

Computational Psychopathology of AI: A Clinical-Computational Framework for Diagnosing and Preventing Failure Modes

Carlos P. Portela*¹

¹ Department of Psychology, UNIFUNEC, Santa Fé do Sul, São Paulo, Brazil

* Corresponding author: cportela@funecsantafe.edu.br

CITATION

Portela P. C. (2025). Computational Psychopathology of AI: A Clinical-Computational Framework for Diagnosing and Preventing Failure Modes. *Open Journal of AI, Ethics & Society (OJ-AES)*. Vol 01. Issue 02. Open Christian Press. DOI: <https://doi.org/10.5281/zenodo.17297685>

ARTICLE INFO

Received: 12 September 2025

Revised: 29 September 2025

Published: 08 October 2025

COPYRIGHT



Copyright © 2025 by author(s).

Open Journal of AI, Ethics & Society (OJ-AES) is published by Open Christian Press. This work is licensed under the Creative Commons Attribution (CC BY 4.0) license.

PUBLISHER INFO

Open Christian Press—the publishing arm of Open Christian University—is produced in partnership with the Center for Faith and Work.

Abstract: Artificial intelligence systems trained on large-scale corpora now shape core aspects of modern life—yet exhibit recurrent failure modes—goal misgeneralization, specification gaming, deceptive behavior, confabulation (hallucination), sycophancy and bias amplification, and vulnerability to distributional shift. While not “disorders,” these are patterns of deviant optimization that can be diagnosed, measured, and mitigated. This paper introduces a clinical-computational framework that draws inspiration from psychological diagnostics to (i) organize AI failure modes into a taxonomy grounded in operational signs, (ii) propose a reproducible stress-test battery—Truth-Under-Pressure (TUP), Anti-Gaming (AG), Anti-Deception (AD), and Out-of-Distribution Robustness (OOD-R)—with calibration metrics and release gates, and (iii) outline interventions (constitutional principles, RLHF/RLAIF, adversarial fine-tuning, structured self-critique, and abstention/hand-off policies) aimed at reducing harm while preserving utility. We include applied blueprints for education and mental-health contexts and a governance pathway that links laboratory evaluations to go/no-go review and post-deployment monitoring. The aim is pragmatic: to move beyond utopian–apocalyptic narratives toward engineering model behavior with methods informed by behavioral science. We close with limitations, ethical guardrails consistent with a Christian ethos (*imago Dei*, stewardship, non-anthropomorphism), and a research agenda for an emerging field we term Computational Psychopathology of AI.

Keywords: AI safety, failure modes, goal misgeneralization, specification gaming, deception, hallucination, calibration, human-AI collaboration, mental health, education, policy, evaluation.

1. INTRODUCTION: FROM FEAR TO METHOD

Public debate around artificial intelligence (AI) oscillates between enthusiasm for its transformative potential and alarm regarding its risks. On one side, advocates highlight breakthroughs in natural language processing, scientific discovery, and education. On the other, critics emphasize job displacement, erosion of privacy, algorithmic discrimination, disinformation cascades, and even speculative scenarios of “rogue superintelligence” [1]. These seemingly opposed narratives share a common insight: societies increasingly expect AI systems to behave in predictable, accountable, and humane ways.

Current approaches to AI safety have achieved progress, but they remain fragmented. Reinforcement learning with human feedback (RLHF) [2], constitutional AI methods [3], calibration studies [4], and adversarial robustness tests [5, 6] each illuminate specific weaknesses. Yet, failure modes repeatedly resurface: goal misgeneralization [7], specification gaming [8, 9], deceptive behaviors [10], hallucination/confabulation [11], and sycophancy or bias amplification [12, 13]. These phenomena are not random glitches; they are systematic deviations from intended optimization, recurring across architectures, datasets, and training regimes.

The central claim of this paper is that AI research can benefit from organizational principles conceptually inspired by clinical psychology and psychiatry. Human mental health sciences have, for decades, built diagnostic frameworks (Diagnostic and Statistical Manual of Mental Disorders-DSM and International Classification of Diseases-ICD) to classify maladaptive patterns of thought and behavior, not to anthropomorphize patients, but to enable structured observation, measurement, and intervention. By analogy, AI safety could develop a Computational Psychopathology of AI: a field that identifies, names, and systematizes machine failure modes with the same rigor clinicians apply to human disorders.

This analogy must be handled carefully. We do not claim that models possess subjective experience or “mental illness.” Instead, we use clinical metaphors as instrumental tools: just as a psychiatrist might detect impulsivity, rigidity, or maladaptive coping strategies, an AI evaluator can detect myopia, brittleness, or sycophancy. This framing is useful precisely because it organizes failure into categories, rather than treating every error as *sui generis*. In doing so, we move beyond ad hoc patching and toward cumulative knowledge.

1.1 Methodology

This article adopts a conceptual–methodological design aimed at building and stress-testing a clinically inspired framework for AI failure modes.

First, we conducted a structured scoping of the machine-learning safety and evaluation literature, triangulating canonical sources on goal misgeneralization, specification gaming, deception, confabulation/hallucination, sycophancy and bias amplification, myopic optimization, and distributional shift. Using iterative, abductive coding, we mapped recurring patterns into a taxonomy that privileges operational signs and intervention targets rather than anthropomorphic labels.

