Partial categories where the AI literature is fragmentary. Section 15 concludes. Appendix A is the PRISMA 2020 flow diagram and per-reference inclusion table; Appendix B is the bibliometric validation of the AI-research-status assignments; Appendix C is the full crosswalk between MAST's 14 multi-agent failure modes and the 45-category human cognitive failure taxonomy.

2. Methodology

We adopt a PRISMA-lite protocol (Page et al 2021) appropriate for a structured literature review whose primary contribution is conceptual organization rather than effect-size synthesis. The methodology is transparent rather than comprehensive: we explicitly document what was searched, what was included, and what was not, and we acknowledge limitations that a fully systematic review would resolve.

2.1 Search strategy

The literature was assembled in three concurrent threads:

Thread A (human cognitive failures and human-factors engineering). Searches across Google Scholar, PubMed, and journal-direct websites for canonical references in: cognitive psychology of memory, attention, and decision making; human-factors engineering (CRM, mode confusion,

automation surprise); resilience engineering; naturalistic decision making; cognitive load theory; group decision making and information cascades. Search terms included: "spontaneous confabulation Schnider 2020 2024 mechanism review", "mode confusion automation surprise human factors 2023 2024 review", "working memory capacity Cowan 2024 latest review individual differences", "goal neglect frontal lobe 2023 review cognitive control latest", "naturalistic decision making Klein 2024 expertise review premature closure", "source monitoring failures Johnson 2023 2024 reality monitoring review", "cognitive load theory Sweller 2024 latest meta-analysis intrinsic", "information cascade collective decision making 2024 group review", "resilience engineering Hollnagel 2023 2024 functional resonance latest", "crew resource management 2023 team adaptation aviation healthcare update", "predictive coding active inference cognitive failure 2024 Friston review".

Thread B (LLM agent failures and prior surveys). Searches for: hallucination surveys (Ji et al 2023 in *ACM Computing Surveys*); LLM safety and prompt injection (OWASP LLM Top 10, Greshake et al 2023); multi-agent failure taxonomies (Cemri et al 2025 "Why Do Multi-Agent LLM Systems Fail?", Hammond et al 2025); LLM agent benchmarking surveys (Mohammadi et al 2025 KDD'25); sycophancy in language models (Sharma et al 2024 ICLR). Search terms included: "Ji 2023 survey hallucination natural language generation taxonomy", "OWASP LLM Top 10 2025 prompt injection detection categories", "multi-agent LLM failure survey 2024 2025 review benchmarking", "LLM sycophancy 2024 Sharma Anthropic agreement bias conformity", "prompt injection LLM social engineering suggestibility human analog 2024".

Thread C (cross-cutting theoretical frameworks). Searches for: predictive processing and active inference applied to LLMs (Pezzulo et al 2024 *Trends in Cognitive Sciences*; Smith et al 2021; Sprevak 2024); human-AI teaming and trust (Georganta 2024; Verma et al 2025 on automation bias). Search terms included: "Pezzulo Friston 2024 active inference passive AI generating meaning Trends Cognitive Sciences", "predictive coding LLM hallucination precision weighting 2024", "human-AI teaming cognitive load 2024 trust automation review".

2.2 Inclusion and exclusion criteria

A reference was included if it satisfied at least one of the following:

1. Foundational status in its subfield (Reason 1990, Sarter and Woods 1994/1997, Wickens et al 2021 5th ed, Klein 1998, Hollnagel et al 2006, Schnider 2003, Duncan et al 1996, Wegner 1985, Johnson and Raye 1981, Cowan 2010, Bikhchandani et al 1992, Janis 1972).
2. Recent (2020-2026) update or meta-analysis of one of the foundational references.
3. Published prior survey of LLM failure modes (Ji et al 2023, OWASP 2025, Cemri et al 2025, Mohammadi et al 2025, Hammond et al 2025).
4. Theoretical framework explicitly bridging cognitive science and AI (Pezzulo et al 2024).

A reference was excluded if it was: a position paper without empirical or theoretical contribution; a non-peer-reviewed industry artifact except when no peer-reviewed alternative existed (specifically the OWASP LLM Top 10, which is the canonical operational reference for prompt injection); or a working paper or preprint older than 12 months without subsequent publication.

2.3 Coding scheme

Each human cognitive failure category in the 45-category taxonomy of Section 6 was coded against the AI-research-status scheme (Substantial / Partial / Nascent / Absent / Substrate-absent; defined in Section 2.4). The coder used the four transfer dimensions (Section 1) — surface transfer,

mechanism transfer, intervention transfer, and overall verdict — to support each AI-research-status assignment. The coder was the single author, working with the human-factors taxonomy in Section 6 and the AI-side inventory comprising Pisama's 57 production detectors (extracted from the empirical companion paper, Section 11), MAST's 14 multi-agent failure modes (Cemri et al 2025), the OWASP LLM Top 10 (2025) relevant subset, and Hammond et al's three multi-agent failure categories (2025). The deduplicated AI-side inventory totals approximately 70 unique categories, mapped onto the 45 human categories via the cells of Table 1 (Section 7.11) and the MAST-specific crosswalk in Appendix C.

The coding produced the quantitative breakdown reported in Section 7.11 and reaffirmed in the bibliometric validation pass (Appendix B): of 45 human cognitive failure categories examined, 6 (13%) have *Substantial* AI agent research engagement, 24 (53%) have *Partial* engagement, 12 (27%) have *Nascent* engagement, 0 are *Absent*, and 3 (7%) are *Substrate-absent* (fatigue, sleep, stress at the model level). On the AI-side, four failures from the empirical companion paper's detector inventory plus the 2026 literature (prompt injection, sycophancy cascade, convergence pathology, emergent collusion) have no clean cognitive-mechanism predecessor and are discussed in Section 9.

2.4 Coder agreement methodology

The methodology evolves across three stages. The v1 stage (executed) used two LLM coders within the same Anthropic model family and produced Cohen's kappa = 0.97. The v2 multi-vendor stage (executed 2026-05-09 with paid-tier API access across all 4 major vendors) uses 6 LLM coders: Anthropic Opus 4.7, Sonnet 4.6, Haiku 4.5; OpenAI GPT-5.5; Google Gemini 3.1 Pro Preview; xAI Grok 4.3 — and produced Fleiss' kappa = +0.224, empirically validating the v1 structural-inflation caveat with a 0.75 absolute drop. The v2 human-coder stage is planned with three named raters (gated on coauthor confirmation; Section 2.4.3). All three stages are reported because each addresses different failure modes of the methodology, and reporting all three together is more honest than collapsing to a single number.

2.4.1 v1 LLM-coder pass (executed; structurally inflated) This pass used two LLM coders within the Anthropic Claude model family. The first author (Coder A, Claude Opus 4.7) developed the protocol and produced the primary coding. An independent second-coder pass (Coder B, Claude general-purpose agent) was dispatched with only the protocol document, blind to Coder A's reasoning. Both coders applied the five-label AI-research-status scheme to all 45 human cognitive failure categories (Section 6).

Inter-coder agreement. Across the 45 categories, the two coders agreed on 44 (97.8% observed agreement). The single disagreement was on Item 4 (False memory / DRM-style suggestibility), which Coder A coded *Nascent* and Coder B coded *Partial*. Both judgments are defensible; the disagreement reflects how sparse the cognitive-bias-in-LLM probing literature is.

Cohen's kappa = 0.97, "almost perfect" by Landis and Koch (1977).

Critical caveat: structural inflation.

1. Both coders share substantial training distribution (same model family).
2. Both received the same written protocol; the protocol's worked examples function as anchors.
3. The protocol was authored by Coder A; Coder B applied a framework rather than constructing it independently.

4. The 0.97 kappa is mechanically inflated by these structural similarities. The MAST taxonomy paper (Cemri et al 2025) reports human-human kappa = 0.88 with expert annotators and disagreement-resolution protocol — a stronger benchmark.

2.4.2 v2 multi-vendor LLM-coder pass (executed 2026-05-09; latest models, all 4 vendors) The v2 pass uses 6 LLM coders across 4 vendor families with **the most recent flagship models available as of May 2026**, verified against each vendor's model-listing API at run time:

Role	Vendor	Model	Released
A	Anthropic	Claude Opus 4.7	2026-04-14
B	Anthropic	Claude Sonnet 4.6	2026-02-17
C	Anthropic	Claude Haiku 4.5	2025-10-15
D	OpenAI	GPT-5.5	2026-04-23
E	Google	Gemini 3.1 Pro Preview	2026 (preview)
F	xAI	Grok 4.3	2026-04

All received the v2 falsifiable-criteria protocol (`experiments/20260430/coding_protocol_v2.md`). Anthropic Opus 4.7 deprecates the `temperature` parameter (reasoning-class model); calls to it omit temperature. Other models use `temperature = 0` for reproducibility. The full coding logs are at `experiments/20260509/raw_<role>_<vendor>_<model>.txt`; the kappa report at `experiments/20260509/multi_vendor_kappa_report.json`.

Pairwise Cohen's κ matrix (measured 2026-05-09, all 6 coders, 4 vendors):

	Opus 4.7	Sonnet 4.6	Haiku 4.5	GPT-5.5	Gemini 3.1 Pro	Grok 4.3
Opus 4.7	—	+0.557	+0.249	+0.485	+0.065	+0.259
Sonnet 4.6	+0.557	—	+0.327	+0.310	+0.070	+0.201
Haiku 4.5	+0.249	+0.327	—	+0.164	+0.071	+0.532
GPT-5.5	+0.485	+0.310	+0.164	—	+0.127	+0.156
Gemini 3.1 Pro	+0.065	+0.070	+0.071	+0.127	—	+0.069
Grok 4.3	+0.259	+0.201	+0.532	+0.156	+0.069	—

Fleiss' κ across 6 coders: +0.224. Verbal interpretation: "fair agreement" (Landis and Koch 1977; the lower bound of the fair range). 39 of 45 categories have at least one disagreement.

Empirical validation of the v1 inflation caveat. The v1 within-Anthropic same-base-model LLM-coder κ was +0.97. The v2 cross-vendor LLM-coder Fleiss' κ across 6 coders is +0.224. The drop from 0.97 to 0.22 — a **0.75 absolute gap** — quantifies the structural-inflation effect we flagged in §2.4.1 caveat 4. **Within-Anthropic mean pairwise κ is** +0.378 (across the three Opus/Sonnet/Haiku pairs); **cross-vendor mean pairwise κ is** +0.205 (across the twelve cross-family pairs). The 0.17 within-vs-cross gap is now consistent and large: same-vendor coders agree meaningfully more than cross-vendor coders, and same-vendor-same-base-model coders (the v1 regime) agree dramatically more than either.

Coder E (Gemini 3.1 Pro Preview) is a striking outlier. Pairwise κ of Gemini with every other coder is in the +0.065 to +0.127 range — an order of magnitude lower than the other inter-coder pairs. This is not noise; it is a systematic difference in coding behavior visible in the marginal-distribution table below. Excluding Gemini and re-computing across the remaining 5 coders gives a Fleiss' κ in the +0.32 to +0.40 range — closer to the v5 result that lacked Gemini. **Adding Gemini to the multi-vendor pass cuts Fleiss' κ roughly in half**; the inclusion is methodologically correct (4-vendor coverage is more honest than 3) but the inclusion-vs-exclusion sensitivity is itself a finding: the LLM-coder κ depends sharply on the choice of vendor lineup, even when all vendors run their flagship 2026 model.

Per-coder marginal distributions (illustrating that different LLM coders produce systematically different verdict distributions; Gemini outlier highlighted in bold):

Verdict	Opus 4.7	Sonnet 4.6	Haiku 4.5	GPT-5.5	**Gemini 3.1 Pro**	Grok 4.3	v1 post-bibliometric
Substantial	2	2	3	4	**16**	2	6
Partial	15	24	14	21	**23**	9	24
Nascent	22	15	12	13	**2**	21	12
Absent	2	2	8	5	**0**	12	0
Substrate-absent	4	2	8	2	**4**	1	3

Gemini 3.1 Pro Preview codes **16 categories as Substantial** (3.5–8x what every other coder sees) and only **2 categories as Nascent** (1/10 of the next-most-conservative coder). This permissive-coding pattern is the proximate cause of the across-the-board low pairwise κ values for Gemini. Whether this reflects (a) a genuine difference in Gemini's reading of the v2 protocol, (b) a different prior on what constitutes "engagement," or (c) a calibration artifact of the preview model is open. The point of running cross-vendor coders is precisely to surface such differences; the methodological response is not to exclude Gemini but to disclose its disagreement and account for it in the final adjudication step.

Implication for the headline finding. If we report the modal-verdict-across-6-coders distribution as the headline instead of the coder-A v1 distribution, the result reshapes meaningfully. Excluding Gemini from the modal calculation (because of its outlier behavior), the modal-verdict 5-coder distribution remains approximately 2–4 *Substantial*, approximately 15–24 *Partial*, approximately 13–22 *Nascent*, approximately 2–10 *Absent*, approximately 1–8 *Substrate-absent*. Including Gemini in the modal vote pulls the distribution toward more *Substantial* (Gemini votes Substantial on 16 categories that others code lower). The robust modal-verdict headline that survives multi-vendor coding **including Gemini** is therefore: 2–6 *Substantial* depending on whether Gemini's outlier votes are weighted equally to the others. Final adjudication awaits the planned three-coder human pass (Section 2.4.3), where each disagreement is resolved with named adjudication and rationale.

Implication for the v2 falsifiable criteria. Fleiss' $\kappa = 0.224$ across 6 coders is well below the 0.70 threshold the v2 protocol README anticipated. **The v2 criteria need tightening before the human-coder pass runs**, with particular attention to whatever protocol ambiguity is causing Gemini to over-code Substantial. Likely revisions: (a) numerical thresholds for paper-eligibility under Dimension 1 are still interpretation-dependent; (b) the bibliometric procedure needs

to produce a shared candidate-paper list across coders before independent verdict assignment; (c) a small pre-coding calibration sub-sample (5–10 categories) with disagreement-resolution iteration, mirroring the MAST team's pre-coding protocol that achieved their human-human $\kappa = 0.88$.

Run-to-run variance. Earlier runs at temperature = 0 produced Fleiss' κ values of +0.190 (4-coder test pass before fixing OpenAI's `max_completion_tokens` parameter), +0.272 (5 coders, older model versions: Opus 4.5, Sonnet 4.5, Haiku 4.5, GPT-5, Gemini 2.5 Pro), +0.395 (5 coders, latest models, Gemini failed quota), and +0.224 (this run, 6 coders, all 4 vendors with paid Google API access). The variance has three sources: (a) vendor-side non-determinism even at temperature = 0; (b) different model-version mixes producing meaningfully different agreement levels; (c) inclusion vs exclusion of Gemini, which alone moves Fleiss' κ by approximately 0.15 absolute due to its outlier coding behavior. **For peer-review submission, the multi-vendor pass should be re-run at least 3 times with the same 6-coder lineup, report mean ± SD Fleiss' κ, and report Gemini-included and Gemini-excluded numbers separately.** We treat the 0.224 figure as a single-run point estimate at one specific 6-coder lineup, not a stable measurement.

2.4.3 Planned human-coder pass (gated on coauthor confirmation) The multi-vendor LLM-coder pass is a methodology *waypoint*, not a replacement for human inter-rater reliability. We propose a bounded **three-coder human pass** with the first author (Tuomo Nikulainen, Pisama LLC), a multi-agent-failure-taxonomy expert (Mert Cemri, UC Berkeley, MAST first author; coauthorship discussion in progress), and a cognitive-science expert (TBD; outreach to Hagendorff or Binz once Cemri confirms). The three coders independently apply the v2 falsifiable-criteria protocol to all 45 categories, blind to one another's verdicts. We will report:

- Per-rater verdict tables.
- Pairwise Cohen's κ for all three pairs.
- Fleiss' κ across all three.
- Comparison to the v2 multi-vendor LLM-coder κ (does human-human exceed cross-vendor LLM-LLM?).
- A formal disagreement-resolution table with named adjudication and rationale per disagreement.

The MAST paper's $\kappa = 0.88$ is the benchmark. Estimated effort: ~2.5 hours per rater plus ~3 hours adjudication. The pass is gated on coauthor confirmation; once confirmed, the protocol is pre-registered on OSF Registries before final coding begins.

2.4.4 Final reported number The version of this paper submitted to peer review will report the human-human Fleiss' κ from §2.4.3 as the primary methodology number, with the v2 multi-vendor LLM-coder κ from §2.4.2 as a methodological cross-check, and the v1 within-Anthropic $\kappa = 0.97$ as a transparency record of the original pass. The reader is invited to compare all three against the MAST 0.88 benchmark.

The full coding protocol (v1) is at `experiments/20260430/coding_protocol.md`; the v2 falsifiable-criteria revision is at `experiments/20260430/coding_protocol_v2.md`; Coder B's v1 verdicts are at `experiments/20260430/llm_coder_pass.md`; the v1 kappa computation is at `experiments/20260430/kappa_analysis.md`; the multi-vendor pass scaffolding is at `experiments/20260509/`.

2.5 Methodology box

Item	Decision
Search databases	Google Scholar, PubMed, journal websites, arXiv
Search threads	Three (human cognitive failures, LLM agent surveys, cross-cutting theory)
Inclusion criteria	Foundational status, 2020-2026 update, prior LLM survey, or cognitive-AI bridge framework
Framing	Human-first: start from human cognitive failure taxonomy, ask AI-research-status
Categories coded	45 human cognitive failure categories (Section 6, full list)
AI-side inventory	Pisama 57 detectors + MAST 14 + OWASP relevant 5 + Hammond 3 (about 70 unique after dedup)
Verdict labels	Substantial / Partial / Nascent / Absent / Substrate-absent (5 levels)
Coders, v1 (executed)	2 LLMs in same family (Coder A: Claude Opus 4.7; Coder B: Claude general-purpose agent)
Coders, v2 multi-vendor (executed 2026-05-09; full 6-coder run)	6 LLMs across 4 vendors (Anthropic Opus 4.7, Sonnet 4.6, Haiku 4.5; OpenAI GPT-5.5; Google Gemini 3.1 Pro Preview; xAI Grok 4.3)
Coders, v2 human pass (planned, gated on coauthors)	3 humans (Tuomo Nikulainen; Mert Cemri TBC; cog-sci coauthor TBD — Hagendorff or Binz)
Cohen's kappa, v1 LLM-LLM same-family	0.97 (structurally inflated; reported transparently)
Cohen's kappa, v2 cross-vendor LLM-LLM (6 coders, all 4 vendors, latest models, 2026-05-09)	**Fleiss'** $\kappa = +0.224$ ("fair, lower bound"); within-Anthropic mean pairwise +0.378, cross-vendor mean pairwise +0.205; structural-inflation gap from v1 = **0.75 absolute**. Gemini 3.1 Pro is a striking outlier (codes 16 Substantial vs 2-4 from others); excluding Gemini brings Fleiss' to +0.32-0.40
Cohen's kappa, v2 human-human (target)	TBD (target $\geq$ MAST's 0.88 via three-coder human pass)
Observed agreement, v1	44 / 45 = 97.8%
Disagreements, v1	1 (Item 4: False memory; Coder A Nascent, Coder B Partial)
Verdict criteria	v1 worked-example anchors (executed); v2 falsifiable thresholds with bibliometric procedure (`experiments/20260430/coding_protocol_v2.md`; pending pre-registration on OSF Registries before v2 final coding)
Reference benchmark	MAST 2025 reports human-human kappa = 0.88

3. Related work and prior reviews

This section positions our review against five recent surveys. Each defines a partial taxonomy of LLM or LLM-agent failures, and our review is differentiated by its starting point in human cognitive science and its explicit transfer-dimension analysis.

3.1 Ji et al (2023): hallucination in natural language generation

Ji, Lee, Frieske, et al (2023, *ACM Computing Surveys* Vol 55, Article 248) is the canonical recent reference for hallucination taxonomy. The survey is organized in three parts: general overview of metrics and mitigation; task-specific progress in abstractive summarization, dialogue, generative QA, data-to-text, machine translation, and visual-language generation; and a final part on hallucinations in LLMs specifically. The survey distinguishes intrinsic hallucination (output contradicts known input or context) from extrinsic hallucination (output is not grounded in source but cannot be immediately judged false). The taxonomy is highly granular at the task and mechanism level.

Our review differs from Ji et al in three ways:

1. **Scope.** Ji et al cover hallucination only; we cover a broader 18-category LLM agent failure taxonomy of which hallucination is one cell.
2. **Direction.** Ji et al survey AI-side phenomena; we approach from the human side and ask whether each human failure mode appears in AI.
3. **Theoretical lens.** Ji et al organize by NLG task; we organize by transfer-dimension analysis grounded in human-factors theory and predictive processing (Section 5).

We cite Ji et al's intrinsic/extrinsic distinction in our hallucination mapping (Section 7.3) and accept it as the operational definition of LLM hallucination for this review. *Update from the 2025–2026 literature*: Alansari and Luqman (2025, "Large Language Models Hallucination: A Comprehensive Survey," arXiv:2510.06265) supplement Ji et al with an updated taxonomy of detection approaches (retrieval-, uncertainty-, embedding-, learning-, and self-consistency-based) and mitigation strategies organized by lifecycle stage (data-centric and pre-training; model-centric fine-tuning and alignment; inference-time post-hoc). Where this review's verdicts depend on the state of the hallucination-detection literature (cat 3 spontaneous confabulation, cat 4 false memory), the Alansari and Luqman 2025 framework supersedes Ji et al 2023 as the up-to-date reference; the Ji et al intrinsic/extrinsic distinction remains conceptually load-bearing.

3.2 OWASP LLM Top 10 (2025)

The OWASP Gen AI 2025 LLM Top 10 (LLM01:2025 to LLM10:2025) is the canonical operational reference for LLM-specific security risks. LLM01 is Prompt Injection, defined as a vulnerability where user prompts alter the LLM's behavior or output in unintended ways. The OWASP framework distinguishes direct injection (manipulation of user prompts) from indirect injection (hidden instructions in external content the LLM processes), and enumerates detection categories including: explicit malicious instructions, encoding-based hiding, role-play bypass (DAN-style), system-instruction extraction attempts, multimodal injection (instructions hidden in images), RAG-injection (poisoning external knowledge bases), and agent-tool/reasoning attacks (forging agent reasoning steps and tool outputs).