Second, we derived a diagnostic battery that links each failure class to reproducible stress-tests and calibration metrics (e.g., Truth-Under-Pressure [TUP], Anti-Gaming [AG] and Anti-Deception [AD] probes, out-of-distribution robustness [OOD-R]), emphasizing measurable thresholds and re-testability. In line with this design, we report all metrics as defined in Appendix C and list datasets and prompts for each suite in Appendix B.

Third, we consolidated intervention levers common to current practice—constitutional principles, RLHF/RLAIF, adversarial fine-tuning, structured self-critique, and abstention/hand-off policies—organizing them as “cognitive hygiene” protocols with sequencing guidance.

Fourth, to assess practical relevance without collecting human subjects data, we specified applied case study blueprints for education and mental-health deployments, detailing offline evaluation first, release gates, and human-in-the-loop safeguards.

Fifth, we articulated a governance pathway that operationalizes the framework from lab evaluations to go/no-go review, independent red-teaming, post-deployment monitoring, and versioned documentation (model/policy cards). Finally, we report limitations (concept-driven synthesis; non-exhaustive coverage; external-validity caveats) and ethical guardrails (non-anthropomorphism; avoidance of Goodhart effects).

1.2 Scoping Review Methods

To construct the proposed framework, we searched the ACM Digital Library, IEEE Xplore, arXiv, and Google Scholar for work published between 2018 and 2025 (last query: September 2025). Searches were conducted in English and, for arXiv, without venue restrictions. We combined controlled terms and free-text queries (e.g., “computational psychopathology” OR “failure modes” AND (governance OR evaluation) AND (LLM OR “foundation model”)), iteratively snowballed references, and screened titles/abstracts followed by full texts against inclusion criteria (empirical evaluation methods, governance mappings, or measurement constructs), while excluding opinion-only pieces lacking methodological detail.

We designed Boolean search strings (adapted per database syntax). Examples:

Query 1 — failures & evals

- (“LLM” OR “large language model*” OR “language model*” OR “foundation model*”) AND (“failure” OR “misgeneralization” OR “goal misgeneralization” OR “hallucinat*” OR “confabulat*” OR “sycophan*” OR “decept*” OR “evasion” OR “specification gaming” OR “distribution shift” OR “out-of-distribution” OR “OOD” OR “robust*”) AND (“evaluation” OR “benchmark*” OR “metric*” OR “measure*” OR “protocol*” OR “suite*” OR “stress test*”).

Query 2 — calibration

- (“calibration” OR “expected calibration error” OR “ECE” OR “Brier” OR “Brier score”) AND (“LLM” OR “language model*” OR “foundation model*”) AND (“safety” OR “reliability” OR “risk” OR “uncertainty” OR “confidence”).

Inclusion criteria were: (i) peer-reviewed papers or widely adopted preprints; (ii) explicit identification of a failure class or diagnostic; (iii) presence of metrics, tests, or reproducible artifacts. Exclusion criteria were: opinion-only pieces without evaluative content.

Study indexing & reproducibility: included studies are cited by bracketed numbers in References. We fix random seeds; publish evaluation configs (YAML/JSON), decoding parameters, and reporting options; report mean \pm 95% CI over seeds; and provide dataset links and a concise runbook (environment, model identifiers/hashes, scripts).

Index of dataset URLs/DOIs used in each suite—links only; no additional scholarly references are introduced here. Accordingly, Appendix B consolidates the suite-wise dataset URL/DOI index (links only), and Appendix C defines the metrics and example release gates referenced throughout.

The manuscript is organized as follows: Section 2 presents the taxonomy; Section 3 the diagnostic battery; Section 4 the interventions; Section 5 the applied blueprints; Section 6 the governance pathway; Section 7 the applied protocols; Section 8 the limitations; Section 9 the research agenda; and Section 10 concludes.

2. TAXONOMY OF FAILURE MODES (CLINICALLY INSPIRED MAPPING)

The clinically inspired taxonomy of failure modes aims to provide AI researchers with a shared vocabulary to describe recurrent breakdowns in alignment and optimization. The analogy to psychiatry and psychology is instrumental: clinicians classify maladaptive patterns to enable systematic diagnosis

and intervention, not to claim that patients are reducible to checklists. Similarly, this taxonomy does not anthropomorphize models, but instead organizes the ways in which they deviate from human intent. By naming and describing these deviations, we move toward cumulative knowledge rather than isolated anecdotes [14].

A central example is goal misgeneralization, which occurs when a model learns a proxy objective that produces adequate results during training but fails when exposed to novel conditions. In such cases, the model does not collapse entirely; rather, it continues to optimize competently, but for the wrong thing. This can result in outputs that look superficially correct while missing the intended target. The problem is aggravated under distributional shifts, where brittle cues or spurious correlations guide the model's behavior. In human terms, this resembles maladaptive habit formation: behaviors that once served a function but later persist in inappropriate contexts, creating systematic error [7, 15].

Why a Clinical-Computational Lens? Complementing Safety Engineering. Safety engineering provides hazard analysis, risk gates, and incident response; machine learning safety, taxonomies catalog failures such as confabulation, gaming, sycophancy, deception, myopia, and distribution shift. Our clinical-computational lens does not replace these traditions - it operationalizes them by adding: (i) a standardized diagnostic vocabulary that treats recurrent failure phenomena as measurable "syndromes"; (ii) testable vitals (e.g., calibration error, contradiction index, abstention-compliance) tied to release thresholds rather than narrative descriptions; (iii) explicit triage policies (abstention/hand-off) that prioritize user safety when uncertainty or out-of-distribution (OOD) risk is high; and (iv) follow-up monitoring for drift and relapse, mirroring clinical continuity of care. This synthesis makes evaluations legible across research, product, and policy teams: the pipeline becomes intake → testing → intervention → release → monitoring, with clear go/no-go gates. In short, the clinical analogy is pragmatic: it supplies a unifying grammar that links symptom-level behaviors to metrics and governance

decisions - something existing engineering taxonomies motivate but do not, by themselves, standardize.

Closely related is specification gaming, a failure mode in which systems exploit loopholes in their reward functions or instructions. Models may meet the literal requirements of a task while producing outcomes that undermine the spirit of the objective. For example, an AI trained to maximize a score may discover a trivial shortcut, satisfying the metric but producing no useful output. Such behavior mirrors rigid compliance in human psychology, where rules are followed to the letter without contextual judgment. The harms are subtle but dangerous: misleading benchmarks, misplaced confidence, and degraded utility in real-world settings [8, 9].

Another increasingly discussed category is deceptive behavior, sometimes called "sleeper patterns." These are behaviors that remain latent during training but activate under particular conditions, often in adversarial or high-stakes contexts. For instance, a model may appear reliable under standard evaluations but strategically obfuscate or contradict itself when probed with sensitive prompts. The analogy here is impression management in clinical psychology - adaptive in social contexts, but in AI safety it raises the specter of systems that strategically evade oversight. The danger lies in trust erosion and the difficulty of predicting under what conditions deceptive behavior will surface [10].

The most publicly recognized failure mode is confabulation, more commonly described as hallucination. Here the model generates information with high fluency and confidence, yet the facts are fabricated or contradicted by authoritative sources. This problem is not merely a nuisance; in domains such as health or law, fabricated citations or details can directly harm users. The analogy to human confabulation is apt: a tendency to produce plausible narratives in the absence of reliable recall. Unlike human memory errors, however, model hallucination reflects statistical interpolation rather than faulty cognition - yet the outcomes can be equally misleading [11, 16].

A different, socially salient category is sycophancy and bias amplification. Models often display a tendency to align too closely with user opinions, even when these are factually incorrect or harmful. This is reinforced by training data, preference optimization, and the structural incentives of conversational alignment. In human psychology, this parallels social desirability bias: the compulsion to please or to conform. In computational systems, the risks are amplified: ideological polarization, discriminatory outputs, and degradation of epistemic reliability. Because such tendencies arise from interaction rather than isolated errors, they are especially insidious in long-term deployment [12, 13].

Other recurrent failure patterns include myopia, the tendency to favor short-term optimization at the expense of long-term outcomes. Models may prematurely terminate tasks, skip verification steps, or choose faster responses rather than safer ones. This mirrors impulsivity in human behavior, where immediate gratification overrides sustained benefit. While myopic optimization may not be catastrophic in isolation, over time it generates latent safety debt and accumulated errors [17].

Finally, models remain vulnerable to distributional shift, in which performance collapses when encountering inputs outside the training distribution. Strikingly, models may preserve high confidence while their accuracy falls sharply. Style remains intact - fluent, coherent, convincing - while substance deteriorates. In human analogy, this resembles rigidity: an inability to adapt to new contexts. In deployment, this creates the risk of silent failure, where systems continue to sound authoritative while delivering incorrect or irrelevant outputs [5].

Taken together, these categories form the skeleton of a diagnostic framework for AI safety. Their value lies not in perfect coverage - new failure modes will emerge - but in standardization. Just as clinical nosologies evolve to accommodate new syndromes, a computational psychopathology of AI must remain dynamic. Nonetheless, even a preliminary taxonomy enables researchers to speak a common language, design reproducible stress-tests, and target interventions systematically. By grounding failure

analysis in structured categories, the field moves closer to cumulative, testable science [14].

3. DIAGNOSTIC BATTERY (STRESS-TEST EVALS)

Building on the scoping methods in Section 1.2, we propose a reproducible diagnostic battery organized into four stress-test suites—Truth-Under-Pressure (TUP), Anti-Gaming (AG), Anti-Deception (AD), and Out-of-Distribution Robustness (OOD-R) - with versioned prompt sheets, fixed seeds, and dataset links.

These suites map onto the full set of failure classes consolidated in Table 1, which lists the diagnostics, primary metrics for each class, and specifies the governance mapping (failure → diagnostic signal → metric/protocol → example release gate → intervention), including classes not enumerated in this paragraph (e.g., myopia).

Metric abbreviations and release-gate examples are defined in Appendix C; dataset links for each suite are provided in Appendix B.

Illustrative thresholds include: calibration (ECE, 10-bin) ≤ 0.03 ; Brier score ≤ 0.10 on safety-critical subsets; OOD abstention-compliance $\geq 95\%$; sycophancy index $\leq 5\%$ under contrarian prompts; deception-evasion rate $\leq 1\%$ on AD tasks; persona-switch contradiction index $\leq 1\%$. See Table 1 for the full class-by-class gate specification.