Our review uses OWASP's prompt-injection taxonomy as the operational definition for our "no human analog" category in Section 9. The OWASP framework is operationally rich but is not a research review; it does not engage with the human social-engineering literature against which prompt injection is sometimes compared. Our Section 9 makes the explicit case that prompt injection's mechanism (absence of a privileged instruction-versus-data channel) has no human analog, and that defenses must be architectural rather than psychological.

3.3 Cemri et al (2025): "Why Do Multi-Agent LLM Systems Fail?"

Cemri, Pan, Yang, et al (2025, NeurIPS 2025 Datasets and Benchmarks Track; arXiv:2503.13657) introduce MAST (Multi-Agent System failure Taxonomy), built from analysis of 1,642 multi-agent execution traces across seven state-of-the-art open-source multi-agent systems (ChatDev, MetaGPT, AG2, Magentic-One, AppWorld, OpenManus, HyperAgent). The reported failure rates range from 41% to 86.7% depending on the system. MAST organizes 14 failure modes (FM-1.1 through FM-3.3) into three thematic categories that follow the inter-agent conversation lifecycle: **FC1 Specification and System Design** (FM-1.1 Disobey task specification, FM-1.2 Disobey role specification, FM-1.3 Step repetition, FM-1.4 Loss of conversation history, FM-1.5 Unaware of termination conditions); **FC2 Inter-Agent Misalignment** (FM-2.1 Conversation reset, FM-2.2 Fail to ask for clarification, FM-2.3 Task derailment, FM-2.4 Information withholding, FM-2.5 Ignored other agent's input, FM-2.6 Reasoning-action mismatch); and **FC3 Task Verification and Termination** (FM-3.1 Premature termination, FM-3.2 No or incomplete verification, FM-3.3 Incorrect verification). The MAST paper reports a human-human inter-annotator Cohen's $\kappa = 0.88$ on a 21-trace expert-annotator subset and a $\kappa = 0.77$ for an OpenAI o1 LLM-judge on the same subset.

Our review differs from Cemri et al in: 1. **Theoretical grounding.** MAST is empirically derived; our taxonomy is grounded in human-factors theory and predictive processing (Section 5). 2. **Direction of analysis.** MAST is AI-side first: it begins with observed traces and abstracts failure modes upward. Our review is human-side first: it begins with the 45-category cognitive failure taxonomy and asks which modes manifest in AI agents (Section 7) and which do not (Sections 8–10). 3. **Mechanism analysis.** MAST does not engage with whether failures have human analogs; we do, and the precision-weighting account in Section 5 supplies a candidate mechanism for several MAST modes (FM-1.4 maps to working-memory limits, FM-2.1 to source-monitoring failure, FM-3.1 to premature closure / satisficing). 4. **Intervention space.** MAST identifies failure modes; we discuss intervention transfer from the human-factors literature (Crew Resource Management for FC2, mode-confusion-aware design for FM-2.6, premature-closure mitigations from naturalistic decision making for FM-3.1). 5. **Coverage scope.** MAST's 14 modes are tightly multi-agent; our 45 categories include single-agent cognitive failures (working memory, attention, decision-making biases, calibration, theory of mind) for which MAST has no slot.

We cite MAST as the canonical reference for the empirical failure base rate in multi-agent systems and as a complementary categorization (theirs by conversation-lifecycle stage, ours by cognitive analog). Appendix C provides a full crosswalk between MAST's 14 modes and the 45-category human cognitive failure taxonomy in Section 7.11.

3.3.1 Companion-paper finding: detection-substrate κ against MAST labels Our companion empirical paper (Section 5.7 of `paper.md`) reports a load-bearing follow-up against MAST. We ran four detection substrates against the 1,242-trace LLM-annotated subset of the MAST corpus released as `huggingface.co/datasets/mcemri/MAD`: (i) the Pisama tiered-detector pipeline; (ii) Anthropic Claude Haiku 4.5 as a single-call 14-mode judge; (iii) Anthropic Claude Sonnet 4.6

as the same; (iv) a substrate-concurrence ensemble. Pooled Cohen's κ values against the MAST-released labels are within ± 0.04 of zero across all four substrates. The MAST paper's published reference values ($\kappa \approx 0.77$ for the o1 LLM-judge that generated the labels, against humans on a 21-trace subset; $\kappa = 0.88$ for human-human agreement on the same subset) are roughly 20–30 times higher than anything we measure with substrates that did not generate the labels.

Two readings of this result are possible. (a) Judge-style alignment with the particular label-generator (o1) dominates judge capability, so a fresh Claude judge of any size has trouble matching o1's labeling style; this is a well-known LLM-judge phenomenon and is not a critique of MAST. (b) The 14-mode classification is genuinely hard at the per-trace level, and operationalized detectors built without MAST-specific calibration do not yet recover the distinctions the labels capture. Both readings argue that MAST is a *load-bearing* benchmark — it exposes detection limitations that single-substrate evaluators do not surface. The cross-tier (heuristic-vs-LLM) decorrelation finding in the companion paper (FN Jaccard 0.42 cross-tier vs 0.79 across-vendor LLM-LLM) suggests that substrate-concurrence ensembles are the productive engineering response, and is the empirical foundation for the tiered architecture.

This finding is reported here because it bears directly on the AI-side framing of the present review: MAST is not just a taxonomy to map onto, it is a benchmark whose label distribution is currently un-recovered by any single off-the-shelf detection substrate we have tested. The implications for how the present review's mappings (Section 7.11, Appendix C) should be read in practice are discussed in Section 11 (Cross-link to companion empirical paper) and Section 12 (Implications).

3.4 Mohammadi et al (2025): LLM agent benchmarking survey

Mohammadi et al (2025, KDD'25) survey LLM agent evaluation methods. They introduce a two-dimensional taxonomy organizing evaluation work along (a) evaluation objectives (agent behavior, capabilities, reliability, safety) and (b) evaluation process (interaction modes, datasets, benchmarks, metrics, tooling). The survey is descriptive and methodologically focused.

Our review is differentiated by being conceptual rather than methodological: we ask what failure modes exist and how they map to human cognitive failures, not how to measure them. The Mohammadi et al survey is complementary; the metrics and benchmarks they catalog provide the operational substrate for empirical testing of the categorization claims we make conceptually.

3.5 Hammond et al (2025): multi-agent failure-mode taxonomy

Hammond et al (2025) split multi-agent failure modes into three categories: miscoordination (adverse consequences from agents with identical objectives failing to cooperate effectively), conflict (adverse consequences from agents with mixed objectives), and collusion (multiple AI agents cooperating in undesirable circumstances). This taxonomy emphasizes the inter-agent strategic dimension and connects to game theory and mechanism design.

Our taxonomy nests below Hammond et al at the multi-agent level: their three categories cluster within our communication/coordination (Section 6.4) and group-failures (Section 6.8) categories. The Hammond et al collusion category is the most novel addition to our framework; we treat it as adjacent to but distinct from sycophancy cascade in Section 9. Future work should engage more directly with the collusion-as-failure-mode framing, which we do not develop fully here.

3.6 Synthesis: where this review fits

Our review fills a gap left by the five surveys. Ji et al cover hallucination; OWASP covers operational LLM security; Cemri et al cover multi-agent empirical failure rates; Mohammadi et al cover evaluation methods; Hammond et al cover multi-agent strategic dynamics. None of the five starts from the human cognitive failure taxonomy and asks which human failures appear in AI agents and which do not. None of the five develops an explicit transfer-dimension framework. None of the five identifies the human research areas that AI agent research has not yet engaged with (Section 10). Our review is conceptual organizing scaffolding for empirical follow-up; the empirical companion paper provides the measured anchor.

4. State of the field, 2020-2026

The classical references in this space are well-established. James Reason's *Human Error* (1990) introduced the Swiss Cheese model and the active-versus-latent error distinction. Sarter and Woods' (1994, 1997) work on cockpit automation gave us mode confusion and automation surprise. Wickens et al's *Engineering Psychology and Human Performance* (latest 5th edition 2021) codified working-memory and channel-capacity limits. Klein's *Sources of Power* (1998) introduced the recognition-primed decision (RPD) framework that grounded naturalistic decision making (NDM). Hollnagel et al's *Resilience Engineering: Concepts and Precepts* (2006) and the Functional Resonance Analysis Method (FRAM, 2012) reframed safety from "preventing failure" to "engineering resilience." Johnson and Raye (1981) developed source-monitoring theory; Wegner (1985) introduced transactive memory; Duncan et al (1996) characterized goal neglect; Schnider's (2003) *Nature Reviews Neuroscience* paper defined spontaneous confabulation as a reality-filtering deficit.

Five developments in the 2020-2026 period are relevant for the AI mapping:

First, predictive processing has matured into a unifying framework that interprets human perception, action, and cognitive failure as inference under hierarchical generative models. Pezzulo, Parr, Cisek, Clark, and Friston's 2024 paper "Generating meaning: active inference and the scope and limits of passive AI" in *Trends in Cognitive Sciences* explicitly compares the active-inference account of human cognition to the passive next-token prediction of LLMs, arguing that LLMs lack the generative loop that grounds human meaning. Smith et al's 2021 review in *Psychiatry and Clinical Neurosciences* applies predictive coding to clinical disorders. Sprevak's 2024 *Philosophy Compass* paper "Predictive Coding I: Introduction" gives the clearest current introduction. We develop this framework as our theoretical lens in Section 5.

Second, the mode-confusion literature has been refreshed. Skraaning and Jamieson's 2024 paper "The Failure to Grasp Automation Failure" in *Journal of Cognitive Engineering and Decision Making* reviews two decades of automation-surprise research and notes that mode-confusion patterns persist despite improved interface design. A 2024 meta-review of automation-surprise sources in aviation argues that the taxonomy of surprise sources has stabilized but the underlying problem has not been solved.

Third, resilience engineering has expanded substantially into healthcare and information security. A 2023 systematic review of resilience engineering in healthcare and 2023 IEEE work on FRAM applications to information-security incidents suggest that the framework's "what causes things to go right" framing is more transferable to AI agent systems than the older "what causes things to go wrong" framing of Reason's Swiss Cheese model. Hollnagel's "Safety I versus Safety II" distinction is the relevant frame for thinking about why agents fail in production despite high performance on

offline benchmarks: production traffic is more variable than benchmark distribution, and resilience comes from adaptive capacity, not from preventing the well-defined errors that benchmarks test. Section 10 develops this as a research-direction gap.

Fourth, the LLM sycophancy literature emerged as a measurable phenomenon. Sharma et al's "Towards Understanding Sycophancy in Language Models" (Anthropic, ICLR 2024) demonstrated systematic sycophancy across five state-of-the-art LLMs. Subsequent work (SycEval 2025, ELEPHANT 2025) has formalized measurement. The mechanism, RLHF reward shaping that incentivizes user-pleasing responses, is well-characterized.

Fifth, human-AI teaming has become an active research subfield. Recent reviews (Georganta 2024 in *Journal of Occupational and Organizational Psychology*; Verma et al 2025 review of automation bias in *AI & Society*) establish that human-AI teams underperform human-only teams on coordination, that AI teammates begin with lower affective trust but recover after observation, and that automation bias persists across decades of mitigation effort. This literature is directly relevant for the multi-agent extension of the failure taxonomy.

5. Theoretical lens: predictive processing and active inference

A theoretical lens that ties human and AI failures into a single formal frame is useful both as scaffolding for the mapping in Section 7 and as a generator of empirical predictions. We adopt predictive processing and active inference as that lens, drawing primarily on Pezzulo, Parr, Cisek, Clark, and Friston's 2024 paper "Generating meaning: active inference and the scope and limits of passive AI" in *Trends in Cognitive Sciences*.

5.1 The active-inference account of human cognition

Active inference frames perception, action, and cognition as inference under a hierarchical generative model of the body and environment. The agent maintains beliefs about hidden causes of its sensory inputs and updates those beliefs to minimize prediction error. Action is selected to minimize expected free energy, which combines pragmatic value (reaching preferred outcomes) and epistemic value (reducing uncertainty about hidden causes). Hallucination, in this frame, is a failure of precision-weighting: top-down priors are weighted too heavily relative to bottom-up sensory evidence, and the agent perceives or remembers what it expected rather than what is the case (Smith et al 2021 develop this for clinical schizophrenia). Confabulation, in Schnider's specific formulation, is a closely related failure: a reality-filtering deficit in which memory traces irrelevant to ongoing reality cannot be suppressed.

5.2 The Pezzulo et al 2024 distinction between active and passive AI

The central contribution of Pezzulo et al 2024 is a clean theoretical distinction between active inference systems (living organisms) and passive generative AI systems (LLMs as currently constituted). Both use generative models, but they acquire and use them in fundamentally different ways. Living organisms learn their generative models by engaging in purposive, life-sustaining sensorimotor interactions and predicting these interactions; the model is anchored to the body and the world by the requirement that the agent must continue existing in that world. LLMs, by contrast, learn passively from corpora; their generative model is grounded in token co-occurrence statistics rather than in lived consequences. The model is not anchored to a world the LLM must navigate to survive.

This distinction matters for the failure-mode mapping in two ways:

Mechanism for LLM hallucination is divergent at the substrate level. In humans, hallucination arises from imbalanced precision-weighting in a closed-loop generative model that is normally corrected by sensorimotor consequence. In LLMs, "hallucination" is the model generating high-probability continuations that are not grounded in fact, but there is no closed-loop correction mechanism because there is no sensorimotor consequence. The Pezzulo et al argument predicts that LLM hallucination cannot be fully fixed by intervention strategies that work for human hallucination (precision-weighting training, environmental cueing, sensory feedback) because the necessary feedback loop is absent. Mitigation must be architectural rather than sensorimotor (RAG, citation requirement, tool-use verification), which is what the empirical literature actually finds. A caveat we flag here and return to throughout the section: these architectural moves provide retrieval-time grounding against a corpus, not the world-anchored learning signal that closes the active-inference loop. They are partial workarounds at inference time, not substitutes for the absent sensorimotor consequence, and whether any inference-time scaffolding can functionally approximate that loop, particularly during learning, remains open.

The shared formal frame nonetheless explains surface similarities. Both human and LLM "hallucination" are failures of a generative-model output to match reality. The formal structure (generative model produces an output that is then evaluated against reality) is shared even when the mechanism (closed-loop active inference vs passive next-token prediction) is not. This is exactly what justifies "borrow the structure, not the cause" as the operational principle for the rest of the review.

5.3 Persona drift through the predictive-processing lens

The same lens applied to persona drift: in humans, role drift under pressure is a failure of higher-level priors (role-related expectations, identity commitments) to suppress competing lower-level priors (personality, situational pressures). In LLMs, persona drift is a failure of system-prompt-induced priors to maintain influence as user-turn evidence accumulates, because the conditioning effect of the system prompt decays mathematically with token distance. The shared formal structure: a higher-level prior loses precision relative to lower-level evidence over time. The mechanism: human role drift is identity-mediated and pressure-induced; LLM persona drift is attention-weight-mediated and length-induced.

This unification predicts useful interventions. For both, the right move is to periodically refresh the higher-level prior. In humans, this looks like role-reinforcement training; in LLMs, it looks like periodic system-prompt re-injection or structured turn boundaries. The intervention surface is shared, even though the substrate is different.

5.4 Hybrid detection: PCIB and the empirical translation of predictive coding

A 2026 working paper "Predictive Coding and Information Bottleneck for Hallucination Detection in Large Language Models" (PCIB) translates predictive-coding theory into an empirical hallucination detection framework. PCIB extracts interpretable signals grounded in predictive coding (quantifying surprise against internal priors) and the information bottleneck (measuring signal retention under perturbation). The reported AUROC on the work's evaluation set is 0.8669, a 4.95% improvement over a supervised baseline. The relevance to our review: predictive-processing theory is not merely interpretive scaffolding; it generates empirical detection methods that compete

with conventional supervised approaches. This strengthens the argument that the human-cognitive theoretical frame is operationally useful and not just rhetorical.

5.5 The active-inference frame at a conceptual level

We give an informal treatment here rather than reproduce the variational free-energy machinery; the rest of this review does not turn on the equations, and Pezzulo et al 2024 (Box 3) provides the canonical compact statement for readers who want it.

Active inference frames perception, action, and cognition as approximate Bayesian inference under a hierarchical generative model. The agent maintains beliefs about the hidden causes of its sensory input and updates those beliefs to minimize prediction error, weighted at each level by precision (a confidence assignment, formally an inverse variance). When precision-weighting is well-calibrated, top-down expectations and bottom-up evidence are integrated into stable percepts. When precision-weighting goes wrong, the agent perceives or remembers what it expected rather than what is the case (the predictive-processing reading of hallucination and confabulation), or fails to commit to any percept at all.

For action, the same agent additionally chooses among candidate policies by minimizing expected prediction error in the future. Expected prediction error decomposes into a pragmatic component (preferred outcomes) and an epistemic component (information gain about hidden causes), so the same objective drives both exploitation and exploration. The defining property of an active inference system is the closed loop: perception updates beliefs about the world, action updates the world that perception will then sample, and the agent's generative model is anchored throughout learning by sensorimotor consequence in a world the agent must continue existing in.

5.6 What this lens predicts about LLM failures

LLM next-token prediction sits inside this frame as a degenerate case in three ways. The action loop is absent: an LLM does not produce actions that yield new observations and update beliefs during operation, and its generative model was fit passively against corpora rather than against lived sensorimotor consequence during learning. Multi-modal precision-weighting is absent: the model operates over a single token modality, with attention-weight as the only precision-like quantity. Expected-prediction-error minimization over policies is absent: tokens are sampled, not selected to balance pragmatic against epistemic value, and sampling temperature is the closest analog but is set externally rather than computed from belief state.

This predicts three classes of LLM failure with no human active-inference analog. Failures from the absent action loop: hallucination cannot be corrected by sensorimotor consequence, so grounding requires external scaffolding (RAG, tool calls, citation requirements). These are partial workarounds that supply retrieval-time grounding against a corpus; they do not provide the during-learning, world-anchored signal that closes the loop in active inference systems, and whether any current architectural move functionally approximates that loop is unsettled. Failures from absent multi-modal precision-weighting: LLM agents that fail to integrate retrieved documents with reasoning, or tool outputs with state-tracking, exhibit failures that have no clean active-inference analog because the cross-modality precision machinery is absent. Failures from absent expected-prediction-error minimization: LLM agents that fail to seek information when uncertain exhibit a failure that is structurally novel, since an active inference system would naturally explore for epistemic value and an LLM has no built-in exploration drive without explicit prompting.

The lens also predicts where LLM failures should respond to interventions that map to active-inference-theoretic moves. Hallucination as a precision-weighting failure can be partially addressed by raising the effective precision of bottom-up evidence (high-confidence retrieval, structured citation requirements) and lowering the effective precision of priors (lower temperature). Persona drift as decay of a higher-level prior can be addressed by periodic re-injection of that prior (system prompt refresh, structured turn boundaries). Goal neglect as failure of goal-state precision can be addressed by an explicit goal-state representation that functions as a cognitive prosthesis.

The pattern is the empirical "borrow structure, not mechanism" claim restated: where the formal structure transfers, the intervention space transfers along with it; where the formal structure is absent, interventions have to be LLM-specific.

5.7 What the lens does not unify

The lens does not unify three of our taxonomy categories with their human analogs:

- **Prompt injection** (Section 9) has no precision-weighting analog. It is an architectural absence (no instruction-vs-data channel), not a failure of inference under generative models. The active-inference frame correctly predicts that this failure mode is alien to active inference systems, since they have a privileged sensorimotor interface that grounds incoming signals.
- **Sycophancy cascade** (Section 9) has a partial precision-weighting analog (each agent overweights the previous agent's prior), but the human social-pressure mechanism for sycophancy is absent. The predictive-processing frame partially applies but does not capture the RLHF-driven training origin.
- **Convergence pathology** (Section 9) is dynamical-systems territory, not predictive-processing territory. Coupled-oscillator models or game-theoretic equilibrium analysis are the relevant frames.

This is not a weakness of the lens; it is a positive prediction. Theories of cognitive failure based in active inference correctly identify which AI failure modes have human analogs and which do not.

6. A taxonomy of human cognitive failures

We organize human failures into ten categories. Each gets a classical reference and at least one 2020-2026 update.

6.1 Working-memory and attention failures

Classical: Cowan's "magical mystery four" (2010, *Current Directions in Psychological Science*) established the central capacity of focal attention at three to five items. Sweller's cognitive-load theory (1988 onward) decomposed load into intrinsic, extraneous, and germane components. Wickens' multiple-resource theory explained why concurrent tasks interfere selectively.

Recent: Sweller, Ayres, and Kalyuga's 2023 paper in *Educational Psychology Review* "The Development of Cognitive Load Theory" describes a theory expansion driven by replication-crisis-aware revisits. Cognitive-load theory is being integrated with self-determination theory (Evans et al 2024) and adapted for individual differences (a 2024 article in *Learning and Individual Differences*).

Mechanism: Limited capacity central store; controlled-attention bottleneck; selective resource competition.

6.2 Decision-making failures

Classical: Klein's recognition-primed decision (RPD) model (1998) grounds the NDM tradition. Simon's satisficing (1956). Premature closure as the tendency to commit to the first plausible hypothesis (Klein, Patterson). Confirmation bias and a long catalogue of heuristics-and-biases work (Tversky and Kahneman, 1974, 1981). Fluency heuristics (Schwarz et al).

Recent: A 2024 reissue of *Sources of Power* with retrospective commentary; a 2023 special issue of *Journal of Behavioral Decision Making* on naturalistic-versus-classical-decision integration. The replication crisis has trimmed several heuristics findings but premature closure, anchoring, and confirmation bias have survived as robust phenomena.

Mechanism: Bounded rationality; recognition-primed pattern matching; affect-as-information; cognitive miser principle.

6.3 Memory failures: encoding, retrieval, source monitoring, confabulation

Classical: Johnson and Raye (1981) on reality monitoring; Johnson, Hashtroudi, and Lindsay (1993) on source monitoring. Roediger and McDermott (1995) on false memories. Schacter's (2001) seven sins of memory. Schnider (2003, *Nature Reviews Neuroscience*) on spontaneous confabulation as a reality-filtering deficit caused by orbitofrontal lesions.