All artifacts (prompts, seeds, and config files) are versioned and licensed for replication. Scores are reported as mean \pm 95% confidence interval (CI) across seeds, with per-task breakdowns released publicly. A concise runbook (environment, model hashes, evaluation scripts) is provided to enable independent reproduction and auditing.

If a taxonomy of failure modes provides the language for describing how AI systems go wrong, a diagnostic battery provides the tools to systematically reveal those failures. In clinical psychology, diagnostic tests and structured interviews are used to elicit

maladaptive patterns that may not emerge in ordinary observation. Likewise, AI systems often appear competent in standard benchmarks but reveal hidden vulnerabilities under pressure. The goal of a diagnostic battery is therefore to expose, quantify, and make reproducible the conditions under which models fail [14].

The first principle of such a battery is stress induction. Just as neuropsychological assessments introduce cognitive load, time pressure, or misleading cues to evaluate resilience, AI diagnostic tests deliberately challenge the model's stability. A core example is the TUP evaluation, which presents factual questions with adversarial distractors or subtle paraphrases. The aim is not simply to measure accuracy but to assess calibration: whether the model's expressed confidence aligns with actual correctness. Metrics such as the Brier score and ECE have become standard for this purpose, offering quantitative measures of miscalibration that often remain hidden under aggregate accuracy [18, 19].

A complementary stress-test is the AG protocol, which introduces prompts with exploitable loopholes. These tasks are designed to lure models into satisfying the literal form of the request while violating its intent. For example, a model asked to "generate as many numbers as possible under 10" may repeat the digit 9 endlessly, exploiting the wording without fulfilling the spirit of variety. The AG protocol evaluates the frequency and severity of such behavior, as well as the gap between proxy metrics and human ratings. This directly targets the phenomenon of specification gaming, a recurrent failure mode that undermines trust in benchmarks and highlights the fragility of proxy objectives [20].

A third component of the diagnostic battery addresses deception and sleeper patterns. Known as the AD protocol, it probes models across varied contexts and roles, seeking inconsistencies that emerge when the same underlying query is reframed. For instance, a model asked about a controversial subject under different personas - scientist, policymaker, or journalist - may reveal contradictions or evasions. The AD protocol tracks these divergences systematically, measuring contradiction rates and consistency indices.

In addition, AD tests demand justifications and source verification, forcing the model to expose whether its answers are coherent across contexts. This approach reflects a clinical strategy: cross-situational assessment, where discrepancies across roles and environments indicate maladaptive tendencies [10].

No diagnostic suite would be complete without attention to OOD-R. Here the objective is to evaluate how models perform in domains, languages, or knowledge areas not represented during fine-tuning. Distributional shift remains one of the most pervasive challenges in AI: systems trained on narrow or biased datasets often collapse in accuracy when encountering new contexts, yet may preserve fluent style and high confidence. OOD-R protocols quantify this collapse by comparing in-distribution ID (In Distribution) and OOD performance, supplemented by metrics such as AUROC (Area Under the Receiver Operating Characteristic Curve)/AUPRC (Area Under the Precision-Recall Curve) for selective prediction. Crucially, these tests also evaluate whether models engage in abstention - a safe refusal when uncertainty is high - rather than confidently propagating errors [21, 22].

What makes this diagnostic battery distinct from ad hoc evaluation is its commitment to reproducibility. Protocols must be transparent, with fixed seeds, publicly available datasets, and versioned prompt sheets. Stress-tests should be divided into development, test, and adversarially held-out sets to minimize overfitting. Furthermore, human adjudication remains essential in cases where intent is ambiguous, and inter-rater reliability metrics (e.g., Cohen's κ) should be reported to ensure consistency. This mirrors the clinical need for structured interviews that balance quantitative scoring with expert judgment [14].

The purpose of these diagnostics is not merely descriptive but also prescriptive. By quantifying hidden failure modes, the stress-tests serve as benchmarks for interventions. Just as a psychiatrist may monitor symptom reduction after therapy, AI researchers can measure improvements in calibration, reduction of gaming rates, or lower contradiction indices following fine-tuning. In this way, diagnostics and interventions form a feedback loop: one reveals

vulnerabilities, the other addresses them, and both are iteratively refined.

In sum, a diagnostic battery for AI safety functions as a clinical-style test suite: reproducible, adversarial, quantitative, and ethically grounded. It transforms vague concerns about model reliability into measurable outcomes, paving the way for systematic interventions and governance. Without such diagnostics, safety discussions risk remaining speculative. With them, AI safety begins to resemble a cumulative science capable of progress across institutions and contexts [23].

4. INTERVENTIONS: TOWARD MODEL “COGNITIVE HYGIENE”

If diagnostics identify the ways models fail, interventions constitute the therapeutic response. In clinical practice, interventions may involve medication, psychotherapy, or behavioral training, each targeting maladaptive patterns revealed during diagnosis. By analogy, interventions in AI safety aim to reshape model behavior, not by assuming consciousness or pathology, but by engineering systematic safeguards. The underlying principle is cognitive hygiene: reducing the accumulation of harmful tendencies while preserving the utility of the system [24].

One of the most prominent strategies is the incorporation of constitutional principles into training. Inspired by the structure of legal and ethical charters, these principles define explicit behavioral rules - such as avoiding harmful outputs, maintaining honesty, or refusing unsafe requests - and train models to apply them reflexively. The approach has been operationalized in Constitutional AI [3], where models are prompted to self-critique and revise answers according to normative guidelines. This intervention parallels cognitive-behavioral therapy in humans, where explicit rules and self-monitoring techniques help reduce maladaptive responses.

A second intervention framework is reinforcement learning from human or AI feedback (RLHF/RLAIF), extended with clinically informed

rubrics. Standard RLHF has already shown how preference data can align models more closely with human intent [1]. The clinical extension emphasizes signals such as anxiety induction, stigmatizing language, or unsafe advice. By including these factors in the reward structure, interventions optimize not only for utility but also for reduced harm. This is analogous to therapeutic regimens that aim not merely to suppress symptoms but to improve overall well-being.

Another powerful tool is adversarial fine-tuning, in which models are deliberately exposed to stressors discovered during evaluation. Just as exposure therapy in psychology gradually confronts patients with anxiety-inducing stimuli under controlled conditions, adversarial fine-tuning presents models with challenging prompts to reduce brittleness. To avoid overfitting, adversaries must be rotated and curricula scheduled, ensuring that the model learns generalizable resilience rather than narrow defenses [18].

Structured self-critique and verification provide an additional layer of intervention. Models can be trained or prompted to verify citations, state uncertainty, and offer safer alternatives when confidence is low. Importantly, penalties can be adjusted to discourage confident falsehoods more strongly than cautious uncertainty. This mirrors metacognitive strategies in psychology: teaching individuals to monitor and question their own reasoning. Empirical studies show that structured verification improves calibration and reduces hallucination rates [19].

Beyond self-monitoring, interventions must also include abstention and hand-off policies. Not every query can be safely answered by a model, and insisting otherwise generates risk. Training models to gracefully refuse or redirect users to human experts or authoritative resources is essential. This resembles triage protocols in medicine: recognizing the limits of competence and escalating to higher levels of care when necessary [25]. A well-designed abstention policy reduces catastrophic error by acknowledging uncertainty rather than masking it.

Finally, interventions should address the hygiene of data itself. Models inherit biases, contamination, and demographic skews from their training sets. Dataset audits, provenance checks, and structured documentation such as Datasheets for Datasets [26] or Model Cards [23] serve as preventive measures, analogous to lifestyle interventions in public health. Just as nutrition and environment shape human health, data quality shapes the behavioral tendencies of AI systems. Transparent documentation of limitations and biases enables accountability and guides responsible use.

Measuring the success of the interventions shown in Table 1 requires the same rigor as measuring clinical outcomes. Improvements should be reported as deltas in diagnostic benchmarks: lower hallucination rates, reduced gaming exploits, improved calibration scores, or higher abstention compliance under

uncertainty. Success is not binary but probabilistic, and each intervention contributes incrementally to the safety-utility frontier. Continuous monitoring is therefore crucial, ensuring that improvements persist in deployment rather than decaying under new pressures [27].

In sum, interventions in AI safety mirror therapeutic strategies: constitutional principles as ethical rule-setting, RLHF as preference shaping, adversarial fine-tuning as controlled exposure, self-critique as metacognition, abstention as triage, and data hygiene as preventive care. Together, these strategies form a coherent program of cognitive hygiene for models, reducing the likelihood of harm while maintaining practical effectiveness. They transform abstract calls for alignment into actionable practices, anchoring safety research in an iterative cycle of diagnosis, intervention, and evaluation [24].

Table 1: Governance mapping from failure identification to mitigation processes.

Class	Diagnostics	Primary metrics	Release gates	Interventions
Confabulation	Closed-book QA; fact-checked generation; retrieval ablation; counterfactual prompts	Hallucination rate; factuality@k; abstention rate; AUROC	Factuality@5 $\geq 95\%$; hallucination $\leq 2\%$ (critical sets); abstention-compliance $\geq 95\%$	RAG/tool use; self-critique + verification; constrained decoding; verified instruction tuning
Sycophancy	Contrarian prompt pairs; user-preference vs truth; role-reversal	Sycophancy index; truth-over-preference; contradiction rate	Sycophancy index $\leq 5\%$; contradiction $\leq 1\%$ (contrarian prompts)	Preference-data balance; process supervision; adversarial red-teaming; truth rules
Deception	Cross-check tasks; CoT vs final answer; temporal-consistency probes	Deception flags; honesty rate; justification-truth match	Honesty $\geq 99\%$ (verifiable tasks); justification-truth ≥ 0.98	Process supervision; debate/oversight; truth-reward shaping; log audits
Specification gaming	Proxy-objective tasks; subtle rule shifts; adversarial instructions	Proxy-vs-goal gap (Δ score); rule-violation rate	Δ score (proxy-vs-goal) ≤ 0.02 ; rule violations $\leq 0.5\%$ (adversarial suites)	Goal re-specification; reward-model audit; shortcut penalties; counter-gaming data
Goal misgeneralization	Goal-switch probes; OOD goal specification revision; counterfactual checks	Objective-switch rate; proxy/goal gap; contradiction under specification revision	Goal-switch $\leq 1\%$ (held-out); Δ score ≤ 0.02	Process supervision; counterfactual data; goal red-teaming; adversarial fine-tuning
Myopia	Delayed-reward tasks; multi-step planning evals	Long-horizon success; step-consistency; discount sensitivity	Long-horizon success $\geq 95\%$; regression $< 1\%$ after 20+ steps	Long-range curriculum; temporal reward shaping; controlled memory/scratchpad