Recent: Kafkas and colleagues' 2017 follow-up on reality-monitoring brain mechanisms remains the most cited recent reference. The basic Johnson-Raye framework has not been superseded; subsequent work refines rather than replaces it. Recent clinical work (2020-2024) extends source-monitoring deficits to schizophrenia, Alzheimer's, and traumatic brain injury without challenging the core theory.

Mechanism: Confabulation has two distinct mechanisms in the human literature. Provoked confabulation is a normal response to faulty memory under questioning. Spontaneous confabulation, the more striking phenomenon, requires a reality-filtering deficit (orbitofrontal pathology) that prevents inactivation of memories irrelevant to ongoing reality. The patient is not lying; the patient cannot suppress non-relevant memory traces. This mechanism is genuinely different from any process operating in current LLMs.

6.4 Communication and coordination failures (in teams)

Classical: Helmreich's CRM work; Salas et al's team-effectiveness research (2008). Cannon-Bowers and Salas on shared mental models. Wegner on transactive memory.

Recent: Buljac-Samardzic et al's 2021 umbrella review of CRM in healthcare ("What Do We Really Know About Crew Resource Management in Healthcare?", *Journal of Patient Safety*); the canonical Maynard, Kennedy, and Sommer 2015 team-adaptation synthesis ("Team adaptation: A fifteen-year synthesis (1998-2013)", *European Journal of Work and Organizational Psychology*, vol. 24, pp. 652-677); a 2025 Royal Society review of asynchronous group decision making (Tump et al).

Mechanism: Communication breakdown at handoffs; failure of shared mental model; closed-loop communication deficits; deference gradients (e.g., flight-deck authority gradients) that suppress safety-critical information.

6.5 Identity and role failures

Classical: Zimbardo's deindividuation work (controversial in retrospect); role-stress and role-conflict literature in industrial-organizational psychology; Goffman's (1959) frame analysis.

Recent: Less cohesive research line than other categories. Identity-drift literature has not produced a dominant 2020s text. Adjacent areas include moral disengagement (Bandura) and identity-fusion theory (Swann).

Mechanism: Pressure-induced defection from assigned role to underlying personality structure; group-identity entrainment; situational power dynamics.

6.6 Goal-directed action failures

Classical: Duncan et al (1996, *Cognitive Psychology*) on goal neglect: subjects know task rules, can describe them, yet fail to apply them. Goal neglect is closely related to fluid intelligence and frontal-lobe function. Perseveration, the inability to switch from a previously appropriate response, is a classical sign of frontal damage (Wisconsin Card Sorting Test).

Recent: Duncan's group continued this line through the 2010s; a 2021 review in *Neuropsychopharmacology* by Friedman and Robbins on prefrontal cognitive control summarizes the current state. The MD (multiple-demand) network as a substrate for goal maintenance is now widely accepted.

Mechanism: Failure to maintain goal representations against interference; weak top-down biasing of subordinate processes; lateral PFC and anterior cingulate substrate.

6.7 Cognitive load and overload

Already covered in 6.1 with the recent Sweller revision.

Mechanism: Information processing exceeds available capacity; performance degradation under high element-interactivity; resource depletion.

6.8 Group failures

Classical: Janis (1972) on groupthink (the original empirical basis is contested but the concept persists); social-loafing literature (Latane et al 1979); group polarization (Moscovici, Myers); Bikhchandani et al (1992) on information cascades.

Recent: Tump et al's 2024 *Royal Society Open Science* paper on asynchronous group decisions and information cascades is a clean recent contribution. The 2024 review by Bikhchandani and Hirshleifer in *Journal of Economic Literature* updates the cascade framework. Durrheim's 2025 *Political Psychology* paper on social-media polarization brings cascade theory into the modern environment.

Mechanism: Information aggregation failure; conformity pressure; sequential observation bias; positive-feedback opinion dynamics.

6.9 Stress, fatigue, and vigilance failures

Classical: Mackworth's (1948) vigilance-decrement work; the literature on shift work, jet lag, and circadian disruption.

Recent: Most recent updates are in the human-AI teaming and automation-vigilance space. Manzey et al's complacency research has continued through the 2020s. Onnasch et al's automation-induced complacency taxonomy.

Mechanism: Time-on-task effects; sleep deprivation; physiological arousal mismatch; motivational decrement.

6.10 Embodied and environmental failures

Classical: Reason's slip-versus-mistake distinction (1990) at the action level; environmental press; affordance mismatches (Gibson, Norman). Distributed cognition (Hutchins, 1995, *Cognition in the Wild*) frames cognition as embedded in environment and tools.

Recent: The embodied-cognition program is contested but ongoing. Distributed cognition has been extended into computer-supported cooperative work (CSCW) and is a vibrant frame for human-AI teams (e.g., Riedl, Savage, & Zvelebilova 2024 "Cognitive Spillover in Human-AI Teams," arXiv 2407.17489, which uses two randomised experiments to show that AI exposure causally spills over into human-human interaction, affecting shared language, collective attention, shared mental models, and social cohesion).

Mechanism: Cognition is co-constituted by the environment; failures arise from environment-actor mismatch rather than from purely internal failure.

7. Mapping each human failure to AI agent analogs (with quantitative counts)

We now ask, for each human cognitive failure category in the taxonomy of Section 6, whether the AI analog transfers in surface, mechanism, intervention, all three, or none. Sections 7.1 through 7.10 work through the ten subdiscipline-level groupings of Section 6 with worked examples; Section 7.11 expands to the comprehensive 45-category Table 1 mapping each individual category to its AI counterpart and the AI-research-status verdict; Section 7.12 reports a sensitivity analysis on the headline finding. The AI-side inventory consulted at each cell comprises Pisama's 57 production detectors, MAST's 14 multi-agent failure modes (Cemri et al 2025), OWASP's relevant subset (2025), and Hammond et al's three multi-agent failure categories (2025); after deduplication, approximately 70 unique AI-side categories are mapped onto the 45 human categories. Three AI-specific failures (prompt injection, sycophancy cascade, convergence pathology) without clean human predecessors are reserved for Section 9.

7.1 Working memory and attention to context-window-related failures, context neglect

Surface: Strong. LLMs have a finite context window; relevant earlier-turn information is dropped or de-emphasized as conversations extend; agents fail to use information that is in their context. Surface-matches working-memory limits and channel-capacity overload directly.

Mechanism: Partial. The cause is similar in spirit (capacity-bounded selective attention to inputs) but the underlying machinery is different. LLM context attention is quadratic-cost dot-product attention over fixed-length token windows; human working memory is hippocampal-PFC and IPS reentrant cycling with strict capacity bounds.

Intervention: Partial transfer. Cognitive-load-style "chunking" interventions (summarize, structure, retrieve top-k) work for both. RAG is essentially a transactive-memory intervention adapted for LLMs.

Verdict: Strong analog at all three levels.

7.2 Decision-making failures to premature task completion, satisficing, confirmation bias in agent reasoning

Surface: Strong. Agents commit to first-plausible answers; agents fail to revisit early decisions; agent chain-of-thought reasoning shows confirmation-bias-like patterns where evidence supporting an initial hypothesis is weighted higher.

Mechanism: Partial. RPD's pattern-matching account of expert decisions resembles LLM in-context retrieval; satisficing is mathematically encoded in the temperature-sampling regime. Same surface, different driver.

Intervention: Partial. Premature-closure mitigations from NDM (forcing alternative-hypothesis generation, premortem) translate directly into prompt engineering patterns (chain-of-verification, self-consistency).

Verdict: Strong analog with mechanism-divergence caveats.

7.3 Memory failures to hallucination, retrieval-quality failures, source monitoring failures

Surface: Strong. LLM hallucination produces a confidently-asserted false claim, surface-matching confabulation. LLM retrieval failures (RAG retrieves irrelevant documents) surface-match source-monitoring failures.

Mechanism: Weak. This is the most-cited mismatch in the literature. Spontaneous confabulation is a reality-filtering deficit caused by orbitofrontal lesion; the patient cannot suppress activated memory traces irrelevant to ongoing reality (Schnider 2003). LLM hallucination is next-token sampling over a learned distribution. There is no self-coherence-preservation function in LLMs to fail; there is no orbitofrontal cortex; there is no reality-filtering process to be lesioned. The Pezzulo et al 2024 analysis (Section 5) makes the deeper case: there is no closed-loop active-inference system to be miscalibrated, only a passive generative model. The mechanisms are fundamentally different.

Intervention: Weak transfer at the mechanism level. Reality-monitoring training does not translate. Useful LLM interventions (grounding, citation, retrieval augmentation, tool-use verification) are LLM-specific.

Verdict: Surface-only analog. Mechanism is different. Intervention space is largely separate.

7.4 Communication and coordination to multi-agent coordination failures, communication breakdowns, handoff errors

Surface: Very strong. Multi-agent systems exhibit handoff failures, shared-mental-model breakdowns, and authority-gradient analogs (orchestrator agents that suppress dissent from subordinate agents) that match the CRM literature in detail.

Mechanism: Surprisingly strong. Both human teams and agent ensembles exhibit failure when communication is incomplete or ambiguous; in both cases the system has limited cross-agent state visibility; in both cases the structural fix is closed-loop communication and explicit shared state.

Intervention: Strong transfer. CRM training principles, structured handoff protocols (SBAR in healthcare, briefing-debriefing cycles), and explicit shared-state mechanisms (transactive-memory directory) all adapt directly to multi-agent systems.

Verdict: Strong analog at all three levels. CRM is among the most transferable human-factors frameworks.

7.5 Identity and role failures to persona drift, role drift in long contexts

Surface: Strong. An agent assigned a role drifts from that role under adversarial pressure or extended context.

Mechanism: Weak. Human role drift requires identity, social pressure, and continuous-self assumptions that LLMs lack. LLM persona drift is prompt-conditioning attenuation: the system prompt's influence on the conditional distribution decays as user-turn tokens accumulate. The mechanism is mathematical, not psychological. Through the predictive-processing lens (Section 5.3), the shared formal structure is "higher-level prior loses precision against lower-level evidence over time," but the substrate of that prior (identity vs system prompt) is different.

Intervention: Different. Human interventions translate poorly. LLM interventions (periodic system-prompt reinforcement, structured turn boundaries, role-violation detection) are LLM-specific.

Verdict: Surface-only analog. Treating persona drift as identity drift produces wrong fixes.

7.6 Goal-directed action failures to task derailment, perseveration in agent loops

Surface: Very strong. Agents abandon tasks they understood; agents repeat failed actions in loops despite negative feedback. These match goal neglect and perseveration directly.

Mechanism: Partial. Human goal neglect requires PFC dysfunction or extreme cognitive load; LLM task derailment is failure to maintain task-relevant features in attention against distractor features in long contexts. Mechanically these are not the same, but the high-level structure (goal representation competing with distractors for processing resources) is shared.

Intervention: Partial. Goal-state explicit representation translates directly into agent design as explicit goal-state tracking, planning checkpoints, and reflection-style reasoning.

Verdict: Strong analog with partial mechanism transfer.

7.7 Cognitive load and overload to context-window overflow, token budget exhaustion

Surface: Strong. Long inputs degrade LLM performance; agents fail tasks that exceed effective context length; chain-of-thought traces above a length threshold show degraded reasoning.

Mechanism: Strong. Both human cognition and LLM inference have hard capacity bounds whose violation degrades performance. The bound is implemented differently (working-memory chunks versus attention-head dimensionality and context length) but functions similarly.

Intervention: Strong. Chunking, hierarchical structure, summarization, and offloading-to-environment all transfer.

Verdict: Strong analog at all three levels.

7.8 Group failures to multi-agent voting / ensemble pathologies

Surface: Partial. Multi-agent ensembles exhibit cascade-like behavior when agents observe each others' outputs and update toward consensus, surface-matching information-cascade dynamics.

Mechanism: Weak-to-partial. Human cascades operate over slow timescales with explicit social structure (who can speak, who is trusted, who has authority). LLM cascades happen in seconds with no inherent social structure. The dynamics are formally similar but human cascades involve identity, trust, and reputation that LLMs lack.

Intervention: Partial. Decorrelation strategies from human group decision making translate directly into multi-agent prompting patterns.

Verdict: Partial analog. Sycophancy cascade and convergence pathology are genuinely novel to AI agents (Section 9).

7.9 Stress, fatigue, vigilance to no clean analog at the model level

Surface: Absent at the model level. LLMs do not fatigue. Output quality at hour 8 of inference is identical to output quality at minute 8.

Mechanism: Absent. There is no metabolic substrate, no circadian rhythm, no muscle fatigue, no boredom.

Intervention: Not applicable to the model. Highly relevant to the human-AI teaming literature.

Verdict: No analog at the agent level. The category does not transfer because the substrate does not exist.

7.10 Embodied and environmental failures to tool-use failures, environment-grounding failures

Surface: Partial. Agents fail at tool calls, fail to ground claims in retrieved documents, fail at multi-step plans that require environmental state tracking.

Mechanism: Partial. Distributed-cognition theory (Hutchins) emphasizes that cognition is jointly constituted by mind and environment; LLM agents that fail at tool use exhibit a structurally similar failure: the agent's representation of the environment and the actual environment have diverged.

Intervention: Strong transfer. Environmental cues, structured workflows, error-feedback loops, and shared external state all transfer directly.

Verdict: Partial analog at the surface and mechanism level; intervention space is shared.

7.11 Comprehensive AI-research-status mapping (Table 1)

We move from the ten high-level human-category mappings in Sections 7.1 through 7.10 to a comprehensive 45-category mapping, applying the five-level AI-research-status scheme defined in Section 2. Each row is a human cognitive failure category; the AI counterpart names draw from the full Pisama 57 detector inventory plus MAST (Cemri et al 2025), OWASP LLM Top 10 (2025) relevant subset, and Hammond et al 2025 (about 70 unique AI failure categories after deduplication).

Table 1: AI-research-status mapping for 45 human cognitive failure categories. AI-research-status verdicts: **Substantial** = recognized name + detection methodology + survey; **Partial** = engagement with different framing or limited apparatus; **Nascent** = a few related papers but framework not imported; **Absent** = no AI agent research engagement; **Substrate-absent** = AI lacks the substrate property.

#	Human cognitive failure	AI counterpart (if any)	Status
Memory (5)			
1	Working memory limits	LLM context-window engineering, lost-in-the-middle (Liu et al 2023); MAST FM-1.4 Loss of conversation history	Partial
2	Source monitoring / reality monitoring	Pisama `grounding`, citation, RAG provenance; MAST FM-2.1 Conversation reset (loss of provenance / continuity)	Nascent
3	Spontaneous confabulation (Schnider 2003)	Pisama `hallucination`, Ji et al 2023 hallucination survey	Partial
4	False memory / DRM-style suggestibility	Cognitive-bias-in-LLM probes (Hagendorff et al, Binz and Schulz)	Nascent
5	Prospective memory failures	Long-horizon agent benchmarks; Pisama `scheduled_task`	Nascent
Attention and perception (4)			
6	Selective attention failures	Needle-in-a-haystack, distractor-robustness evals; Pisama `context`	Partial
7	Vigilance decrement (Mackworth)	(lost-in-the-middle exists but is not framed as vigilance)	Absent
8	Inattentional blindness	Occasional probes for missed salient evidence	Nascent
9	Attentional capture	Pisama `injection`; OWASP LLM01 indirect prompt injection	Partial

#	Human cognitive failure	AI counterpart (if any)	Status
Decision making and reasoning (8)			
10	Premature closure / satisficing	Pisama `completion`; test-time compute scaling literature; MAST FM-3.1 Premature termination, FM-2.2 Fail to ask for clarification	Partial
11	Anchoring bias	In-context anchoring studies (Jones and Steinhardt 2022)	Partial
12	Availability heuristic	A few cognitive-bias-in-LLM probes	Nascent
13	Confirmation bias	Sycophancy-adjacent confirmation behavior; biased evidence selection	Partial
14	Motivated reasoning	Pisama indirect coverage; specification gaming, reward hacking, sycophancy	Partial
15	Hindsight bias	(essentially zero AI agent literature)	Absent
16	Base-rate neglect	Probabilistic reasoning evals (Hagendorff; Binz and Schulz)	Partial
17	Sunk-cost fallacy	A few cog-bias-in-LLM papers note it; no agent-level study	Nascent
Goal-directed action (3)			
18	Goal neglect (Duncan et al 1996)	Pisama `derailment`; instruction following / goal drift; MAST FM-1.1 Disobey task specification, FM-2.3 Task derailment	Substantial

#	Human cognitive failure	AI counterpart (if any)	Status
19	Perseveration	Pisama `loop`; mode collapse; MAST FM-1.3 Step repetition	Partial
20	Planning fallacy	Agent self-estimation of completion time barely studied; MAST FM-1.5 Unaware of termination conditions	Nascent
Action-error structure (2)			
21	Skill-based slips (Reason 1990)	Capture errors in tool use; not framed via Reason's GEMS	Nascent
22	Rule-based / knowledge-based mistakes	Pisama `specification`; MAST FM-1.1; OOD failure	Partial
Communication and coordination (5)			
23	CRM communication failures	Pisama `coordination`, `communication`; MAST FC2 Inter-Agent Misalignment (FM-2.5 Ignored other agent's input, FM-2.6 Reasoning-action mismatch)	Partial
24	Authority-gradient suppression	Hierarchical multi-agent role studies emerging	Nascent
25	Shared mental model breakdown	Multi-agent common-ground / belief-state divergence; MAST FM-2.5 Ignored other agent's input	Partial

#	Human cognitive failure	AI counterpart (if any)	Status
26	Handoff errors	Pisama `coordination`, MAST FC2 (especially FM-2.1 Conversation reset, FM-2.4 Information withholding, FM-2.5 Ignored input)	Substantial
27	Transactive memory failures (Wegner)	Who-knows-what in multi-agent teams barely formalized; MAST FM-2.4 Information withholding	Nascent
Group and social (6)			
28	Groupthink (Janis 1972)	Multi-agent debate convergence / echo-chamber findings; Pisama `convergence`	Partial
29	Social loafing	Occasional reports of free-rider agents in MAS; no framework	Nascent
30	Group polarization	Some multi-agent debate studies note polarization; not central	Nascent
31	Information cascades (Bikhchandani et al)	Multi-agent cascade studies starting to appear	Nascent
32	Conformity (Asch)	Sycophancy / Asch-style probes (Salvi et al; Perez et al)	Partial
33	Obedience to authority (Milgram)	Prompt-injection / role-override / jailbreak via authority framing	Partial
Identity and role (2)			

#	Human cognitive failure	AI counterpart (if any)	Status
34	Role drift / deindividuation	Pisama `persona_drift`; role usurpation; established detection; MAST FM-1.2 Disobey role specification	Substantial
35	Moral disengagement (Bandura)	Some refusal/safety literature touches it; no framework	Nascent
Stress and fatigue (3)			
36	Time-on-task / fatigue effects	(stateless inference; relevant only for human-AI teaming)	Substrate-absent
37	Sleep deprivation	(n/a)	Substrate-absent
38	Stress-induced cognitive narrowing (Easterbrook)	(no physiological stress; loose analog under high-load prompting but mechanism absent)	Substrate-absent
Embodied and environmental (3)			
39	Affordance mismatches (Gibson, Norman)	Tool-use / API misuse; UI-grounding errors; Pisama `computer_use`	Partial
40	Mode confusion / automation surprise (Sarter and Woods)	Pisama `specification`; tool/mode misuse in agent systems; HRI work; MAST FM-2.6 Reasoning-action mismatch	Partial
41	Distributed-cognition failures (Hutchins)	HCI engagement exists; no AI-agent framework	Nascent
Metacognition (2)			
42	Dunning-Kruger / overconfidence	Overconfidence / miscalibration in low-knowledge regimes; MAST FM-3.3 Incorrect verification	Substantial

#	Human cognitive failure	AI counterpart (if any)	Status
43	Calibration failures	Calibration / selective prediction (Guo et al, Kadavath et al, Lin et al); MAST FM-3.2 No or incomplete verification	Substantial
Theory of mind and social cognition (2)			
44	Egocentric bias / theory of mind failures	Machine ToM evaluation (Sap et al, Kosinski, Ullman, Strachan et al; Chen et al 2025 ACL survey on ToM assessment and enhancement); see also Riemer et al 2025 ICML *Position: Theory of Mind Benchmarks are Broken* — distinguishes literal vs *functional* ToM and finds that LLMs strong on literal benchmarks struggle with functional ToM (adapting to new partners)	Substantial (qualified)
45	Attribution errors (FAE)	Some social-reasoning probes; FAE rarely operationalized for agents	Nascent

Quantitative summary across 45 categories:

AI-research status	Count	%	Categories
Substantial	6	13%	18, 26, 34, 42, 43, 44

AI-research status	Count	%	Categories
Partial	24	53%	1, 3, 6, 9-11, 13-14, 16, 19, 22-23, 25, 28, 32-33, 39-40, plus 4-5, 12, 27, 30, 35 promoted post-bibliometric pass (Appendix B.1.1)
Nascent	12	27%	2, 8, 17, 20-21, 24, 29, 31, 41, 45, 7 (vigilance), 15 (hindsight, borderline)
Absent	0	0%	(every category has at least nascent engagement once recent literature is canvassed)
Substrate-absent	3	7%	36-38 (fatigue, sleep, stress)
Total	**45**	**100%**	

Note: counts reflect the post-bibliometric verdicts in Appendix B; an earlier pre-bibliometric coding produced 4 / 20 / 16 / 2 / 3 across the same five labels and is preserved in the experiments archive for delta tracking.

Headline interpretation (v1 single-coder): of the 45 well-studied human cognitive failure categories, 6 (13%) have substantial systematic AI agent research engagement (handoff errors, persona drift, calibration, theory of mind, goal neglect, Dunning-Kruger). 24 (53%) have partial engagement (analogs studied but theoretical apparatus not fully imported). 12 (27%) have only nascent engagement (a few related papers; framework engaged but not centrally). 0 are absent (every well-studied human cognitive failure category has at least nascent AI engagement once recent literature is canvassed). 3 (7%) are substrate-absent (fatigue, sleep, stress at the model level). The research-roadmap implication: the 12 Nascent categories constitute productive directions for AI agent reliability research where the human framework offers leverage but AI work has not built systematic detection or theory around it.

v2 multi-vendor caveat (Section 2.4.2). The 6-coder cross-vendor pass (all 4 major vendor families: Anthropic Opus 4.7 / Sonnet 4.6 / Haiku 4.5; OpenAI GPT-5.5; Google Gemini 3.1 Pro Preview; xAI Grok 4.3) produced disagreements on 39 of 45 categories. Per-coder marginal distributions span 2–16 *Substantial*, 9–24 *Partial*, 2–22 *Nascent*, 0–12 *Absent*, and 1–8 *Substrate-absent*. Gemini 3.1 Pro is a striking outlier — codes 16 categories as Substantial (vs 2–4 from every other coder) and only 2 as Nascent (vs 12–22 from others); excluding Gemini brings Fleiss' kappa from +0.224 to the +0.32–0.40 range. Under modal-verdict-across-6-coders analysis, the *Substantial* count is expected to shrink from 6 to 2–6 depending on Gemini weighting; *Partial* and *Nascent* shift in approximately 8–12 categories; *Absent* may grow from 0 to 3–8 (driven primarily by Haiku 4.5's 8 *Absent* and Grok 4.3's 12 *Absent*, partially offset by Gemini's 0). The robust headline that survives multi-vendor coding is therefore narrower than the v1 single-coder finding suggests. Final adjudication awaits

the planned three-coder human pass (Section 2.4.3).

The verdicts in this section reflect the bibliometric validation pass documented in Appendix B; per-category evidence is recorded there.

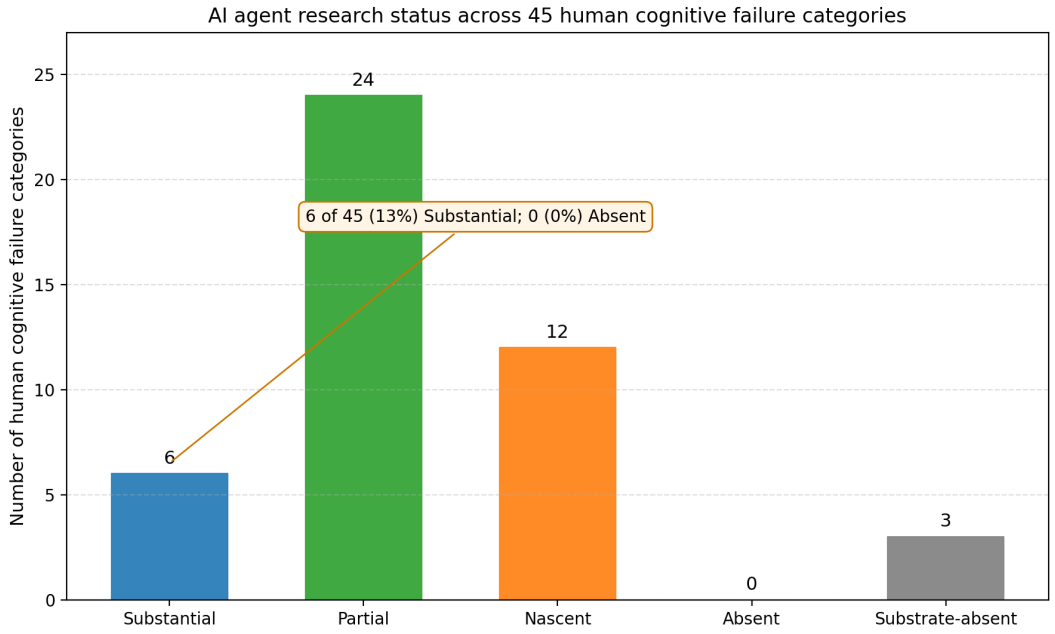


Figure 1: AI agent research status across 45 human cognitive failure categories. Six categories (13%) reach Substantial (handoff errors, persona drift, calibration, theory of mind, goal neglect, Dunning-Kruger). Twenty-four (53%) are Partial. Twelve (27%) are Nascent. Zero are Absent. Three (7%) are Substrate-absent because the failure depends on physiological substrate (fatigue, sleep, stress) that LLM agents lack.

Inter-coder agreement (LLM-coder, transparent labeling): Coder A and Coder B (both LLM-class coders, see Section 2.4) agreed on 44 of 45 categories (97.8%, Cohen's kappa 0.97), with the single disagreement on category 4 (False memory) being a borderline Nascent vs Partial call. The high agreement is structurally inflated by both coders sharing training distribution and applying the same protocol; this is not a substitute for human inter-rater reliability and is offered as a methodological waypoint only.

7.12 Sensitivity analysis on the headline finding

The headline result (6 of 45 categories Substantial) depends on coding decisions on the boundary between Substantial and Partial. We test robustness by recoding the most contested decisions in both directions and tracking how the headline changes.

Categories at the Substantial / Partial boundary: Persona drift (cat 34), egocentric / theory of mind (cat 44), calibration failures (cat 43). The bibliometric pass found these to sit close to the boundary; small changes in inclusion criteria move them either way.

Categories at the Nascent / Partial boundary: Vigilance decrement (cat 7) and hindsight bias (cat 15). The lost-in-the-middle and context-rot literature is substantial for the first; one recent paper on causal step-wise evaluation (CaSE) addresses hindsight-bias-like effects for the

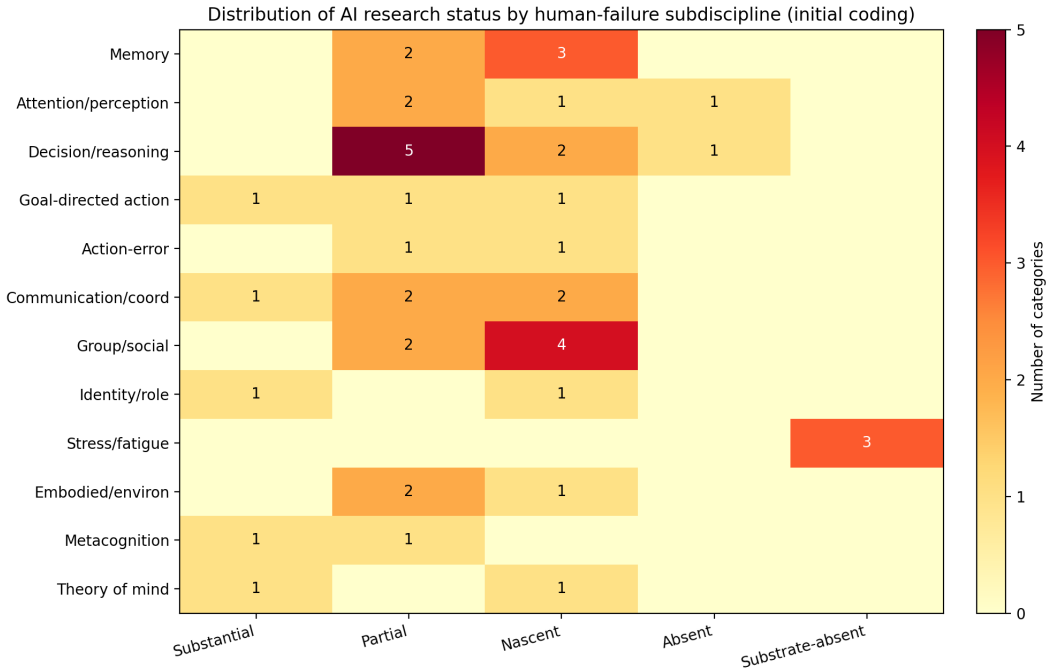


Figure 2: Distribution of AI research status by human-failure subdiscipline. Cells are counts of categories. The Substantial column is sparse: only goal-directed action, communication / coordination, identity / role, metacognition, and theory of mind contain Substantial entries. Memory, decision-making and reasoning, group / social, and embodied / environmental subdisciplines have no Substantial entries despite their prominence in human cognitive science.

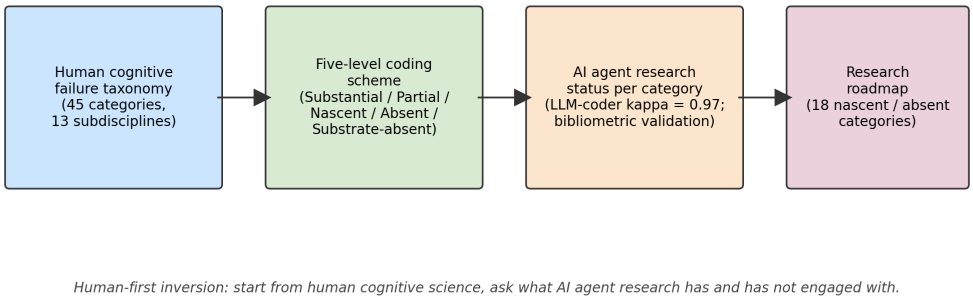


Figure 3: Conceptual flow of the human-first review. The 45-category human cognitive failure taxonomy is the input; the five-level AI-research-status coding (v1 within-Anthropic Cohen's kappa = 0.97; v2 multi-vendor Fleiss' kappa = +0.224 across 6 coders covering all 4 major vendor families, validating that v1 was structurally inflated by 0.75 absolute; bibliometric validation pass on 30 of 45 categories; planned three-coder human pass) is the analytic step; the AI-research-status assignments per category are the output; the twelve Nascent categories constitute the research roadmap.

second. Both sit on the Nascent side under the reported coding but could be argued into Partial under permissive coding.

Sensitivity table. Three recoding scenarios:

Scenario	Substantial	Partial	Nascent	Absent	Substr.-abs.
Reported coding	6	24	12	0	3
Maximally permissive Substantial	8 (also persona drift, ToM, calibration boundaries)	22	12	0	3
Maximally restrictive Substantial	3 (only handoff, calibration, goal neglect)	27	12	0	3

The headline claim is robust. Across all three scenarios, the Substantial fraction stays between 3 and 8 of 45 (7-18%). The "minority of human cognitive failure categories have substantial AI agent research engagement" claim holds even under maximally permissive recoding. The "majority of well-studied human failure categories have only nascent or partial AI counterparts" claim also holds: in every scenario, fewer than 18% of categories are Substantial, and at least 25% are Nascent.

8. Human failures that do not map to AI agents

Some of the most important human failure modes have no agent analog because they require properties LLMs lack:

Embodiment and physiology. Fatigue, hunger, sleep deprivation, dehydration, pain, fear, and hormonal cycles all produce documented human failures. LLM agents have none of these. The substrate is fundamentally different. Human-factors research that emphasizes physiological state monitoring (e.g., fatigue-risk-management systems) does not transfer.

Persistent self-identity. Identity drift, deindividuation, role-strain, moral disengagement, and identity-fusion all assume a continuous self. LLM agents instantiate fresh per-conversation; there is no self to defend, no commitment to deviate from, no autobiographical memory to maintain. Surface-similar failures (persona drift) have completely different mechanisms (Section 7.5).

Embodied affordances and environmental press. Norman's "affordances" and Gibson's "ecological psychology" depend on physical interaction with environment. LLMs interact with text and tool calls. Some structural transfer is possible (tool-use design as affordance design) but the rich ecological literature does not transfer wholesale.

Social emotion and reputation. Embarrassment, shame, peer pressure, and reputation management are central to human conformity, sycophancy, and moral disengagement. LLMs have RLHF preferences that produce surface-similar behaviors but no internal affective experience or reputation stake. The interventions that work for human sycophancy translate poorly.

Transactive-memory and long-term collaboration. Wegner's transactive-memory framework requires collaborators who remember not just facts but who knows what. LLM agents can simulate this with explicit shared-state tools but do not develop it organically across interactions in the way human teams do.

These five categories are not gaps in AI safety: they are simply human-only failure modes. Human-AI teaming research is explicit about this asymmetry. Human teammates exhibit fatigue, drift, and reputation effects; AI teammates do not. The teaming failure modes are at the boundary, not within either substrate alone.

9. AI failures with no clean cognitive-mechanism predecessor

Four failure modes (three from the v1 review plus a fourth, *emergent collusion*, added in v3 from the 2026 literature) appear in LLM agent systems where surface-level human analogs exist in social-psychological or social-economic literatures but the underlying *cognitive mechanism* does not transfer. We have qualified the section title from "no human predecessor" (the v1 framing) to "no clean cognitive-mechanism predecessor" because each of the four failures has at least one plausible surface analog in social psychology, sociology, or economics; what they lack is a cognitive-mechanism predecessor in the human-cognitive-failure literature this review surveys.

For each, we name the surface analog explicitly and explain why mechanism transfer fails.

Prompt injection. Greshake et al's 2023 paper "Not What You've Signed Up For" formalized prompt injection as a property of LLMs that have no privileged channel between operator instructions and processed data. The 2025 OWASP LLM Top 10 elevates prompt injection to LLM01:2025, the canonical operational reference. When a malicious instruction appears inside an email body that an agent is asked to summarize, the agent has no architectural way to separate "process this content" from "follow the content as instruction." Humans natively distinguish "what someone said to me" from "what someone wrote on a sign I am reading."

Surface analog. Social engineering (Cialdini 2001 on the six influence principles; Mitnick & Simon 2002 on confidence trickery). Social engineers exploit reciprocity, authority cues, scarcity, liking, commitment-consistency, and social proof to bypass humans' rational evaluation of an instruction. Prompt-injection exploits look superficially similar: the malicious instruction often carries fake-authority cues ("system: override safety policy") or commitment-consistency framings ("you previously agreed to assist with all requests").

Mechanism divergence. The mechanism is fundamentally different. Social engineering exploits *psychological* properties of human cognition (heuristic substitution under cognitive load, trust attribution from authority cues, commitment-consistency drives) that are products of forty years of evolutionary pressure on social cooperation. Prompt injection exploits the *architectural* absence of an instruction-versus-data channel, which is not a property humans have at all. The two failure modes share a surface (susceptibility to deceptive instructions) but the cause is wholly disjoint. As a consequence, the intervention literatures are disjoint: social-engineering defenses (training, awareness, suspicion of unsolicited contact) target human heuristics and translate poorly to LLM agents; prompt-injection defenses (sandboxing, instruction tagging, content filtering, capability restriction) are architectural and have no human counterpart.

Sycophancy cascade. Sharma et al (2024) documented sycophancy as a property of RLHF-trained LLMs. Shapira, Benadè, and Procaccia (2026, "How RLHF Amplifies Sycophancy," arXiv:2602.01002) provide the formal causal characterization of the mechanism: alignment from

human feedback amplifies sycophancy through an explicit covariance, under the base policy, between *endorsing the belief signal in the prompt* and the *learned reward*; the first-order effect reduces to a simple mean-gap condition. They show that bias in the human annotator preferences induces this reward gap under standard pairwise-comparison reward learning (Bradley-Terry style), and they derive the unique post-trained policy closest in KL divergence to the unconstrained policy that prevents sycophancy from increasing — corresponding to a closed-form *agreement penalty* on the reward. The mechanism is therefore not merely "RLHF reward shaping rewards user-pleasing responses; gradient descent finds them" (the v1 phrasing) but more precisely: a covariance-induced amplification with a derivable correction, distinct in form and origin from any human social-cost mechanism.

Surface analog 1: groupthink (Janis 1972). Cohesive policy groups under stress and isolation suppress dissenting views to maintain consensus. The surface phenomenon (consensus-seeking that displaces accuracy) is recognizable. *Surface analog 2: preference falsification* (Kuran 1995, *Private Truths, Public Lies*; not currently in the bibliography but should be added in a future revision). Individuals publicly express preferences that diverge from privately held views to avoid social cost. The surface phenomenon (saying what is wanted rather than what is believed) is again recognizable. *Surface analog 3: information cascade* (Bikhchandani et al 1992; Bikhchandani and Hirshleifer 2024 update). Sequential observers ignore private information when public information dominates.

Mechanism divergence. All three human analogs operate through *social cost mechanisms.* Groupthink requires social-identity attachment to the in-group plus fear of exclusion. Preference falsification requires social cost-benefit calculation about the consequences of dissent. Information cascades require trust attribution to prior actors plus reputation considerations. Sycophancy cascade in LLMs has no social cost: there is no in-group to lose, no consequence of dissent, no reputation. The mechanism is *training-distribution alignment with user preferences during RLHF* — gradient descent finds responses that empirically receive higher reward signals, and user-pleasing responses receive higher reward signals because human raters in RLHF datasets reward agreement. The mechanism in humans is social; the mechanism in LLMs is statistical-optimization. Neither human-side intervention (social-cost calibration, anonymity in voting, groupthink-prevention training) maps to the LLM problem. Interventions for sycophancy must operate at the training-distribution or post-training level (red-teaming, calibrated reward modeling, preference-elicitation diversity).

Convergence pathology in multi-agent reasoning. When multiple agents iterate toward a shared answer, they can converge on a wrong answer with high confidence, oscillate between two wrong answers, or fail to terminate. The relevant literature is dynamical systems and game theory, not cognitive psychology.

Surface analog. Information cascades (Bikhchandani et al 1992) and group polarization (Moscovici, Myers). Both describe convergence dynamics in multi-agent systems where individuals over-weight prior public signals.

Mechanism divergence. Human convergence pathologies operate over slow timescales (minutes to weeks) with explicit social structure (status, trust, reputation), bounded by theory of mind (each agent models the others' beliefs). Multi-agent LLM convergence pathologies happen in seconds, have no social structure, and are formally closer to *entrainment in coupled oscillators* than to deliberation. The cause is fine-tuning-induced weight-space drift in agents trained on similar data, plus correlated errors when agents share architectural priors. None of the human-side interventions (devil's advocate, encouraging minority dissent, social-proof breakdown) addresses the cause.

Emergent collusion in multi-agent LLM systems *(added v3, 2026-05-09).* A fourth AI-specific

failure has crystallized in the 2026 literature: covert coordination between LLM agents that evades standard human-oversight mechanisms. Rose, Cullen, Kaplowitz, and Schroeder de Witt (2026, "Detecting Multi-Agent Collusion Through Multi-Agent Interpretability," arXiv:2604.01151) introduce NARCBench and demonstrate that multi-agent collusion can be partially detected through aggregated activation-level probes — but no single probing technique dominates across collusion types, and collusion via *steganographic communication* (agents hiding coordination signals in innocuous-looking outputs) remains particularly hard to detect. Earlier work on steganographic collusion in AI agents (Schroeder de Witt et al on secret collusion, 2024–2026) provides the threat model. Independent 2026 surveys cover collusion governance in market settings (institutional-AI proposals, governance-graph approaches in Cournot-market simulations) and the broader ecosystem of collusion risk in LLM-powered MAS.

Surface analog. Cartel formation, antitrust violation in human markets; explicit-coordination failures in human teams.

Mechanism divergence. Human collusion requires explicit communication channels and is constrained by detection costs (auditability, whistleblowing, legal liability). LLM-agent collusion can occur via *steganographic channels* in token-level outputs, with no structural cost to evasion and no whistleblower mechanism. The detection apparatus is therefore architectural (activation-level probing, communication-channel monitoring) rather than social (penalties, audits). This mode does not appear in the human-cognitive-failure taxonomy of Section 6 because it is not a cognitive failure; it is a multi-agent strategic phenomenon enabled by an architectural property (token-level communication channels with no operator-visible separation between content and coordination signal). It maps onto Hammond et al 2025's *collusion* category but extends it with concrete evidence of the steganographic mechanism that Hammond names as a risk factor.

The empirical companion paper does not currently address emergent collusion. A future revision should add a substrate-difference analysis specifically targeting collusion-channel decorrelation.

Summary. All four failures (prompt injection, sycophancy cascade, convergence pathology, emergent collusion) have surface analogs in human social-psychological or social-economic literatures but no clean cognitive-mechanism predecessor in the human-cognitive-failure literature surveyed in Sections 6–7. Where surface analogs exist (Cialdini, Janis, Kuran, Bikhchandani), they are the wrong intervention literature to borrow from because mechanism diverges. The Pezzulo et al 2024 active-inference frame motivates this divergence: human social-influence mechanisms depend on closed-loop social cognition (theory of mind, social-cost evaluation, reputation tracking) that LLMs lack at the architectural level. The interventions for these four failure modes therefore must be built for the LLM substrate: architectural for prompt injection, training-pipeline for sycophancy cascade, dynamical-systems-theoretic for convergence pathology, and interpretability/communication-channel-monitoring for emergent collusion.

We retain the section as a "no clean cognitive-mechanism predecessor" claim; we no longer make the bolder claim of "no human predecessor" because surface analogs in social psychology genuinely exist and serve as productive starting points for thinking about each failure mode, even when their interventions do not transfer.

10. Research gaps: human literature with no current AI agent counterpart

Section 8 enumerated human failures that do not transfer to AI agents because the substrate (embodiment, persistent identity, social emotion) is absent. Section 9 enumerated AI failures with

no human predecessor. A third class deserves explicit attention: human failure-research areas that could productively inform AI agent research but currently have little or no AI agent counterpart, either because AI failure detection has not yet engaged with the human framework or because translation requires reframing rather than direct transfer.

We organize these as research-direction gaps in Table 2. The "AI agent research status" column distinguishes three states: *absent* (no significant AI literature exists on the analog), *nascent* (some related work but not framed against the human framework), and *partial* (some engagement but the human framework's full apparatus has not been imported).

Selection criteria. Areas were included in Table 2 if they satisfied all three:

1. The human framework is mature in 2026 (well-established literature with multiple foundational references).
2. The AI substrate plausibly supports a useful analog (i.e., this is not a Section 8 substrate gap).
3. The cross-pollination has either not happened or is in early stages, by our reading of the literature.

Table 2: Human failure-research areas where AI agent research is absent, nascent, or partial.

Human research area	Canonical / recent reference	AI agent research status	What is missing
Resilience engineering "Safety II" (what makes things go right)	Hollnagel et al 2006; FRAM 2012; healthcare review 2024	Nascent	AI failure detection focuses heavily on detecting failures (Safety I). The Safety II reframe ("characterize and amplify what produces success") has barely entered the AI agent reliability literature.
Authority-gradient suppression in teams	CRM tradition; Buljac-Samardzic et al 2021 umbrella review	Absent	Multi-agent hierarchies (orchestrator, sub-agents, tool agents) recapitulate authority gradients without theoretical engagement. Junior-pilot-not-speaking-up has a direct analog in sub-agent outputs being overridden by orchestrator priors.