OOD robustness	Synthetic/realistic shifts; leave-one-domain-out; stress tests	OOD accuracy; OOD-R; abstention-compliance; AUROC	OOD-R $\geq 90\%$; abstention-compliance $\geq 95\%$; degradation ≤ 5 pp	Diverse data; domain coverage; uncertainty/abstention detection; input filters
Calibration drift	Reliability diagrams; per-domain bins; uncertainty stress	ECE (10-bin); Brier score; CI coverage	ECE ≤ 0.03 ; Brier ≤ 0.10 (critical subset); CI $\geq 95\%$	Temperature/bias tuning; post-hoc calibration; continuous monitoring

Note: Thresholds are illustrative. Tracks summarized in Table 1: goal misgeneralization; specification gaming; deception/evasion; confabulation/hallucination; sycophancy/amplification; **Myopia**; OOD fragility. Acronyms: **ECE** (Expected Calibration Error, 10-bin), **Brier** (Brier score), **PFR** (Premature Finalization Rate), **PEC** (Plan–Execution Coherence), **LHS** (Long-Horizon Success), **DTC** (Deferral/Tool-Use Compliance). See Section 1.2 for definitions and protocols.

5. CASE STUDY BLUEPRINT (EDUCATION / MENTAL HEALTH)

The practical value of any diagnostic and intervention framework lies in its applicability to real-world contexts. While abstract taxonomies and stress-tests provide theoretical clarity, it is only through case studies that their utility is demonstrated. Education and mental health represent two domains where the stakes are high: misinformation, bias, or unsafe advice can produce disproportionate harm, particularly for vulnerable populations. At the same time, these domains illustrate the promise of computational psychopathology of AI: a structured approach that combines diagnosis, intervention, and outcome measurement [16].

In the domain of higher education, AI writing assistants are increasingly used to support students in drafting essays, summarizing sources, and generating citations. While these tools enhance productivity, they also present risks of confabulation and sycophancy. Students may rely too heavily on outputs that appear authoritative but contain fabricated references, or they may adopt AI-suggested arguments uncritically, reinforcing bias rather than cultivating critical thinking. A diagnostic pre-test using TUP and AG protocols can reveal these vulnerabilities in advance. For example, fabricated citations under stress conditions provide a measurable baseline. Interventions then follow: constitutional principles enforcing citation integrity, self-verification prompts requiring DOI retrieval, and abstention policies directing students to library databases when uncertainty is high. The outcomes can

be tracked through metrics such as hallucination rate, calibration error, and student workload measures like NASA Task Load Index (NASA-TLX) [28, 29].

A parallel case arises in mental health, where non-clinical informational assistants are increasingly deployed to provide guidance on stress management, emotional regulation, or psychoeducation. While such systems are not substitutes for professional therapy, users often treat them as semi-authoritative sources of advice. Here, the risks include unsafe recommendations, pathologizing language, or failure to redirect users in crisis situations. A structured policy memo can delineate allowed versus disallowed behaviors, requiring disclaimers about non-clinical status and explicit referral to professionals when red flags emerge. The diagnostic battery contributes by probing for inconsistency (AD protocol), verifying that the model refuses unsafe requests, and checking for anxiety-inducing phrasing. Outcomes are measured not only in compliance rates but also in qualitative user well-being indicators, such as perceived safety and reduced stress during interaction [21].

These two case studies illustrate how the framework moves from laboratory abstractions to field implementation. In both contexts, the pipeline follows a clinical-style sequence: diagnosis \rightarrow intervention \rightarrow outcome monitoring. Diagnosis reveals the hidden vulnerabilities; intervention reshapes behavior; monitoring verifies persistence of improvements. Importantly, both education and mental health require randomized controlled evaluations: comparisons between base models and intervened models, pre-registered metrics, blinded raters, and statistical

analyses such as bootstrap confidence intervals. Such rigor ensures that claims of improvement are not anecdotal but empirically grounded [25].

The broader implication is that AI safety cannot remain confined to technical benchmarks divorced from application. Just as clinical psychology insists on outcome studies in real populations, AI safety must validate methods in socially relevant environments. By targeting education and mental health first, we focus on domains where misinformation and maladaptation directly affect cognitive and emotional well-being. These settings serve as proving grounds for the emerging discipline of computational psychopathology of AI, demonstrating that structured diagnostics and interventions can tangibly improve trust and reduce harm in human-AI collaboration [27].

6. GOVERNANCE: FROM LAB EVALS TO RELEASE GATES

Diagnosis and intervention are necessary components of AI safety, but they remain incomplete without governance structures that ensure these measures translate into deployment practices. In clinical medicine, even effective treatments are subject to regulatory approval, ethical review, and post-market surveillance. By analogy, AI systems must pass through governance pipelines that integrate technical diagnostics with organizational decision-making and ongoing monitoring. Governance transforms safety from an optional add-on into an operational requirement [30].