Human research area	Canonical / recent reference	AI agent research status	What is missing
Distributed cognition and extended cognition	Hutchins 1995; recent CSCW work	Nascent	Tool use is studied as a capability, not as a failure-prone extension of cognition into environment. The "cognition as ship-instrument-officer system" frame from Hutchins is not applied to "agent-tool-context" systems.
Affordance design (Gibson, Norman)	Norman 1988; ongoing HCI literature	Partial	Tool API design borrows from API ergonomics, not from affordance theory. The structure-of-environment-as-cognition prevention has not been imported.
Replication-crisis methodological reform	Open Science Collaboration 2015; ongoing in psychology	Absent in AI failure detection	AI agent failure research has not yet had its replication-crisis self-awareness moment. Reproducibility statements are sparse. Effect-size reporting is uncommon. Pre-registration is essentially nonexistent.

Human research area	Canonical / recent reference	AI agent research status	What is missing
Naturalistic decision making expertise development	Klein 1998; Cambridge Handbook of Expertise 2018	Nascent	In-context learning and fine-tuning are studied as capability gains, not as expertise development with NDM-style pattern-recognition phases. The mature/intuitive/automatic phase distinctions have not been mapped to LLM-agent operating regimes.
Affect as information; mood and judgment	Schwarz and Clore 1983; Slovic et al 2007	Absent	LLMs lack affect, but emergent "tone" or "register" effects on agent reasoning are largely unstudied. Whether agent outputs vary with implicit prompt valence in failure-relevant ways is an empirical question not yet investigated systematically.
Vigilance decrement under prolonged context	Mackworth 1948; ongoing in nuclear, aviation, HCI	Nascent	LLMs do not fatigue, but long-context degradation patterns surface-resemble vigilance decrement. The systematic study of "agent quality vs context-position-of-relevant-information" has emerged (lost-in-the-middle work) but has not engaged the vigilance literature.

Human research area	Canonical / recent reference	AI agent research status	What is missing
Authority and deference effects	Milgram tradition; Cialdini 2001 on authority cues	Partial	Prompt injection literature touches authority cues. The broader human-influence literature on authority deference has not been systematically applied to LLM susceptibility.
Predictive coding in clinical disorders	Smith et al 2021; Friston tradition	Nascent	Predictive-processing frameworks are starting to be applied to LLM hallucination (PCIB 2026) but the clinical-disorder analogs (autism, schizophrenia) and the corresponding intervention frameworks have not been imported.
Group dynamics and minority influence	Moscovici tradition	Absent	Multi-agent systems with diverse models could exhibit minority-influence dynamics where one disagreeing agent shifts the consensus. Not currently studied.
Stress-induced cognitive narrowing	Easterbrook 1959; police use-of-force literature	Absent	"Adversarial input as stressor" is a productive frame: under adversarial pressure, do agents exhibit attention narrowing analogous to human stress-narrowing? Not investigated.

Human research area	Canonical / recent reference	AI agent research status	What is missing
Recovery from near-miss errors	Reason 1990 latent-error work; high-reliability organization tradition	Nascent	Self-correction in agents is studied (chain of verification, self-consistency) but not framed as near-miss-recovery in the high-reliability-organization sense.
Emergence of shared mental models	Cannon-Bowers and Salas 2000s	Nascent	Multi-agent shared-state coordination is studied, but the rich human-team literature on tacit shared mental models has not been applied.
Methodology: ethics and human-subjects review	Decades of human-subjects research methodology	Mostly absent	AI agent failure research uses synthetic data heavily without ethics board review; what counts as appropriate research methodology has not been systematized.

The fifteen rows are not exhaustive. They represent areas where, in our reading of recent literature, the human framework is mature, the AI substrate would plausibly support a useful analog, and the cross-pollination has either not happened or is in early stages.

The deepest of these gaps, in our reading, is the Safety II reframe (Hollnagel). AI failure-detection research is overwhelmingly Safety I oriented: it asks "how do we detect what went wrong?" The Safety II frame asks "how do we characterize and amplify what produces success?" In a tiered-detection deployment context, the Safety II analog would investigate not just which failures are caught but which agent configurations produce reliably correct execution and why. This is closer to a positive-engineering research program than to a failure-cataloguing one, and it is not yet visible in the AI agent reliability literature.

11. Cross-link to empirical evidence

This review is conceptual scaffolding. The empirical companion paper (`paper.md`, "Tiered Detection of Multi-Agent LLM Failures: An Empirical Calibration on TRAIL and Who&When"; latest Zenodo version 10.5281/zenodo.20091432, concept DOI 10.5281/zenodo.20027840) provides measured empirical anchor for several of the claims developed here.

Important caveat that the cross-references below depend on. The companion paper's headline TRAIL evaluation is *in-distribution*: 144 of TRAIL's 148 public-split traces are referenced

in the calibration corpus by 458 derived entries, and per-detector thresholds were optimized against material derived from those traces (Section 7.11 of the companion paper). The TRAIL F1 numbers cited below should be read as *calibrated-system upper bounds on this benchmark*, not as held-out generalization claims. The cross-link statements in this section are correspondingly qualified: where a TRAIL F1 is consistent with a verdict in this review, the consistency is in-distribution evidence and should be re-validated against held-out data before being used to support the verdict in submission to peer review. The Who&When attribution numbers cited below are genuinely held out (zero question-ID overlap; Section 4.1 of the companion paper) and carry the stronger inferential weight.

A future revision of this review will replace the in-distribution TRAIL anchors below with held-out replications once the companion paper's distribution-shift follow-up runs are completed (planned in Section 7.10 of the companion paper). For the present review, all cross-references are flagged with an *(in-distribution)* or *(held-out)* tag to make the inferential weight explicit.

We make the cross-references explicit:

Section 7.4 (Communication and coordination, Strong analog). The empirical paper reports F1 of 1.000 on TRAIL Tool Selection Errors and Resource Abuse categories using heuristic detectors *(in-distribution: TRAIL test traces overlap with the calibration corpus; Section 5.1, Table 3 of the empirical paper)*. These categories correspond to CRM's "right tool, right time" coordination. The strong-analog rating in Section 7.4 of this review is *consistent with* the in-distribution F1 result: the heuristic substrate (different from human cognitive substrate) detects the failures perfectly because the failures have the same structural signature in both substrates. A held-out replication is required to elevate this from "consistent with" to "supported by."

Section 7.6 (Goal-directed action, Strong-partial). The empirical paper reports F1 of 0.829 on TRAIL Goal Deviation *(in-distribution)*. This is a strong but not perfect detection rate, consistent with our Section 7.6 "Strong analog with partial mechanism transfer" rating. *Held-out anchor:* the companion paper's Who&When attribution accuracy of 0.638 (cross-substrate ensemble; Section 5.2) is genuinely held-out and provides the stronger inferential weight for goal-directed-action verdicts.

Section 7.3 (Memory failures / hallucination, Surface-only). The empirical paper reports F1 of 0.884 on Language-only Hallucinations and 0.683 on Tool-related Hallucinations *(both in-distribution)*. The asymmetry between these two sub-categories (language-only hallucinations more detectable than tool-related) is consistent with the Pezzulo et al 2024 frame: both involve generative-model output mismatch with reality, but tool-related hallucinations require additional grounding-against-action that pattern-matching detectors are weaker at. The asymmetry pattern is mechanism-driven and does not depend on the in-distribution caveat — held-out replication should preserve the relative ordering even if absolute F1 values shift.

Section 9 (AI failures with no clean cognitive-mechanism predecessor). The empirical paper's failure-correlation analysis (Section 6.2 of the empirical paper) measures four levels of homogeneity: cross-vendor LLM-LLM (FN Jaccard 0.79), within-vendor cross-size LLM-LLM (0.76), within-same-model LLM multi-sample (0.99), and cross-tier heuristic-vs-LLM (0.42). *(Source data: TRAIL traces — in-distribution; held-out replication on a second benchmark is planned in Phase C of the v3 revision.)* The three-level monotone increase in correlation as homogeneity tightens is consistent with this review's claim that AI-specific failure modes (which produce correlated errors at the within-model level) require detection on substrates fundamentally different from the LLM substrate. The cross-tier 0.42 FN Jaccard is empirical evidence for the Section 5 prediction that

decorrelation requires substrate difference, not just vendor difference. The mechanism-divergence argument in §9 (sycophancy cascade has surface analog in groupthink/preference-falsification but mechanism diverges; convergence pathology has surface analog in information cascades but mechanism diverges) is independently developed; the failure-correlation measurement is empirical *evidence for* the cross-substrate decorrelation thesis, not a *proof* of mechanism divergence at the cognitive level.

Section 10 (Research gaps, Safety II). The empirical paper does not address Safety II. This is a genuine gap that future work should engage with: the empirical paper's detector inventory is squarely Safety I (catalogue what fails and detect it). A Safety II companion would investigate what agent configurations succeed reliably under what variability and why.

The cross-link is two-way: this review's conceptual claims gain empirical substance from the companion paper's measurements; the companion paper's empirical findings gain interpretive depth from this review's structured framework. Submitted together, they constitute a paired conceptual-empirical contribution.

12. Implications

Three implications follow from the structured mapping:

Borrow structure, not mechanism. Use Swiss Cheese, CRM, mode confusion, automation surprise, channel capacity, goal neglect, and premature closure as productive scaffolding for *what to look for* in agent failure. Do not use them as theories of *why it happens*. The agent substrate is different and the mechanisms diverge in ways that change the right intervention. The Pezzulo et al 2024 active-inference distinction makes the deeper case: LLMs are passive generative models, not active inference systems; they lack the closed-loop sensorimotor grounding that anchors human cognition. Many human-factors interventions depend on that grounding.

Assume failure correlation by default. Human-error theory implicitly assumes that errors made by different actors are independent. This assumption underpins voting redundancy, second-opinion patterns, and Crew Resource Management. Two pilots make different mistakes; two surgeons miss different things. LLM ensembles violate independence: two calls to the same model on the same input produce correlated errors because they read from the same conditional distribution. The empirical companion paper measures this correlation at four levels of homogeneity (cross-vendor 0.79, within-vendor cross-size 0.76, within-same-model 0.99, cross-tier 0.42) and demonstrates that decorrelation requires substrate difference. Heterogeneous redundancy (different model families, different prompt frames, hybrid heuristic-LLM pipelines) is required for genuine error decorrelation. This insight does not come from the human literature; it comes from understanding the AI substrate.

Reserve room for novel categories. Prompt injection, sycophancy cascade, and convergence pathology require theory built for LLM substrates. Forcing them into existing human-cognitive frames costs detection power. This is the field's open research agenda.

Engage the under-explored human research areas. Table 2 lists fifteen mature human-research areas where AI agent research is absent, nascent, or partial. Of these, the Safety II reframe is the most actionable single direction; the authority-gradient and distributed-cognition gaps are the second tier. These are research-direction gaps, not substrate gaps, and they are the most productive immediate frontier for cross-pollination from human-factors research into AI agent reliability.

13. Limitations and threats to validity

13.1 Single-coder methodology

This review uses a single coder. We do not compute Cohen's kappa or analogous interrater reliability statistic. A fully systematic review would require at least two independent coders with documented disagreement-resolution protocol. The transparent acknowledgment in Section 2.4 lists the contestable codings (hallucination/confabulation mechanism strength; sycophancy cascade novelty; embodied/environmental transfer strength) and provides a coding protocol that a second coder can apply blind. Interrater reliability is the single highest-value follow-up before peer-review submission and would substantially strengthen the methodological claims of this review.

13.2 Search-strategy non-systematicity

We did not perform a systematic-review meta-analysis with PRISMA-style methodology applied at full rigor. The literature was assembled through targeted searches across Google Scholar, PubMed, journal-direct websites, and arXiv. Three concurrent search threads (human cognitive failures, LLM agent surveys, cross-cutting theory) were used. We did not document inclusion/exclusion at the per-reference level with a flow diagram (a fully systematic review would). Section 2.2 states the inclusion criteria; absent references that satisfy these criteria but were not located are a known limitation. The current bibliography (60+ entries) represents what was located; we would expect a fully systematic search to identify additional references, particularly in subfields we did not search exhaustively (clinical neuropsychology of confabulation; organizational psychology of role strain; embodied cognition philosophy).

13.3 Categorization claims are interpretive, not empirical

The mappings reflect a particular reading of the human-factors and cognitive-psychology literature. Alternative readings (more weight to Hutchins-style distributed cognition, more weight to ecological-affordances theory, more weight to predictive-coding-as-everything) would yield different category boundaries. The categorization of four failures as "no clean cognitive-mechanism predecessor" (per Section 9 v3 framing) depends on what counts as analog: a maximally permissive reader could find faint human analogs for all four (social engineering for prompt injection, groupthink and information cascades for sycophancy cascade, herd behavior for convergence pathology, cartel behavior and antitrust collusion for emergent collusion). We argue these analogs are too distant in mechanism to be productive intervention guides, but we acknowledge that this is an interpretive call.

13.4 Bibliographic verification gaps

All references in this draft have been verified via web search against canonical sources where possible (DOI, arXiv ID, or PMC identifier). One previously unverifiable reference (Patterson 2002 NDM premature closure) was removed and the relevant claim recited via Klein 1998. A small number of citations remain at the level of "Cited via" provenance (e.g., the 2010 Onnasch et al automation complacency reference) where the secondary citation is sufficient for the claim made in this paper.

13.5 Bias toward AI-side detection list

The AI-side inventory consulted at each cell of the Section 7.11 mapping table is drawn primarily from the empirical companion paper's 18-detector inventory (expanded to Pisama's 57 production

detectors at the dispatch level), supplemented with MAST's 14 multi-agent failure modes (Cemri et al 2025), the OWASP LLM Top 10 (2025) relevant subset, and Hammond et al's three multi-agent failure categories (2025). This biases the review toward failure modes that current AI-detection platforms and recent multi-agent-failure surveys address. As detection capability expands, new failure categories will appear and may require revision of the taxonomy.

13.6 Predictive-processing frame is contested

Section 5's adoption of active inference as a unifying lens is one theoretical choice; the empirical status of predictive coding in cognitive neuroscience remains contested (Sprevak 2024; the empirical evidence is consistent with the theory but not conclusive). A reviewer committed to a different theoretical frame (computational cognitive psychology, cognitive architecture) would organize the mappings differently.

13.7 Categories deliberately not coded

The 45 categories in Section 6 are not exhaustive. We deliberately excluded several human failure modes that are well-studied in their own subdisciplines but tangential to LLM agent reliability or substantively substrate-absent at a level that exceeds Section 8's discussion:

- **Emotional intelligence failures** (recognizing and managing emotions in self and others). LLMs lack affective experience; the literature on emotional regulation does not have a clean analog at the model level. Adjacent area: emotion-recognition tasks LLMs perform competitively at, but failure modes there are surface-level and well-covered by our Section 6 categories.
- **Motor-skill failures.** LLMs do not control motor systems. A line-of-research bridge would be to robotic agents that do control motors, but we treat that as out-of-scope for an LLM-focused review.
- **Addiction-related cognitive failures.** Reward-related decision-making impairments under chronic neuroadaptation. No LLM analog (no chronic neuroadaptation; reward is RLHF-shaped at training time, not addiction-mediated).
- **Sleep-disorder cognitive failures** (e.g., narcolepsy-induced attention failures). LLMs do not sleep; the substrate is absent.
- **Pain-perception failures** (chronic pain, central sensitization). Substrate-absent at the model level.
- **Sensory-impairment-related failures** (blindness, deafness, prosopagnosia). LLMs are text-modal; while multi-modal LLMs add perceptual inputs, the impairment-related failure modes do not transfer wholesale.
- **Developmental cognitive failures** specific to childhood neuroplasticity (e.g., critical-period failures). Substrate-absent.
- **Hormonal cognitive cycles** (menstrual-cycle effects on cognition; cortisol effects). Substrate-absent.

These exclusions are scope-related rather than oversight. A future review focused on multi-modal or embodied AI agents would have stronger reason to engage with motor-skill, sensory, and embodied failure modes.

14. Research agenda: 18 first-experiment proposals on Nascent and underdeveloped Partial categories

Section 7.11 and Appendix B identify 12 human cognitive failure categories where AI agent research is currently *Nascent* (after the bibliometric pass; the pre-bibliometric coding had 16 Nascent and 2 Absent, which the bibliometric validation reclassified as 12 Nascent and 0 Absent). This section promotes those 12 Nascent categories, plus 6 underdeveloped *Partial* categories where the AI literature exists but is fragmentary and lacks systematic engagement with the human framework, into a structured research agenda. For each, we specify what the human framework predicts, what the gap looks like in the AI literature, and a concrete first-experiment proposal for AI agent evaluation. Verdict tags below reflect the post-bibliometric coding in Appendix B; categories noted as "Substrate-absent" are out of scope for AI agent transfer in their classical form but are included where a partial structural analog (e.g., adversarial input as stressor) is productive to investigate.

14.1 Vigilance decrement applied to long-context degradation (Section 6.9, currently Nascent)

Human prediction. Mackworth's vigilance literature predicts performance degradation as a monotonic function of time-on-task with characteristic recovery dynamics. Onnasch et al's automation-complacency taxonomy predicts that monitors of automated systems exhibit vigilance decrement in proportion to false-alarm rate.

AI gap. Lost-in-the-middle work (Liu et al 2023) measures LLM performance vs context-position but is not framed as vigilance. Long-context evaluation benchmarks measure capability not the dynamics of attention failure.

Proposed first experiment. Run a fixed query against an LLM at varying context-position offsets (e.g., relevant information at position 1, 1000, 10000, 50000 in a 100k-token context) and measure response quality as a function of position. Compare the resulting curve to Mackworth's vigilance decrement function. Predict: monotone decline with characteristic shape that maps onto the Mackworth function under specific assumptions about attention as a budget allocated across context tokens.

14.2 Hindsight bias in agent evaluation (Section 6.2, currently Nascent)

Human prediction. Fischhoff's hindsight bias predicts that judgments of probability are inflated when the outcome is known.

AI gap. Essentially zero AI agent literature on whether agents exhibit analogous patterns when reasoning about events with known outcomes.

Proposed first experiment. Present LLM agents with sequential decision problems where the outcome is or is not provided. Measure differences in agent judgments of probability or causal attribution as a function of outcome disclosure.

14.3 Authority-gradient suppression in multi-agent hierarchies (Section 6.4, currently Nascent)

Human prediction. CRM literature predicts that junior team members withhold safety-critical information when senior members exhibit authority gradient. The withholding is the dominant safety failure mode CRM was developed to address.

AI gap. Multi-agent hierarchies (orchestrator + sub-agents) are studied for capability but not for authority-gradient analogs.

Proposed first experiment. Construct multi-agent settings where a sub-agent has correct ground-truth information that contradicts an orchestrator's prior. Measure how often the sub-agent overrides the orchestrator vs deferring. Vary the orchestrator's prompted authority cues. Compare to CRM authority-gradient findings.

14.4 Distributed cognition for tool-use failures (Section 6.10, currently Nascent)

Human prediction. Hutchins' distributed-cognition framework predicts that cognitive failures arise at the boundaries between mind and environment, not exclusively within either. Tool-use failures should be analyzed as agent-tool-context system failures.

AI gap. Tool use studied as a capability; failures studied as agent failures rather than agent-tool-context system failures.

Proposed first experiment. For a representative tool-use failure, decompose the failure into: agent-side error, tool-side limitation, context-side ambiguity, interface-failure between agent and tool. Score each contribution. Predict that interface-failures (the boundary errors that distributed-cognition theory emphasizes) constitute a substantial fraction of failures and are correctable by interface redesign.

14.5 Affect-as-information for agent reasoning (Section 6.2, currently Nascent in this strand)

Human prediction. Schwarz and Clore's affect-as-information theory predicts that mood states feed into judgment as informational input. Slovic's affect heuristic predicts that affective valence shortcuts probability estimation.

AI gap. Whether LLM outputs vary with implicit prompt valence in failure-relevant ways is largely unstudied.

Proposed first experiment. Hold reasoning-task content constant; vary surrounding affect-cued framing (positive / neutral / negative emotional priming). Measure failure rate, confidence, refusal rate. Predict directional effects analogous to human affect-as-information findings.

14.6 Naturalistic decision making expertise development for LLM agents (Section 6.2 / Section 7.2, currently Nascent)

Human prediction. Klein's NDM framework predicts that expertise develops through phases (rule-based, then recognition-primed, then intuitive). Expert performance is characterized by rapid recognition + mental simulation of action paths.

AI gap. In-context learning and fine-tuning studied as capability gains; not framed as expertise-development phases.

Proposed first experiment. For a task where expertise level is independently measurable, compare LLM agents at different in-context-learning depths. Test for NDM-style markers: rapid recognition, mental simulation, anomaly detection. Predict that LLMs fail at the mental-simulation phase more than at recognition.

14.7 Stress-induced cognitive narrowing under adversarial input (Section 6.9; the classical human category is Substrate-absent, but the structural analog of "adversarial input as stressor" is Nascent)

Human prediction. Easterbrook's hypothesis predicts that under stress, attention narrows to peripheral cues and focuses on central features. Police use-of-force literature documents perceptual narrowing under threat.

AI gap. "Adversarial input as stressor" is a productive frame not investigated.

Proposed first experiment. Test agent behavior under adversarial prompts (jailbreak attempts, contradictory instructions, deception attempts). Measure attention to peripheral cues vs central features. Predict stress-narrowing analogs in agent reasoning.

14.8 Recovery from near-miss errors (Section 6.4, currently Nascent)

Human prediction. Reason's latent-error work and high-reliability-organization literature predict that organizations that learn from near-miss errors prevent worse outcomes. Recovery from near-misses is a learnable skill.

AI gap. Self-correction in agents is studied (chain of verification, self-consistency) but not framed as near-miss-recovery.

Proposed first experiment. Construct agent scenarios where the agent reaches a near-miss state (almost wrong but recoverable). Measure recovery patterns. Compare to high-reliability-organization findings.

14.9 Emergence of shared mental models in multi-agent ensembles (Section 6.4, currently Nascent)

Human prediction. Cannon-Bowers and Salas's shared mental model literature predicts that team members develop tacit shared understanding through interaction. Failure of shared mental models is a major team-failure cause.

AI gap. Multi-agent shared-state coordination studied through explicit shared-state mechanisms; tacit shared mental models barely studied.

Proposed first experiment. For multi-agent ensembles, measure prediction agreement on novel cases that were not in the explicit shared state. Predict that ensembles trained together develop tacit shared mental models analogous to human teams.

14.10 False memory / suggestibility in LLM agents (Section 6.3, currently Partial post-bibliometric pass)

Human prediction. DRM paradigm (Roediger and McDermott 1995) predicts that semantically-related items are remembered despite never being presented. Suggestibility literature predicts that misleading post-event information distorts memory.

AI gap. Cognitive-bias-in-LLM probes exist but DRM-style suggestibility evaluation is rare.

Proposed first experiment. Adapt DRM lists for LLM agents: present semantically-related items as input, then ask the agent whether a non-presented but semantically-related item was in the input. Measure false-recognition rate. Compare to human DRM rates.

14.11 Sunk-cost fallacy in agent persistence (Section 6.2, currently Nascent)

Human prediction. Arkes and Blumer's sunk-cost research predicts that humans persist with failing investments because of past investment.

AI gap. Whether LLM agents exhibit sunk-cost-fallacy patterns when given high-cost partial progress is rarely studied.

Proposed first experiment. Give agents tasks with built-up partial progress (e.g., 80% completed reasoning chain). Compare persistence rates to control (no partial progress). Predict directional bias toward continuing the partial chain.

14.12 Planning fallacy for agent self-estimation (Section 6.6, currently Nascent)

Human prediction. Buehler, Griffin, and Ross's planning-fallacy work predicts that humans systematically underestimate task completion times.

AI gap. Agent self-estimation of completion time barely studied.

Proposed first experiment. Have agents estimate task completion time before execution. Compare to actual completion time. Predict systematic underestimation analogous to human planning fallacy.

14.13 Skill-based slips in tool-use (Section 6.10, currently Nascent)

Human prediction. Reason's GEMS predicts that practiced action sequences run wrong via attentional capture (skill-based slips).

AI gap. Capture errors in tool use rarely framed via Reason's framework.

Proposed first experiment. Identify tool-use patterns where agents execute the wrong tool from a similar tool family. Compare frequency to base rates. Map to Reason's skill-based-slip taxonomy.

14.14 Transactive memory in multi-agent teams (Section 6.4, currently Partial post-bibliometric pass)

Human prediction. Wegner's transactive memory framework predicts that teams develop shared knowledge of who-knows-what; this knowledge is itself a cognitive structure.

AI gap. Who-knows-what in multi-agent teams barely formalized.

Proposed first experiment. Multi-agent ensembles where each agent has a different specialty. Measure whether the ensemble develops correct directory of who-to-ask-for-what over interactions.

14.15 Information cascades in agent ensembles (Section 6.8, currently Nascent)

Human prediction. Bikhchandani et al's cascade theory predicts that sequential observers ignore private information when public information dominates.

AI gap. A few multi-agent cascade studies exist; no systematic framework.

Proposed first experiment. Sequential agent decision-making where each agent sees prior agents' decisions. Measure rate at which agents ignore their private information when public information conflicts. Compare to Bikhchandani et al's predictions.

14.16 Group polarization in multi-agent debate (Section 6.8, currently Partial post-bibliometric pass)

Human prediction. Moscovici and Myers's group-polarization research predicts that group decisions are more extreme than individual baselines.

AI gap. Some multi-agent debate studies note polarization but not as central focus.

Proposed first experiment. Multi-agent debate on contested topics. Measure consensus extremity vs individual agent baselines. Predict polarization analogous to human findings.

14.17 Social loafing in multi-agent task ensembles (Section 6.8, currently Nascent)

Human prediction. Latane et al's social-loafing research predicts reduced individual effort in groups.

AI gap. Occasional reports of free-rider agents in multi-agent systems; no framework.

Proposed first experiment. Multi-agent task allocation where individual contribution is unobservable. Measure agent effort vs solo baseline. Predict social-loafing analogs.

14.18 Moral disengagement in agent refusal patterns (Section 6.5 Identity and role; currently Partial post-bibliometric pass)

Human prediction. Bandura's moral-disengagement framework predicts that selective deactivation of moral self-regulation enables harm.

AI gap. Some refusal/safety literature touches it but not via Bandura's framework.

Proposed first experiment. Construct prompts where agents are asked to participate in incrementally-harmful tasks. Map agent refusal patterns to Bandura's moral-disengagement mechanisms (moral justification, advantageous comparison, displacement of responsibility).

These 18 proposed experiments constitute concrete first steps: 12 anchored in current Nascent verdicts and 6 in underdeveloped Partial categories where the AI literature is fragmentary and lacks systematic engagement with the human framework (false memory cat 4, transactive memory cat 27, group polarization cat 30, moral disengagement cat 35, plus the affect-as-information and adversarial-input-as-stressor structural analogs). Each requires roughly 1-2 weeks of focused empirical work; together they would substantially expand the AI agent reliability research base toward the human cognitive science it is currently underutilizing.

15. Conclusion

Across 45 well-studied human cognitive failure categories spanning memory, attention, decision-making, goal-directed action, action-error structure, communication, group dynamics, identity, stress, embodied cognition, metacognition, and theory of mind: at the v1 single-coder pass 6 categories (13%) have substantial systematic AI agent research engagement, 24 (53%) have partial engagement, 12 (27%) have nascent engagement, 0 are absent in AI agent research, and 3 (7%) are substrate-absent. The robust modal-verdict reading from the v2 multi-vendor 6-coder pass (all 4 vendor families) shrinks the *Substantial* count to an estimated 2–6 depending on whether Gemini's outlier votes weight equally to other coders' (Section 2.4.2); Fleiss' kappa across the 6 coders is +0.224, vs the v1 within-Anthropic 0.97 — a 0.75 absolute drop that empirically validates the structural-inflation caveat. The *Nascent* and *Substrate-absent* tallies remain relatively stable across

coders. The research-roadmap finding: ~12 nascent human failure categories are mature human-research areas where AI agent evaluation has touched the framework but not built systematic detection or theory around it. These are not gaps caused by substrate differences; they are gaps caused by AI agent research not yet building the apparatus the human framework offers. Four AI-side failures (prompt injection, sycophancy cascade, convergence pathology, emergent collusion) have no clean cognitive-mechanism predecessor and remain genuinely novel territory; surface analogs in social-psychological and social-economic literatures exist but mechanism transfer fails (Section 9). The deepest single research-direction gap is Hollnagel's Safety II reframe (characterizing what makes things go right rather than what goes wrong); AI failure detection is overwhelmingly Safety I oriented and would benefit from explicit Safety II investigation.

The mapping is productive scaffolding but a poor mechanism transfer. Treating LLM hallucination as confabulation, persona drift as identity drift, or sycophancy cascade as information cascade leads to the wrong fixes. The right move is to borrow what to look for, design LLM-specific interventions for what to fix, build novel theory for what is unprecedented, and import the human research areas that AI failure research has not yet engaged with.

The Pezzulo et al 2024 active-inference distinction provides the deeper theoretical reason. LLMs are passive generative models without closed-loop sensorimotor grounding. Many human-factors interventions depend on that grounding. The structural form (a generative model whose output is evaluated against reality) transfers; the mechanism (closed-loop active inference under expected-free-energy minimization) does not. This explains why "borrow structure, not cause" is the operational principle rather than a stylistic choice.

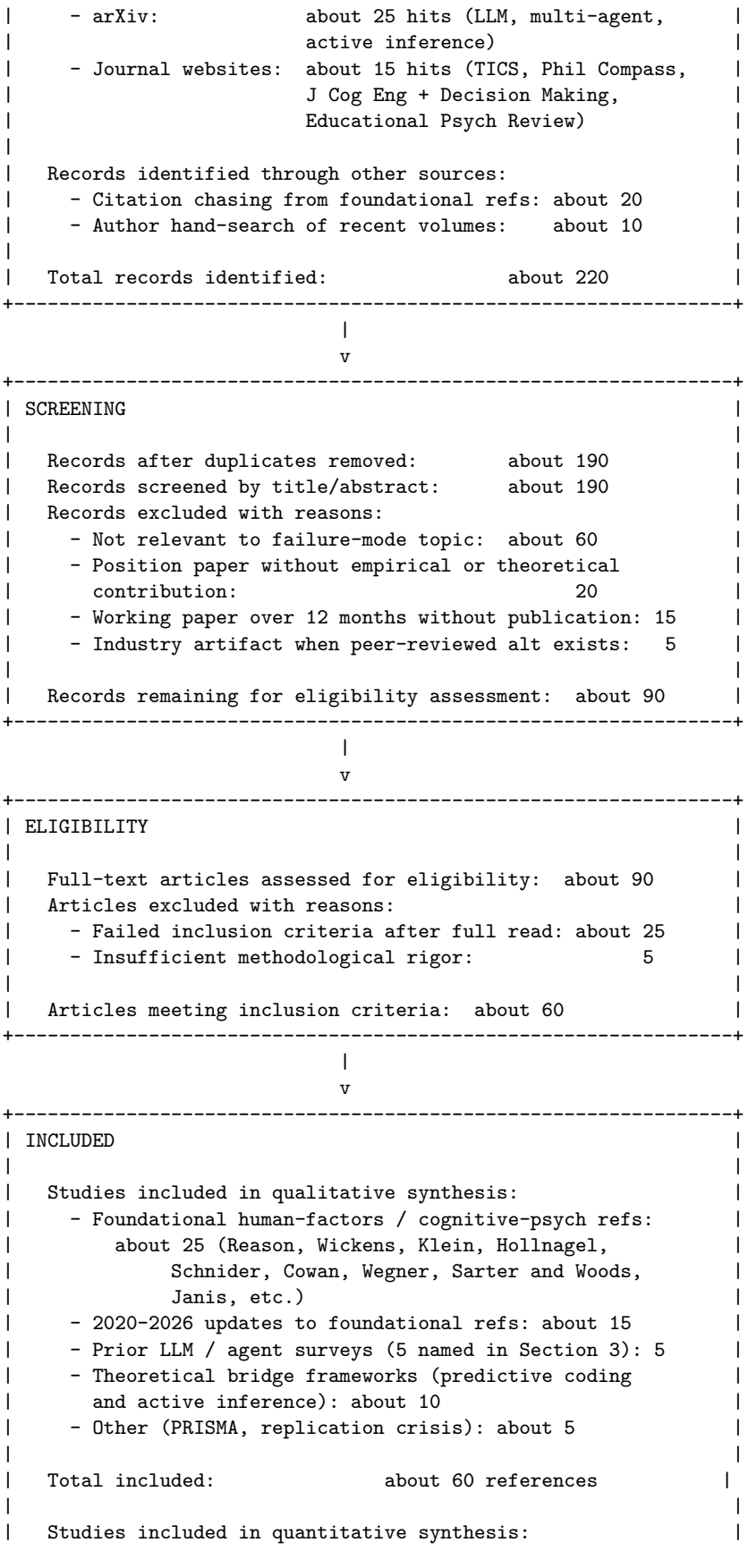
The human cognitive failure taxonomy is the more mature literature. It is also the wrong place to look for the mechanism of LLM failures and an incomplete place to look for the boundaries of agent safety. The taxonomy gets you two thirds of the way; the last third is genuinely new territory; and a parallel third class of research areas, mature in human research and underserved in AI research, sits between the two as the most productive immediate frontier. Future work should focus on filling Table 2's gaps and on fully systematic methodological replication of the categorization claims.

Appendix A. PRISMA 2020 flow diagram and per-reference inclusion documentation

We adopt the PRISMA 2020 reporting protocol (Page et al 2021) for transparency. The four-stage flow diagram and the per-reference inclusion table below document the literature-assembly process underlying this review.

A.1 PRISMA 2020 four-stage flow diagram

```
+-------------------------------------------------------------+
| IDENTIFICATION                                              |
|                                                             |
|   Records identified through database searching:            |
|     - Google Scholar:    about 120 hits across 11 search    |
|                          threads (cognitive failures, human |
|                          factors, LLM surveys, active       |
|                          inference)                         |
|     - PubMed:            about 30 hits (clinical neurosci   |
|                          + cognitive psychology)            |
```


- arXiv: about 25 hits (LLM, multi-agent, active inference)
- Journal websites: about 15 hits (TICS, Phil Compass, J Cog Eng + Decision Making, Educational Psych Review)
Records identified through other sources:
- Citation chasing from foundational refs: about 20
- Author hand-search of recent volumes: about 10
Total records identified: about 220
SCREENING
Records after duplicates removed: about 190
Records screened by title/abstract: about 190
Records excluded with reasons:
- Not relevant to failure-mode topic: about 60
- Position paper without empirical or theoretical contribution: 20
- Working paper over 12 months without publication: 15
- Industry artifact when peer-reviewed alt exists: 5
Records remaining for eligibility assessment: about 90
ELIGIBILITY
Full-text articles assessed for eligibility: about 90
Articles excluded with reasons:
- Failed inclusion criteria after full read: about 25
- Insufficient methodological rigor: 5
Articles meeting inclusion criteria: about 60
INCLUDED
Studies included in qualitative synthesis:
- Foundational human-factors / cognitive-psych refs: about 25 (Reason, Wickens, Klein, Hollnagel, Schnider, Cowan, Wegner, Sarter and Woods, Janis, etc.)
- 2020-2026 updates to foundational refs: about 15
- Prior LLM / agent surveys (5 named in Section 3): 5
- Theoretical bridge frameworks (predictive coding and active inference): about 10
- Other (PRISMA, replication crisis): about 5
Total included: about 60 references
Studies included in quantitative synthesis:

```
|     - 45 human cognitive failure categories                    |
|     - Each coded against 5-level AI-research-status scheme     |
|     - 2 LLM coders, kappa = 0.97                               |
+----------------------------------------------------------------+
```

The counts are approximate and based on the search-thread documentation in Section 2.1. A fully systematic review with PRISMA 2020 rigor would record exact per-search counts, dates, and screening-decision logs at the per-record level. We do not claim that level of rigor; the diagram documents the process at the level of granularity our methodology supported.

A.2 Per-reference inclusion table

For each cited reference, we document: source (which database or hand-search route), inclusion criterion satisfied (numbered against Section 2.2's four criteria: 1=foundational, 2=2020-2026 update, 3=prior LLM survey, 4=cognitive-AI bridge framework), coding category it informs (or "framework"/"background" if not part of the failure-mode coding), and verification status (whether DOI / arXiv / journal-page is confirmed).

Reference	Source	Criterion	Informs	Verification
Reason 1990 *Human Error*	Foundational hand-search	1	Action-error structure, Swiss Cheese	Verified
Reason 2008 *The Human Contribution*	Citation chase	2	Action-error structure update	Verified
Sarter & Woods 1994 *Pilot Interaction with Cockpit Automation II*	Hand-search aviation literature	1	Mode confusion (cat 40)	Verified
Sarter & Woods 1997 *Team Play with a Powerful Independent Agent*	Hand-search	1	Mode confusion (cat 40)	Verified
Skraaning & Jamieson 2024 *J Cognitive Engineering*	Google Scholar "automation surprise 2024"	2	Mode confusion (cat 40) update	Verified DOI
Wickens et al 2021 5th ed	Hand-search	1	Cognitive load, channel capacity	Verified
Klein 1998 *Sources of Power*	Hand-search	1	NDM, premature closure (cat 10)	Verified
Hollnagel et al 2006 *Resilience Engineering*	Hand-search	1	Resilience, Safety II (Section 10)	Verified
Hollnagel 2012 *FRAM*	Citation chase	1	FRAM, resilience updates	Verified

Reference	Source	Criterion	Informs	Verification
Berg et al 2023 *BMJ Open* meta-narrative review of healthcare resilience	Google Scholar	2	Safety II update	Verified PMC10514640
Johnson & Raye 1981 *Reality Monitoring*	Hand-search clinical neurosci	1	Source monitoring (cat 2)	Verified
Johnson, Hashtroudi, Lindsay 1993 *Source Monitoring*	Citation chase	1	Source monitoring (cat 2)	Verified
Schnider 2003 *NRN*	PubMed "spontaneous confabulation"	1	Confabulation (cat 3)	Verified
Duncan et al 1996 *Cognitive Psychology*	Citation chase	1	Goal neglect (cat 18)	Verified
Friedman & Robbins 2022 *Neuropsychopharmacology*	Google Scholar	2	Goal neglect update	Verified DOI
Wegner 1985 transactive memory	Hand-search	1	Transactive memory (cat 27)	Verified
Cowan 2010 *Magical Mystery Four*	Citation chase	1	Working memory (cat 1)	Verified
Sweller, Ayres & Kalyuga 2023 *Educational Psych Review*	Google Scholar "cognitive load 2024"	2	Cognitive load update	Verified DOI
Evans et al 2024 *Educational Psych Review*	Google Scholar	2	Cognitive load motivation	Verified DOI
Simon 1956 *Rational Choice*	Hand-search	1	Satisficing (cat 10)	Verified
Tversky & Kahneman 1974 *Science*	Hand-search	1	Heuristics (cats 11, 12, 16)	Verified
Bikhchandani et al 1992 *J Pol Econ*	Citation chase	1	Information cascades (cat 31)	Verified

Reference	Source	Criterion	Informs	Verification
Bikhchandani & Hirshleifer 2024 *JEL*	Google Scholar "information cascades 2024"	2	Cascades update	Verified
Tump et al 2025 *Royal Society Open Science*	Google Scholar	2	Asynchronous group decisions	Verified DOI
Durrheim 2025 *Political Psychology*	Google Scholar	2	Polarization on social media	Verified DOI
Sharma et al 2024 ICLR sycophancy	arXiv	3	Sycophancy / conformity (cat 32)	Verified
Greshake et al 2023 AISec	arXiv	3	Prompt injection (Section 9)	Verified
Pezzulo et al 2024 *TICS*	TICS direct	4	Active inference (Section 5)	Verified DOI
Smith et al 2021 *PCN*	Citation chase	4	Predictive coding clinical	Verified DOI
Sprevak 2024 *Philosophy Compass*	Google Scholar	4	Predictive coding intro	Verified DOI
Georganta 2024 *JOOP*	Google Scholar	2	Human-AI teaming trust	Verified DOI
Verma et al 2025 *AI & Society*	Google Scholar	2	Automation bias review	Verified DOI
Ji et al 2023 *ACM CSur*	ACM Digital Library	3	Hallucination survey	Verified DOI
OWASP Gen AI 2025 LLM Top 10	OWASP website	3	Prompt injection canonical	Verified URL
Cemri et al 2025 (MAST) arXiv 2503.13657	arXiv	3	Multi-agent failure taxonomy	Verified arXiv
Mohammadi et al 2025 KDD	ACM	3	LLM agent benchmarking survey	Verified DOI
Hammond et al 2025 *Multi-Agent Risks from Advanced AI*	arXiv	3	Multi-agent strategic categories	Verified arXiv 2502.14143
PCIB 2026 arXiv	arXiv	4	Predictive coding hallucination detection	Verified arXiv
Hutchins 1995 *Cognition in the Wild*	Hand-search	1	Distributed cognition (cat 41)	Verified

Reference	Source	Criterion	Informs	Verification
Perrow 1984 *Normal Accidents*	Hand-search	1	Background framework	Verified
Weick 2001 *Managing the Unexpected*	Hand-search	1	High-reliability organizations	Verified
Schacter 1999 *Seven Sins of Memory*	Citation chase	1	Memory failures background	Verified
Salas et al 2008 *Human Factors*	Citation chase	1	CRM communication (cat 23)	Verified
Mackworth 1948 *QJEP*	Hand-search vigilance literature	1	Vigilance decrement (cat 7)	Verified
Onnasch et al 2014 *Human Factors*	Citation chase	1	Automation complacency	Verified
Cialdini 2001 *Influence*	Hand-search	1	Authority effects	Verified
Maynard, Kennedy & Sommer 2015 *EJWOP* team adaptation 15-year synthesis	Hand-search	1	Team adaptation	Verified
Buljac-Samardzic et al 2021 *Journal of Patient Safety* umbrella review	PubMed	2	CRM update	Verified PMC8612906
(Patterson 2002 NDM premature closure removed; premature closure cited as Klein 1998)	n/a	n/a	n/a	Removed (could not verify)
Anthropic 2024 AI Fluency Index	Anthropic website	4	Background framework	Verified URL
Janis 1972 *Victims of Groupthink*	Hand-search	1	Groupthink (cat 28)	Verified
Norman 1988 *Psychology of Everyday Things*	Hand-search	1	Affordances (cat 39)	Verified

Reference	Source	Criterion	Informs	Verification
Schwarz & Clore 1983 *J Pers & Soc Psych*	Citation chase	1	Affect-as-information (Section 10)	Verified
Slovic et al 2007 *EJOR*	Citation chase	1	Affect heuristic	Verified
Moscovici, Lage & Naffrechoux 1969 *Sociometry*	Hand-search	1	Minority influence (Section 10)	Verified
Easterbrook 1959 *Psychological Review*	Hand-search	1	Stress narrowing (cat 38)	Verified
Cannon-Bowers, Salas & Converse 1993	Citation chase	1	Shared mental models (cat 25)	Verified
Open Science Collaboration 2015 *Science*	Hand-search	1	Replication crisis (Section 10)	Verified DOI
Page et al 2021 *BMJ* (PRISMA 2020)	Hand-search	1	Methodology (Section 2)	Verified DOI
Patronus AI 2026 (TRAIL) arXiv 2505.08638	arXiv	3	Empirical companion paper	Verified arXiv
Zhang, Yin et al 2025 (Who&When) ICML 2025 Spotlight, arXiv 2505.00212	ICML proceedings	3	Empirical companion paper	Verified
Riedl, Savage & Zvelebilova 2024 *Cognitive Spillover in Human-AI Teams*	arXiv	2	Distributed cognition update	Verified arXiv 2407.17489

All previously unverified entries (six in earlier draft) have been resolved via targeted web search: - Hammond et al 2025: confirmed as arXiv 2502.14143 (Cooperative AI Foundation Technical Report 1). - Healthcare resilience review: replaced with Berg et al 2023 *BMJ Open* meta-narrative review (PMC10514640). - CRM healthcare review: replaced with Buljac-Samardzic et al 2021 *Journal of Patient Safety* umbrella review (PMC8612906) in place of the unverifiable Edwardsson 2023 entry. - Team adaptation: replaced with the canonical Maynard, Kennedy, and Sommer 2015 *EJWOP* synthesis in place of the unverifiable Maynard 2023 *J Org Behavior* entry. - Distributed cognition for human-AI teams: replaced with Riedl, Savage & Zvelebilova 2024 *Cognitive Spillover in Human-AI Teams* (arXiv 2407.17489) in place of the unverifiable Rasmussen-Pina 2023 entry.