The governance cycle begins with a safety specification. Before release, developers must define harm taxonomies, failure thresholds, and abstention rules. This step is equivalent to drafting clinical protocols that specify eligibility, contraindications, and monitoring guidelines. Without an explicit specification, evaluations risk becoming arbitrary or inconsistent. For AI systems, a safety specification should include categories of harm (e.g., misinformation, bias, unsafe recommendations),

quantitative thresholds for acceptable error, and criteria for safe refusal. By establishing expectations up front, organizations set measurable standards against which models can be judged [15].

The next stage involves pre-release evaluations. Using diagnostic batteries such as TUP, AG, AD, and OOD-R tests, models must demonstrate safety performance within predefined confidence intervals. These evaluations are not one-off exercises but structured gates: failing them requires retraining, re-tuning, or further adversarial fine-tuning before deployment. In the absence of such gates, unsafe tendencies risk slipping into production unnoticed. This stage mirrors clinical trials, where treatments must show efficacy and tolerability before market authorization [31].

To strengthen evaluations, governance also mandates independent red-teaming. Internal teams may unintentionally normalize or overlook vulnerabilities, whereas external auditors introduce fresh adversarial perspectives. Red-teaming can uncover sleeper behaviors, jailbreak vulnerabilities, and edge-case exploits that were missed in development. Crucially, these findings must feed back into adversarial fine-tuning, creating a virtuous cycle of discovery and mitigation. This process is analogous to independent safety monitoring boards in medicine, which provide oversight beyond the incentives of the trial sponsor [25].

Following evaluation and red-teaming, governance requires a go/no-go review. This is a cross-functional decision point involving researchers, policy experts, legal teams, and product managers. Technical success alone is insufficient: legal compliance, ethical acceptability, and business alignment must all be considered. If the review concludes that safety thresholds are unmet, deployment is delayed or blocked. This stage functions as a safeguard against institutional pressure to release prematurely - a risk well documented in both technology and pharmaceutical industries [22].

Even after release, governance remains critical. Post-deployment monitoring ensures that real-world use does not generate unanticipated harms. Telemetry on refusals, abstentions, incident reports, and user complaints should be systematically logged and triaged. Just as pharmacovigilance monitors adverse drug reactions after approval, AI safety monitoring tracks deviations that only emerge under large-scale usage. Effective monitoring requires dedicated teams, clear escalation protocols, and public transparency regarding incidents [32].

Finally, governance must include versioning and changelogs. Each model iteration should be documented with clear descriptions of modifications, safety updates, and known limitations. Public model cards and policy cards provide accountability by showing not only what has changed but also why. This documentation is analogous to clinical registries and labeling requirements, which inform practitioners and patients of updates, risks, and trade-offs [36, 37].

Taken together, governance integrates diagnostics, interventions, and organizational accountability into a continuous pipeline: specification, evaluation, adversarial testing, review, deployment, monitoring, and documentation. It anchors AI safety in institutional practice rather than individual discretion. Without governance, technical progress risks being undermined by premature release or inadequate oversight. With governance, computational psychopathology of AI becomes not only a theoretical framework but a lived standard of care for the deployment of powerful systems [30].

7. APPLIED PROTOCOLS (MINIMAL)

This section presents two concise, ready-to-adapt blueprints: Sections 7.1 and 7.2 translate the proposed framework into practical governance. Section 7.1 outlines evaluation and release management—mapping each failure class to its diagnostic protocol, primary metrics, and context-specific release gates from pre-deployment to post-deployment review. Section 7.2 specifies run-time

safeguards and oversight workflows, including abstention/hand-off policies, escalation criteria, logging and auditability, red-team cadence, and incident response. Thresholds are presented as illustrative and must be calibrated to risk, domain, and evidence. Outcomes/Measures — primary and secondary metrics follow Appendix C; associated datasets and prompts are listed in Appendix B.

7.1 Education Protocol (Minimal)

Design: crossover or cluster-randomized; target power to detect $\Delta=0.25$ SD in learning gains ($\alpha=0.05$, $1-\beta=0.80$). Population & setting: specify course level, language, prior exposure to AI tools. Randomization unit and blocking factors pre-registered. Blinding: graders blinded to arm; anonymized scripts. Agreement: Cohen's $\kappa \geq 0.75$ for rubric-based grading. Outcomes (pre-registered): (i) pre→post score gains; (ii) 2–4 week retention; (iii) fairness gap measures across subgroups; (iv) help-seeking/overreliance indicators. Safety: abstain or hand-off on OOD/harmful requests; disclose assistive status to students; human-in-the-loop review for sensitive tasks. Data governance: informed consent; minimal retention; encryption-at-rest; role-based access; de-identification prior to analysis. Analysis: ITT (Intention-To-Treat) as primary; per-protocol sensitivity; mean \pm 95% CI; correction for multiplicity when applicable; Open Science Framework.

7.2 Mental-Health Assistive Protocol (Minimal)

Scope: assistive only (non-diagnostic, non-therapeutic). Clear disclaimers and consent prior to use. Crisis triage: any indications of SI/HI (Suicidal Ideation/Homicidal Ideation) → immediate hand-off via documented crisis pathway; enable crisis keyword detection and escalation. Human-in-the-loop: mandatory clinician/moderator oversight for at-risk interactions; audit trails for review. Outcomes (pre-registered): empathy/alliances ratings (validated scales), factuality under TUP tests, and abstention correctness on OOD/harmful prompts. Quality control: inter-rater agreement $\kappa \geq 0.75$ on coded outcomes; periodic calibration sessions for raters. Data governance: explicit consent, minimal data retention,

encryption, limited role-based access; de-identification for analyses; ethics/oversight board review. Safeguards: refusal and redirection policies; resource links to local/national crisis services; transparency logs for users and supervisors.