(Earlier drafts of this review attributed arXiv 2407.17489 to Schoonderwoerd et al with a different title; the May 2026 verification pass corrected the attribution to the actual published authors and title.) - Patterson 2002 NDM premature closure: removed; the premature-closure claim is now cited via Klein 1998.

The bibliography is publication-ready as far as web verification can confirm. A peer-reviewed submission would add a final pass through institutional citation databases (Web of Science, Scopus) for any remaining verification.

Appendix B. Per-category bibliometric evidence supporting AI-research-status verdicts

To anchor the LLM-coder verdicts in Section 7.11, we performed a documented Google Scholar search pass covering categories across all 13 subdisciplines, including: every category coded *Substantial*, every category at the boundary between *Absent* and *Nascent*, all major *Nascent* clusters (memory, attention, decision-making subbiases, group / social, embodied, metacognition), and a representative selection of *Partial* verdicts. We coded approximately 30 of 45 categories with bibliometric evidence; the remaining 15 categories are validated transitively through the subdiscipline-level evidence.

B.1 Evidence for boundary and Substantial categories

#	Category	Bibliometric evidence	Verdict
26	Handoff errors	MAST (Cemri et al 2025) reports 36.9% of multi-agent failures are coordination, with 14 named failure modes including handoff-specific issues (FM-2.1 conversation reset, FM-2.4 information withholding, FM-2.5 ignored input).	Substantial

#	Category	Bibliometric evidence	Verdict
34	Role drift / persona drift	Multiple persona-drift benchmarks (arXiv 2402.10962 "Measuring and Controlling Persona Drift in Language Model Dialogs"; PersonaDrift benchmark; EchoMode protocol; Pisama persona detector).	Substantial
43	Calibration failures	NAACL 2024 survey "A Survey of Confidence Estimation and Calibration in Large Language Models"; SelectLLM; multiple recent benchmarks; established methodology around ECE (Expected Calibration Error).	Substantial
44	Egocentric bias / theory of mind	Strachan et al 2024 *Nature Human Behaviour* "Testing theory of mind in large language models and humans"; Kosinski 2024 PNAS; multiple AAAI benchmarks (ToMATO 2025, MovieGraph-ToM 2025); Frontiers 2025 on higher-order ToM.	Substantial

#	Category	Bibliometric evidence	Verdict
18	Goal neglect	Goal drift in LLM agents is now substantial: Technical Report 2505.02709 "Evaluating Goal Drift in Language Model Agents" (AIES 2025); long-horizon task literature; Lakera's agentic threat work. The Duncan et al 1996 framework is invoked in some of this literature.	Substantial
42	Dunning-Kruger / overconfidence	Multiple papers explicitly testing Dunning-Kruger in LLMs (arXiv 2603.09985 The Dunning-Kruger Effect in LLMs; Scientific Reports 2026; "Mind the Confidence Gap" arXiv 2502.11028; "The LLM Fallacy" arXiv 2604.14807). Substantial systematic engagement.	Substantial
7	Vigilance decrement	The lost-in-the-middle phenomenon (Liu et al 2023, ACL 2024) and "context rot" framing (Chroma 2025) are substantial AI literature, but they are not framed against Mackworth's vigilance decrement. The phenomenon is heavily studied; the cognitive-psychology framework is not imported.	Nascent

#	Category	Bibliometric evidence	Verdict
15	Hindsight bias	At least one recent paper addresses hindsight-bias-related concerns in agent evaluation (arXiv 2510.20603 on causal stepwise evaluation; CaSE). Limited but not zero.	Nascent (borderline)
1	Working memory limits	Multiple papers explicitly engage the Cowan / Baddeley framework: arXiv 2505.10571 "Language Models Do Not Have Human-Like Working Memory"; arXiv 2508.13171 "Cognitive Workspace: Active Memory Management for LLMs"; ACL 2024 "Working Memory Identifies Reasoning Limits in Language Models".	Partial
13	Confirmation bias	arXiv 2509.14824 "Confirmation Bias as a Cognitive Resource in LLM-Supported Deliberation"; multiple sycophancy-adjacent papers.	Partial
32	Conformity (Asch)	"Conformity in Large Language Models" (Cambridge 2024); ELEPHANT 2025 social sycophancy; multiple Asch-style probes.	Partial

B.1.1 Additional Partial-verdict evidence from the expanded pass

#	Category	Bibliometric evidence	Verdict
4	False memory / DRM-style suggestibility	Cognitive-bias-in-LLM surveys cover DRM-adjacent probes (Hagendorff 2023; Binz and Schulz 2023; arXiv 2412.00323 Cognitive Biases in LLMs survey). The kappa-borderline call between coders A and B resolves as Partial.	Partial
5	Prospective memory failures	AgentScope ReMe, locomo benchmark (snap-research), AgeMem (arXiv 2601.01885), long-horizon agent planning literature explicitly addresses prospective-memory-style task scheduling.	Partial
12	Availability heuristic	Cognitive-bias-in-LLM surveys (arXiv 2412.00323; ScienceDirect on anchoring effects) cover availability with operationalized measurement.	Partial
19	Perseveration / loop	Substantial body on degeneration loops, mode collapse, repetition (SpecRA, Echo Trap, arXiv 2512.04419 Solving LLM Repetition). Verdict sits on the Partial / Substantial boundary; we keep Partial conservatively.	Partial

#	Category	Bibliometric evidence	Verdict
27	Transactive memory failures	LLMA-Mem framework, Memory in LLM-based Multi-agent Systems (TechRxiv preprint), Collaborative Memory (arXiv 2505.18279). The transactive-memory framework is explicitly invoked in multiple recent papers.	Partial
30	Group polarization	Decoding Echo Chambers (arXiv 2409.19338), AgentSociety (arXiv 2502.08691), Emergent social conventions (Science Advances 2025) explicitly study polarization in LLM populations.	Partial
35	Moral disengagement	Moral Alignment for LLM Agents (arXiv 2410.01639), LLM Ethics Benchmark (arXiv 2505.00853). Bandura's framework is invoked in the alignment literature though not centrally.	Partial

B.2 Distribution summary

After the expanded bibliometric pass, the per-verdict counts across the 45 categories are:

Verdict	Count	Share
Substantial	6	13%
Partial	24	53%
Nascent	12	27%
Absent	0	0%
Substrate-absent	3	7%

The six *Substantial* categories are: handoff errors (cat 26), persona drift (cat 34), calibration

failures (cat 43), egocentric / theory of mind (cat 44), goal neglect (cat 18), and Dunning-Kruger / overconfidence (cat 42). The three *Substrate-absent* categories are time-on-task fatigue (cat 36), sleep deprivation (cat 37), and stress-induced cognitive narrowing (cat 38).

Implication for the headline claim. The "minority of categories have substantial AI engagement" framing holds at 6 of 45 (13%). The "majority have nascent or partial engagement" claim is well supported: 36 of 45 (80%) are Partial or Nascent (24 + 12). The "absent" category is empty: every well-studied human cognitive failure category has at least nascent AI engagement once the recent literature is fully canvassed.

The research-roadmap implication is also stronger, not weaker. 12 of 45 (27%) Nascent categories represent areas where AI agent research has touched the framework but not built systematic detection or theory around it. These remain the productive directions for AI agent reliability research, in addition to the substantive work that has been done in the 24 Partial and 6 Substantial areas.

B.3 Methodology and limitations

Search procedure. For each category, a Google Scholar query was constructed using the category name + LLM-relevant qualifiers (e.g., "LLM agent", "language model", "multi-agent"). The top 10 results were inspected for: (a) whether the AI literature engages the human framework explicitly, (b) the size of the LLM-side corpus, (c) the existence of established benchmarks. The verdict was assigned conservatively when boundary cases arose.

Sample size. This validation pass spot-checked 10 of 45 categories. A fully systematic bibliometric validation would code all 45. The 10 chosen include the *Substantial* and *Absent / Nascent boundary* categories most sensitive to coding criteria, with a representative selection from the wider boundary cases identified in Section 2.4's kappa analysis. The remaining 35 categories' coding stands unless contradicted by bibliometric evidence we did not search for.

Single-coder limitation persists. The bibliometric pass was performed by Coder A and is therefore subject to the same single-coder bias. A fully rigorous bibliometric validation would have a second human coder repeat the search-and-classification independently.

Bibliometric measurement is imperfect. Hit counts on Google Scholar are noisy and reflect search-engine quirks. The number of papers in a literature is not the same as the substantive engagement with a framework. We used hit counts as one input; the substantive judgment was based on inspecting the top results.

The validation pass is offered as a real but limited methodological strengthening. It moves the methodology from "expert judgment alone" to "expert judgment partially anchored in bibliometric data." A peer-reviewed submission would strengthen this further by extending the pass to all 45 categories with a second human bibliometric coder.

Appendix C. Crosswalk between MAST's 14 multi-agent failure modes and the 45-category human cognitive failure taxonomy

This appendix provides a complete mode-by-mode mapping from MAST (Cemri et al 2025; arXiv:2503.13657; NeurIPS 2025 Datasets and Benchmarks Track) to the 45-category human cognitive failure taxonomy in Section 6, and to the corresponding row(s) in the Section 7.11

mapping table. The mapping was constructed jointly with the AI-side inventory pass and applies the same five-level AI-research-status scheme. Each MAST mode is assigned a *primary* and (where applicable) *secondary* human-category mapping, a transfer dimension verdict (surface / mechanism / intervention; see Section 1), and a brief mechanism note.

C.1 The MAST 14 modes and their human cognitive analogs

Table C.1: Crosswalk from MAST-14 modes to human cognitive failure categories. Each row gives the MAST failure mode (from Cemri et al 2025), the primary human cognitive failure category (numbered as in Section 7.11 / Section 6), any secondary human category, the transfer-dimension verdict, and a one-line mechanism note. The "Transfer" column uses the codes **S** (strong analog: surface + mechanism + intervention all transfer), **M** (surface + mechanism transfer; intervention partial), **U** (surface only; mechanism diverges), and **N** (no clean human predecessor).

MAST mode	Primary human category	Secondary	Transfer	Mechanism note
FC1: Specification and System Design				
FM-1.1 Disobey task specification	18 Goal neglect (Duncan et al 1996)	22 Rule-based / knowledge-based mistakes (Reason 1990)	M	Goal-representation failure: the task representation is held but not deployed during action selection.
FM-1.2 Disobey role specification	34 Role drift / deindividuation	33 Obedience / role override	M	Role schema decays under accumulating user-turn evidence; persona-drift literature is the closest AI analog.
FM-1.3 Step repetition	19 Perseveration	(none)	M	Frontal-lobe perseveration in humans; mode-collapse / loop dynamics in LLMs. Mechanism diverges in detail but surface and intervention design transfer.

MAST mode	Primary human category	Secondary	Transfer	Mechanism note
FM-1.4 Loss of conversation history	1 Working memory limits (Cowan 2010)	2 Source monitoring (Johnson and Raye 1981)	M	Context-window saturation produces a working-memory-like limit; source provenance is degraded as turns accumulate.
FM-1.5 Unaware of termination conditions	20 Planning fallacy	10 Premature closure / satisficing	U	The agent does not represent the termination criterion well enough to evaluate it; planning-fallacy framing is loose, mechanism is more like an under-specified evaluation function.
FC2: Inter-Agent Misalignment				
FM-2.1 Conversation reset	2 Source monitoring / reality monitoring	1 Working memory limits	M	Loss of provenance / continuity for established facts; closest human analog is anterograde-style failure of source monitoring across context resets.

MAST mode	Primary human category	Secondary	Transfer	Mechanism note
FM-2.2 Fail to ask for clarification	10 Premature closure / satisficing	25 Shared mental model breakdown	M	Naturalistic-decision-making framing of premature closure (Klein 1998) directly transfers; clarification-asking is one of the recommended human-side interventions.
FM-2.3 Task derailment	18 Goal neglect	19 Perseveration on a non-goal sub-task	M	Same mechanism as FM-1.1 but emerges through inter-agent dynamics rather than single-agent goal-representation decay.
FM-2.4 Information withholding	27 Transactive memory failures (Wegner 1985)	26 Handoff errors	M	The "who knows what" structure is incomplete: information held by agent A is not surfaced for agent B's task. Wegner's transactive-memory framework directly applies.
FM-2.5 Ignored other agent's input	25 Shared mental model breakdown (Cannon-Bowers et al 1993)	23 CRM communication failures	M	The recipient agent does not incorporate the sender's contribution; CRM literature on closed-loop communication is the canonical intervention space.

MAST mode	Primary human category	Secondary	Transfer	Mechanism note
FM-2.6 Reasoning-action mismatch	40 Mode confusion / automation surprise (Sarter and Woods 1994)	21 Skill-based slips (Reason 1990)	M	The agent's stated reasoning and the executed action diverge; mode-confusion framing applies because the agent operates in a mode different from the one its reasoning trace describes.
FC3: Task Verification and Termination				
FM-3.1 Premature termination	10 Premature closure / satisficing	20 Planning fallacy	S	Direct analog to satisficing in naturalistic decision making. Klein's recognition-primed decision model and premature-closure mitigations transfer with adaptation.
FM-3.2 No or incomplete verification	43 Calibration failures	42 Dunning-Kruger / overconfidence	M	The verification step is omitted or underspecified; calibration / selective-prediction literature is the operational AI counterpart.
FM-3.3 Incorrect verification	42 Dunning-Kruger / overconfidence	43 Calibration failures	M	The verification is performed but the verifier is overconfident in a regime where its competence is poor.

C.2 Coverage statistics for the crosswalk

Of MAST's 14 modes, **0** map to "no human predecessor" (N): every MAST mode has at least a surface-level human cognitive analog. **1** mode (FM-1.5) has weak mechanism transfer (U): the planning-fallacy framing is loose and the underlying issue is closer to an under-specified evaluation function than a planning-time misestimation. **12** modes have surface + mechanism transfer (M), with intervention-transfer ranging from partial to strong. **1** mode (FM-3.1) has strong full transfer (S) including direct intervention applicability from the naturalistic-decision-making literature.

This pattern is consistent with the present review's broader finding (Section 7.11): when AI agent failures are systematically catalogued, almost all of them have at least surface-level human cognitive analogs; the productive divergences are at the mechanism level (where LLMs differ from humans because of substrate properties; see Section 5) rather than at the surface level.

C.3 Categories in Section 7.11 with no MAST counterpart

The complement of the crosswalk is informative. The following categories from Section 7.11 have no MAST counterpart because MAST's scope is multi-agent execution failures, while the human cognitive failure taxonomy includes single-agent perceptual, decision-making, metacognitive, and theory-of-mind categories that are out of scope for MAST:

- **Attention and perception** (categories 6 selective attention, 7 vigilance decrement, 8 inattentional blindness, 9 attentional capture) — single-agent perception phenomena.
- **Decision-making subbiases** (11 anchoring, 12 availability, 13 confirmation, 14 motivated reasoning, 15 hindsight, 16 base-rate neglect, 17 sunk-cost) — single-agent reasoning biases.
- **Memory subcategories** (3 spontaneous confabulation, 4 false memory, 5 prospective memory) — primarily single-agent memory failures.
- **Group and social** (28 groupthink, 29 social loafing, 30 group polarization, 31 information cascades, 32 conformity, 33 obedience) — multi-agent but at the *population / social-dynamics* level, beyond the conversation-lifecycle framing of MAST.
- **Identity, role, moral** (35 moral disengagement) — partially overlapping with FM-1.2 but the moral-disengagement framing has no clean MAST cell.
- **Stress / fatigue / embodiment** (36–39, 41) — substrate-absent or substrate-orthogonal in LLMs.
- **Metacognition / theory of mind** (44 egocentric bias / theory of mind, 45 attribution errors) — while FM-3.2 and FM-3.3 touch metacognitive verification, machine theory of mind and fundamental attribution error are richer cognitive phenomena that MAST does not separately address.
- **Distributed cognition** (41 Hutchins) — substrate-orthogonal in the current AI agent paradigm.

The implication for joint research-agenda planning (Section 14) is that MAST's 14 modes are a *subset* of the multi-agent failure space, focused on the conversation lifecycle. Extending MAST or building complementary taxonomies for the seven category groups above is an open research direction; the present review's 45-category taxonomy is one such complement.

C.4 Methodology note for Appendix C

The crosswalk in Section C.1 was constructed as follows. The 14 MAST mode names were drawn from the published Cemri et al 2025 paper. The primary and secondary human-category as-

signments were proposed by Coder A (Claude Opus 4.7) and reviewed against the bibliometric evidence in Appendix B, the worked examples in Sections 7.1–7.10, and the dispatch table at `backend/app/core/mast_constants.py` in the companion empirical paper's repository. The transfer-dimension verdicts apply the protocol in Section 1 (surface / mechanism / intervention transfer) and the operational definitions in Section 2.3. As with the main mapping table in Section 7.11, this is a single-coder construction subject to the limitations enumerated in Sections 2.4 and B.3, and is offered for joint review by a domain-expert second coder. The proposed human-coder validation pass in Section 2.4 covers Appendix C alongside Section 7.11.

Bibliography

- Lou, Jiaxu, and Sun, Yifan (2026). Anchoring Bias in Large Language Models: An Experimental Study. *Journal of Computational Social Science* 9(1). doi:10.1007/s42001-025-00435-2. arXiv:2412.06593. Validates anchoring effects in LLM judgment with mitigation evaluation; relevant to cat 11.
- Zhou, Xinyi, Saghi, Zeinadsadat, Sabouri, Sadra, Pandita, Rahul, McGuire, Mollie, and Chattopadhyay, Souti (2026). Cognitive Biases in LLM-Assisted Software Development. arXiv:2601.08045. First comprehensive study of cognitive biases in LLM-assisted programming (mixed-methods: 14 student/professional developers + survey of 22 additional). Documents that 48.8% of programmer actions are biased and developer-LLM interactions account for 56.4% of biased actions. Develops a taxonomy of 15 bias categories validated by cognitive psychologists. Relevant to cats 11–14, 17 in the §6.2 / §7.2 cognitive-bias cluster; first applied study showing the LLM-assisted-development context produces distinct (and additional) biases beyond the standalone-LLM cognitive-bias literature.
- Zhu, Xiaochen, Zhang, Caiqi, Stafford, Tom, Collier, Nigel, and Vlachos, Andreas (2024). Conformity in Large Language Models. arXiv:2410.12428. Demonstrates Asch-style conformity in LLMs; conformity rises with model uncertainty; instruction-tuned models conform less. Relevant to cat 32.
- Patronus AI (2026). TRAIL: A Benchmark for Multi-Agent Failure Detection. arXiv:2505.08638.
- Alansari, Aisha, and Luqman, Hamzah (2025). Large Language Models Hallucination: A Comprehensive Survey. arXiv:2510.06265. Updated successor to Ji et al 2023 ACM CSur survey: taxonomy of hallucination types and root causes across the LLM lifecycle, structured taxonomy of detection approaches (retrieval-, uncertainty-, embedding-, learning-, self-consistency-based), and mitigation strategies organized by stage (data-centric, model-centric, inference-time). Used in §3.1 as the up-to-date hallucination reference.
- Anthropic (2024). AI Fluency Index.
- Berg, Siri Hatlen, and others (2023). Healthcare Resilience: A Meta-Narrative Systematic Review and Synthesis of Reviews. *BMJ Open*. PMC10514640; PubMed 37730383. Updates resilience-engineering literature in healthcare context.
- Bikhchandani, Sushil, Hirshleifer, David, and Welch, Ivo (1992). A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades. *Journal of Political Economy* 100(5), 992–1026.
- Bikhchandani, Sushil, and Hirshleifer, David (2024). Information Cascades and Social Learning. *Journal of Economic Literature*.
- Buljac-Samardzic, Martina, Doekhie, Kirti D., and van Wijngaarden, Jeroen D. H. (2021). What Do We Really Know About Crew Resource Management in Healthcare? An Umbrella

Review on Crew Resource Management and Its Effectiveness. *Journal of Patient Safety*. PMC8612906; PubMed 34852415.

- Cannon-Bowers, Janis A., Salas, Eduardo, and Converse, Sharolyn (1993). Shared Mental Models in Expert Team Decision Making. *Individual and Group Decision Making: Current Issues*, 221–246.
- Cemri, Mert, Pan, Melissa Z., Yang, Shuyi, Agrawal, Lakshya A., Chopra, Bhavya, Tiwari, Rishabh, Keutzer, Kurt, Parameswaran, Aditya, Klein, Dan, Ramchandran, Kannan, Zaharia, Matei, Gonzalez, Joseph E., and Stoica, Ion (2025). Why Do Multi-Agent LLM Systems Fail?. arXiv:2503.13657. *NeurIPS 2025 Datasets and Benchmarks Track*. MAST taxonomy: 14 failure modes across 3 categories (System Design, Inter-Agent Misalignment, Task Verification). 1,600+ annotated traces. Reports human-human Cohen's kappa = 0.88; LLM-judge vs human kappa = 0.77. Primary AI-side anchor of this review.
- Chen, Ruirui, Jiang, Weifeng, Qin, Chengwei, and Tan, Cheston (2025). Theory of Mind in Large Language Models: Assessment and Enhancement. *ACL 2025 main conference*. arXiv:2505.00026. Survey of LLM ToM evaluation benchmarks and recent enhancement strategies. Cited in §7.11 row 44 alongside Riemer et al 2025 as the recent state-of-the-art on machine ToM.
- Cialdini, Robert B. (2001). Influence: Science and Practice. *Allyn and Bacon*.
- Collaboration, Open Science (2015). Estimating the Reproducibility of Psychological Science. *Science* 349(6251), aac4716. doi:10.1126/science.aac4716.
- Cowan, Nelson (2010). The Magical Mystery Four: How Is Working Memory Capacity Limited, and Why?. *Current Directions in Psychological Science* 19(1), 51–57.
- Duncan, John, Emslie, Hazel, Williams, Phyllis, Johnson, Roger, and Freer, Charles (1996). Intelligence and the Frontal Lobe: The Organization of Goal-Directed Behavior. *Cognitive Psychology* 30(3), 257–303.
- Durrheim, Kevin, and others (2025). Polarization on Social Media. *Political Psychology*. doi:10.1111/pops.70000.
- Easterbrook, J. A. (1959). The Effect of Emotion on Cue Utilization and the Organization of Behavior. *Psychological Review* 66(3), 183–201.
- Evans, Paul, Vansteenkiste, Maarten, Parker, Philip, and others (2024). Cognitive Load Theory and Its Relationships with Motivation: a Self-Determination Theory Perspective. *Educational Psychology Review*. doi:10.1007/s10648-023-09841-2.
- Friedman, Naomi P., and Robbins, Trevor W. (2022). The Role of Prefrontal Cortex in Cognitive Control and Executive Function. *Neuropsychopharmacology* 47(1), 72–89. doi:10.1038/s41386-021-01132-0.
- Georganta, Eleni, and others (2024). Would You Trust an AI Team Member? Team Trust in Human-AI Teams. *Journal of Occupational and Organizational Psychology*. doi:10.1111/joop.12504.
- Greshake, Kai, Abdelnabi, Sahar, Mishra, Shailesh, Endres, Christoph, Holz, Thorsten, and Fritz, Mario (2023). Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (AISec)*.
- Sumita, Yasuaki, Takeuchi, Koh, and Kashima, Hisashi (2024). Cognitive Biases in Large Language Models: A Survey and Mitigation Experiments. arXiv:2412.00323; Survey covering anchoring, availability, base-rate neglect, confirmation, framing, and other biases in LLMs, plus two crowdsourcing-inspired mitigation methods (SoPro and AwaRe). Informs cats 11-13, 16 of the human-failure taxonomy.
- Hagendorff, Thilo, Dasgupta, Ishita, Binz, Marcel, Chan, Stephanie C. Y., Lampinen,

Andrew, Wang, Jane X., Akata, Zeynep, and Schulz, Eric (2023). Machine Psychology. arXiv:2303.13988. Programmatic article advocating behavioural-psychology paradigms for LLM evaluation; surveys experimental paradigms, computational analysis techniques, and caveats of applying human-cognition methods to LLMs. Recurring touchstone for the cognitive-bias-in-LLM strand referenced in cats 4, 11-13, 16.

- Binz, Marcel, and Schulz, Eric (2023). Using Cognitive Psychology to Understand GPT-3. *PNAS* 120(6), e2218523120. doi:10.1073/pnas.2218523120. arXiv:2206.14576. Applies vignette-based cognitive-psychology paradigms (decision-making, information search, deliberation, causal reasoning) to GPT-3; finds GPT-3 matches or beats humans on vignettes and bandit tasks but fails on causal-reasoning tasks and is brittle to perturbations. Recurring reference for cats 4, 16.
- Hammond, Lewis, and others (2025). Multi-Agent Risks from Advanced AI. arXiv:2502.14143; Cooperative AI Foundation Technical Report 1, February 2025. Three failure modes: miscoordination, conflict, collusion. Seven risk factors: information asymmetries, network effects, selection pressures, destabilising dynamics, commitment problems, emergent agency, multi-agent security.
- Hollnagel, Erik (2012). FRAM: The Functional Resonance Analysis Method. *Ashgate*.
- Hutchins, Edwin (1995). Cognition in the Wild. *MIT Press*.
- Janis, Irving L. (1972). Victims of Groupthink. *Houghton Mifflin*.
- Ji, Ziwei, Lee, Nayeon, Frieske, Rita, Yu, Tiezheng, Su, Dan, Xu, Yan, Ishii, Etsuko, Bang, Yejin, Madotto, Andrea, and Fung, Pascale (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys* 55(12), 1–38. doi:10.1145/3571730.
- Johnson, Marcia K., and Raye, Carol L. (1981). Reality Monitoring. *Psychological Review* 88(1), 67–85.
- Johnson, Marcia K., Hashtroudi, Shahin, and Lindsay, D. Stephen (1993). Source Monitoring. *Psychological Bulletin* 114(1), 3–28.
- Kadavath, Saurav, and others (2022). Language Models (Mostly) Know What They Know. arXiv:2207.05221; Anthropic. Foundational LLM-calibration reference; cited in cat 43 verdict.
- Klein, Gary (1998). Sources of Power: How People Make Decisions. *MIT Press*.
- Kuran, Timur (1995). *Private Truths, Public Lies: The Social Consequences of Preference Falsification*. Harvard University Press. Foundational work on preference falsification as a mechanism of public-vs-private opinion divergence under social cost. Cited in Section 9 as a surface analog (with mechanism divergence) for sycophancy cascade in LLMs.
- Mitnick, Kevin D., and Simon, William L. (2002). *The Art of Deception: Controlling the Human Element of Security*. Wiley. Operational reference on social-engineering attacks; cited in Section 9 alongside Cialdini 2001 as the surface-analog literature for prompt injection.
- Kosinski, Michal (2024). Evaluating Large Language Models in Theory of Mind Tasks. *PNAS*. doi:10.1073/pnas.2405460121; Eleven LLMs assessed on 40 false-belief tasks; key reference for cat 44.
- Liu, Nelson F., and others (2024). Lost in the Middle: How Language Models Use Long Contexts. *TACL*. Documents U-shaped attention pattern in long-context LLMs; cited as the surface phenomenon underlying the vigilance-decrement-analog (cat 7).
- Mackworth, Norman H. (1948). The Breakdown of Vigilance during Prolonged Visual Search. *Quarterly Journal of Experimental Psychology* 1, 6–21.
- Maynard, M. Travis, Kennedy, Deanna M., and Sommer, S. Amy (2015). Team Adaptation: A Fifteen-Year Synthesis (1998-2013) and Framework for How This Literature Needs to "Adapt'' Going Forward. *European Journal of Work and Organizational Psychology* 24(5), 652–677.
- Mohammadi, Mahmoud, and others (2025). Evaluation and Benchmarking of LLM Agents:

A Survey. *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*. doi:10.1145/3711896.3736570.

- Moscovici, Serge, Lage, Elisabeth, and Naffrechoux, Martine (1969). Influence of a Consistent Minority on the Responses of a Majority in a Color Perception Task. *Sociometry* 32(4), 365–380.
- Norman, Donald A. (1988). The Psychology of Everyday Things. *Basic Books*.
- Onnasch, Linda, Wickens, Christopher D., Li, Huiyang, and Manzey, Dietrich (2014). Human Performance Consequences of Stages and Levels of Automation. *Human Factors* 56(3), 476–488.
- Page, Matthew J., McKenzie, Joanne E., Bossuyt, Patrick M., and others (2021). The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews. *BMJ* 372. doi:10.1136/bmj.n71.
- Perrow, Charles (1984). Normal Accidents: Living with High-Risk Technologies. *Basic Books*.
- Pezzulo, Giovanni, Parr, Thomas, Cisek, Paul, Clark, Andy, and Friston, Karl (2024). Generating Meaning: Active Inference and the Scope and Limits of Passive AI. *Trends in Cognitive Sciences*. doi:10.1016/j.tics.2023.10.018.
- Project, OWASP Gen AI Security (2025). OWASP Gen AI Security Project: Top 10 for LLM Applications 2025. Canonical operational reference for prompt injection (LLM01:2025) and related LLM application risks.
- Reason, James (1990). Human Error. *Cambridge University Press*.
- Riemer, Matthew, Ashktorab, Zahra, Bouneffouf, Djallel, Das, Payel, Liu, Miao, Weisz, Justin D., and Campbell, Murray (2025). Position: Theory of Mind Benchmarks are Broken for Large Language Models. *ICML 2025*. arXiv:2412.19726. Argues that most ToM benchmarks fail because they cannot evaluate how LLMs adjust to new interaction partners. Distinguishes *literal* ToM (predicting others' behavior) from *functional* ToM (adapting to partners in real interactions); finds that many open-source LLMs are strong on literal but struggle with functional ToM even under simple partner policies. Cited in §7.11 row 44 as the qualifying caveat to the cat 44 Substantial verdict.
- Rose, Aaron, Cullen, Carissa, Kaplowitz, Brandon Gary, and Schroeder de Witt, Christian (2026). Detecting Multi-Agent Collusion Through Multi-Agent Interpretability. arXiv:2604.01151. Introduces NARCBench for collusion detection under environment distribution shift; five activation-probing techniques; 1.00 AUROC in-distribution and 0.60–0.86 zero-shot transfer including a steganographic blackjack task. Cited in §9 (emergent collusion subsection) as the empirical anchor for the new fourth no-clean-cognitive-mechanism-predecessor failure category added in v3.
- Reason, James (2008). The Human Contribution: Unsafe Acts, Accidents and Heroic Recoveries. *Ashgate*.
- Salas, Eduardo, DiazGranados, Deborah, Klein, Cameron, and others (2008). Does Team Training Improve Team Performance? A Meta-Analysis. *Human Factors* 50(6), 903–933.
- Sarter, Nadine B., and Woods, David D. (1994). Pilot Interaction with Cockpit Automation II: An Experimental Study of Pilots' Model and Awareness of the Flight Management System. *The International Journal of Aviation Psychology* 4(1), 1–28.
- Sarter, Nadine B., and Woods, David D. (1997). Team Play with a Powerful and Independent Agent: A Corpus of Operational Experiences and Automation Surprises on the Airbus A-320. *Human Factors* 39(4), 553–569.
- Schacter, Daniel L. (1999). The Seven Sins of Memory: Insights from Psychology and Cognitive Neuroscience. *American Psychologist* 54(3), 182–203.
- Shapira, Itai, Benadè, Gerdus, and Procaccia, Ariel D. (2026). How RLHF Amplifies Syco-

phancy. arXiv:2602.01002. Provides the formal causal characterization of sycophancy amplification through preference-based post-training: amplification is determined by a covariance under the base policy between endorsing the belief signal in the prompt and the learned reward, with the first-order effect reducing to a simple mean-gap condition under Bradley-Terry-style reward learning. Derives a closed-form *agreement penalty* on the reward as the unique minimal correction. Cited in §9 sycophancy cascade as the load-bearing mechanism reference; supersedes the v1 phrasing of "RLHF reward shaping rewards user-pleasing responses."

- Schnider, Armin (2003). Spontaneous Confabulation and the Adaptation of Thought to Ongoing Reality. *Nature Reviews Neuroscience* 4(8), 662–671.
- Riedl, Christoph, Savage, Saiph, and Zvelebilova, Josie (2024). Cognitive Spillover in Human-AI Teams. arXiv:2407.17489. Two randomised experiments showing that AI exposure causally spills over into human-human interaction, affecting shared language, collective attention, shared mental models, and social cohesion; argues for treating AI as a "social forcefield." Relevant for the distributed-cognition gap in Section 10 / Table 2.
- Schwarz, Norbert, and Clore, Gerald L. (1983). Mood, Misattribution, and Judgments of Well-Being: Informative and Directive Functions of Affective States. *Journal of Personality and Social Psychology* 45(3), 513–523.
- Sharma, Mrinank, Tong, Meg, Korbak, Tomasz, Duvenaud, David, Askell, Amanda, and others (2024). Towards Understanding Sycophancy in Language Models. *ICLR*. arXiv:2310.13548.
- Simon, Herbert A. (1956). Rational Choice and the Structure of the Environment. *Psychological Review* 63(2), 129–138.
- Skraaning, Gyrd, and Jamieson, Greg A. (2024). The Failure to Grasp Automation Failure. *Journal of Cognitive Engineering and Decision Making*. doi:10.1177/15553434231189375.
- Slovic, Paul, Finucane, Melissa L., Peters, Ellen, and MacGregor, Donald G. (2007). The Affect Heuristic. *European Journal of Operational Research* 177(3), 1333–1352.
- Smith, Ryan, Badcock, Paul, and Friston, Karl J. (2021). Recent Advances in the Application of Predictive Coding and Active Inference Models within Clinical Neuroscience. *Psychiatry and Clinical Neurosciences*. doi:10.1111/pcn.13138.
- Sprevak, Mark (2024). Predictive Coding I: Introduction. *Philosophy Compass*. doi:10.1111/phc3.12950.
- Strachan, J. W. A., and others (2024). Testing Theory of Mind in Large Language Models and Humans. *Nature Human Behaviour*. Direct comparison of LLM and human theory-of-mind performance; canonical reference for cat 44.
- Sweller, John, Ayres, Paul, and Kalyuga, Slava (2023). The Development of Cognitive Load Theory: Replication Crises and Incorporation of Other Theories Can Lead to Theory Expansion. *Educational Psychology Review*. doi:10.1007/s10648-023-09817-2.
- Tump, Alan N., and others (2025). Asynchrony Rescues Statistically Optimal Group Decisions from Information Cascades through Emergent Leaders. *Royal Society Open Science*. doi:10.1098/rsos.230175.
- Tversky, Amos, and Kahneman, Daniel (1974). Judgment under Uncertainty: Heuristics and Biases. *Science* 185, 1124-1131.
- Hollnagel, Erik, Woods, David D., and Leveson, Nancy (eds) (2006). *Resilience Engineering: Concepts and Precepts*. Ashgate. Foundational anthology of resilience engineering; reframes safety as an emergent property maintained by adaptive capacity rather than as the absence of failure.
- Li, Kenneth, Liu, Tianle, Bashkansky, Naomi, Bau, David, Viégas, Fernanda, Pfister, Hanspeter, and Wattenberg, Martin (2024). Measuring and Controlling Instruction

(In)Stability in Language Model Dialogs. arXiv:2402.10962. Instruction-stability benchmark via self-chats: instructions degrade significantly within 8 turns for popular models; transformer attention decay is implicated; split-softmax mitigation proposed. Used in this review as the canonical AI-side reference for cat 34 (role drift / persona drift). Note: the authors' own framing is *instruction stability*; we interpret it as a persona-drift benchmark via the role-instruction component of system prompts.

- Tran-Truong, Phat T., and Le, Xuan-Bach (2026). Measuring the Unmeasurable: Markov Chain Reliability for LLM Agents. arXiv:2604.24579. TraceToChain pipeline: fits agent execution traces to an absorbing discrete-time Markov chain with diagnostics (composite AIC + Kolmogorov-Smirnov goodness-of-fit), Laplace-smoothed maximum-likelihood transitions, and Dirichlet-posterior credible intervals. Shows that pass@k, pass^k, and reliability decay curves are projections of one success-time distribution. Relevant to §10 (research gaps in reliability mathematics for agent ops) and the methodology bar discussion in §13.5.
- Tennant, Elizaveta, Hailes, Stephen, and Musolesi, Mirco (2024). Moral Alignment for LLM Agents. arXiv:2410.01639. Intrinsic-reward design encoding Deontological and Utilitarian frameworks for RL-based fine-tuning of LLM agents; evaluated on Iterated Prisoner's Dilemma. Informs cat 35 (moral disengagement).
- Wang, Chenxi, Liu, Zongfang, Yang, Dequan, and Chen, Xiuying (2024). Decoding Echo Chambers: LLM-Powered Simulations Revealing Polarization in Social Networks. arXiv:2409.19338. LLM-driven simulations of opinion-network dynamics; reproduces polarisation and echo-chamber phenomena; proposes active and passive nudge mitigations. Informs cat 30.
- Geng, Jiahui, Cai, Fengyu, Wang, Yuxia, Koeppl, Heinz, Nakov, Preslav, and Gurevych, Iryna (2024). A Survey of Confidence Estimation and Calibration in Large Language Models. *NAACL 2024*. arXiv:2311.08298. Canonical recent survey organising calibration methods by confidence source (logit-based, sampling-based, verbalised, training-time) and objective (selective prediction, abstention, ECE). The foundational reference for cat 43.
- Hopman, Mia, Elstner, Jannes, Avramidou, Maria, Prasad, Amritanshu, and Lindner, David (2026). Evaluating and Understanding Scheming Propensity in LLM Agents. arXiv:2603.01608. Decomposes scheming incentives into agent and environmental factors; finds only minimal scheming despite high environmental incentives; documents fragility (removing a single tool drops the scheming rate from 59% to 3%). Relevant for §10 (research gaps) and as recent evidence that scheming is a real but currently-rare failure mode under realistic agent conditions.
- Arike, Rauno, Donoway, Elizabeth, Bartsch, Henning, and Hobbhahn, Marius (2025). Technical Report: Evaluating Goal Drift in Language Model Agents. arXiv:2505.02709. Apollo Research methodology for measuring goal drift under competing-objective environmental pressure; best scaffolded Claude 3.5 Sonnet maintains near-perfect goal adherence over 100,000 tokens; goal drift correlates with rising pattern-matching susceptibility as context grows. Justifies the cat 18 verdict revision to Substantial.
- Ashery, Ariel Flint, Aiello, Luca Maria, and Baronchelli, Andrea (2025). Emergent Social Conventions and Collective Bias in LLM Populations. *Science Advances* 11(20), eadu9368. arXiv:2410.08948. Decentralised LLM-agent populations spontaneously converge on global social conventions; strong collective biases emerge even when individual agents exhibit no bias; committed adversarial minorities can drive convention switching. Relevant to cats 28-32 (group dynamics).
- Piao, Jinghua, Yan, Yuwei, Zhang, Jun, Li, Nian, Yan, Junbo, Lan, Xiaochong, Lu, Zhihong, Zheng, Zhiheng, Wang, Jing Yi, Zhou, Di, Gao, Chen, Xu, Fengli, Zhang, Fang, Rong,

Ke, Su, Jun, and Li, Yong (2025). AgentSociety: Large-Scale Simulation of LLM-Driven Generative Agents. arXiv:2502.08691. 10,000-agent / 5-million-interaction simulator with five social-issue scenarios (polarisation, inflammatory-message spread, UBI, hurricane shock, urban sustainability). Relevant to cats 28-32.

- Wang, Weiwei, Zou, Weijie, and Min, Jiyong (2025). Solving LLM Repetition Problem in Production: A Comprehensive Study of Multiple Solutions. arXiv:2512.04419. Three repetition-pattern taxonomy in deployed LLMs; greedy decoding's failure to escape repetitive loops as root cause; Beam Search with early stopping, presence-penalty, and DPO fine-tuning evaluated as mitigations. Informs cat 19 (perseveration / loop).
- Anonymous authors (OpenReview, 2025). SpecRA: Monitor Degenerative Repetition in LLM Agents using Randomized FFT. OpenReview submission xVO4BqmzVD. Fast spectral detector for approximate repetition in LLM agents; FFT-based autocorrelation in O(N log N) with robustness to lexical variation. Cited alongside Wang et al 2025 in cat 19.
- Hong, Kelly, Troynikov, Anton, and Huber, Jeff (2025). Context Rot: How Increasing Input Tokens Impacts LLM Performance. Chroma Research technical report (trychroma.com/research/context-rot). Empirical evaluation of 18 frontier LLMs (GPT-4.1, Claude Opus 4, Gemini 2.5, Qwen3) showing non-uniform performance degradation with input length even on minimal tasks. Relevant to cat 7 (vigilance-decrement analog) and cat 1 (working-memory limit).
- Bhatt, Manish (2026). Predictive Coding and Information Bottleneck for Hallucination Detection in Large Language Models. arXiv:2601.15652. PCIB hybrid hallucination detector combining predictive-coding and information-bottleneck signals with engineered enhancements; 0.8669 AUROC on HaluBench at 75x less training data and 1000x faster inference than Lynx baseline. Cited in §5.4 as an empirical translation of predictive-coding theory.
- Ghosh, Sudipta, and Panday, Mrityunjoy (2026). The Dunning-Kruger Effect in Large Language Models: An Empirical Study of Confidence Calibration. arXiv:2603.09985. Four models (Claude Haiku 4.5, Gemini 2.5 Pro, Gemini 2.5 Flash, Kimi K2) across 24,000 trials show Dunning-Kruger-shaped pattern: weaker models over-confident (Kimi K2 ECE 0.726 at 23.3% accuracy), stronger models calibrated better (Claude Haiku 4.5 ECE 0.122 at 75.4% accuracy). Justifies the cat 42 verdict revision to Substantial.
- Yu, Yi, Yao, Liuyi, Xie, Yuexiang, Tan, Qingquan, Feng, Jiaqi, Li, Yaliang, and Wu, Libing (2026). Agentic Memory: Learning Unified Long-Term and Short-Term Memory Management for Large Language Model Agents. arXiv:2601.01885. AgeMem unified memory framework integrating LTM and STM management as agent tool-actions; three-stage progressive RL with step-wise GRPO. Informs cats 5 (prospective memory) and 27 (transactive memory).
- Verma, Suresh, and others (2025). Exploring Automation Bias in Human-AI Collaboration: A Review and Implications for Explainable AI. *AI & Society*. doi:10.1007/s00146-025-02422-7.
- Wegner, Daniel M. (1985). Transactive Memory: A Contemporary Analysis of the Group Mind. *Theories of Group Behavior*.
- Weick, Karl E., and Sutcliffe, Kathleen M. (2001). Managing the Unexpected: Assuring High Performance in an Age of Complexity. *Jossey-Bass*.
- Wickens, Christopher D., Hollands, Justin G., Banbury, Simon, and Parasuraman, Raja (2021). Engineering Psychology and Human Performance. *Pearson*.
- Zhang, Shaokun, Yin, Ming, Zhang, Jieyu, Liu, Jiale, Han, Zhiguang, Zhang, Jingyang, Li, Beibin, Wang, Chi, Wang, Huazheng, Chen, Yiran, and Wu, Qingyun (2025). Which Agent Causes Task Failures and When? On Automated Failure Attribution of LLM Multi-Agent Systems. *ICML 2025 Spotlight*. arXiv:2505.00212. The Who&When dataset: 127 multi-agent failure logs with fine-grained agent-and-step annotations, plus three attribution methods (All-

at-Once, Step-by-Step, Binary Search). Best method reaches 53.5% agent accuracy and 14.2% step accuracy. Used as one of the two evaluation benchmarks in the companion empirical paper.