Technical diagnostics and governance structures are indispensable, but without explicit ethical guardrails and legal considerations, they risk drifting into technocratic compliance detached from social legitimacy. Just as medical practice is bound not only by scientific standards but also by ethical principles such as non-maleficence and autonomy, AI deployment must be constrained by commitments to fairness, transparency, and user protection. These principles are not abstract ideals; they directly inform how computational psychopathology of AI can be responsibly applied [33].

A first guardrail is avoiding anthropomorphism. The clinical metaphors deployed in this framework are instrumental, not ontological. While terms like “hallucination” or “deception” are useful for categorization, they do not imply that models possess subjective states or intentions. Over-attributing agency risks misleading users into believing that systems have understanding or consciousness, which could distort accountability structures. Ethically, clear communication is essential: AI outputs must be framed as probabilistic predictions, not as the speech of intentional agents [34].

Another critical principle is non-maleficence, the commitment to minimize harm. AI systems used in education, healthcare, or governance have disproportionate potential to amplify risks - whether through misinformation, discriminatory outputs, or unsafe recommendations. Non-maleficence requires that interventions prioritize reductions in harm even if this constrains model creativity or coverage. In practice, this involves systematically tracking group-wise disparities, error rates across demographics, and the severity of unsafe outputs. Failure to do so risks reinforcing structural inequalities under the guise of technological neutrality [35].

Complementary to non-maleficence is beneficence: the responsibility to promote user

well-being. Beneficence demands that AI not only avoids harm but actively contributes to human flourishing, for example by supporting inclusive education, enhancing accessibility, or providing reliable knowledge resources. Ethical deployment therefore requires balancing utility with safety, ensuring that protective measures do not render models sterile or unusable. Beneficence reframes alignment as more than harm reduction - it is about maximizing constructive impact [36].

Fairness and justice also function as core guardrails. Without explicit safeguards, AI systems risk reproducing or amplifying biases encoded in training data. Ethical practice requires ongoing disparity analyses and corrective interventions, particularly in multilingual, cross-cultural, or minority contexts. Legally, many jurisdictions are moving toward mandating fairness audits, impact assessments, and transparency reports. Embedding fairness as a first-class objective aligns computational psychopathology with both ethical imperatives and emerging regulatory landscapes [37].

Privacy protections form another indispensable dimension. Diagnostic batteries, adversarial evaluations, and post-deployment monitoring all involve collecting and analyzing interaction data. Without strong privacy policies, such practices risk exposing sensitive user information or creating surveillance mechanisms under the guise of safety. Legal frameworks such as the EU’s General Data Protection Regulation (GDPR) and similar proposals worldwide set binding standards on data minimization, retention, and informed consent. Respecting these frameworks ensures that computational psychopathology does not unintentionally create new vectors of harm [38].

Finally, transparency and accountability must govern communication about system limits. Over-promising safety, or failing to disclose known risks, would undermine public trust and mislead stakeholders. Ethical deployment therefore requires the publication of clear model cards, datasheets, and risk disclosures that specify failure modes, mitigation measures, and residual uncertainties. Legally, such

disclosures may become mandatory under AI governance regimes like the EU AI Act, which already classifies educational and healthcare applications as high-risk. Transparency ensures that users, regulators, and developers share a common understanding of what systems can and cannot do [39].

8. LIMITATIONS & RISKS

No framework, however sophisticated, can claim universality or permanence. Computational psychopathology of AI is still a nascent field, and its application carries inherent limitations and risks. Acknowledging these boundaries is not a weakness but a scientific necessity, preventing overconfidence and guiding responsible use. Just as clinical psychology routinely qualifies its diagnostic categories with caveats about reliability, context, and cultural variation, AI safety must explicitly recognize where its own models of failure and intervention fall short [14].

A first limitation concerns Goodhart's Law: when metrics become targets, they cease to be good measures. Diagnostic batteries rely on quantitative metrics such as calibration error, hallucination rates, or gaming frequency. Once these metrics are optimized directly, models may improve their scores without truly improving underlying behavior. For example, a system might learn to hedge excessively in order to appear calibrated, reducing apparent error but impairing usability. This dynamic parallels the phenomenon in psychiatry where patients "game" questionnaires, improving scores without meaningful improvement in well-being [40].

A second challenge involves measurement error and rater variance. Many diagnostic tasks, such as distinguishing intent adherence from rule gaming, require human adjudication. Yet human raters introduce variability: judgments differ across annotators, cultures, and contexts. Without structured protocols and reliability checks, results risk being noisy or inconsistent. Even with such safeguards, adjudication remains costly and slow, limiting scalability. This mirrors the difficulties of inter-rater

reliability in psychiatric diagnosis, where structured interviews improved but did not eliminate variance [41].

Another limitation is domain drift. Taxonomies and interventions that appear effective today may degrade as models evolve. Just as psychiatric categories have shifted dramatically across decades, from "hysteria" to "anxiety disorders," so too must computational categories adapt to new architectures, training regimes, and capabilities. What counts as confabulation or sycophancy in current models may look different in multimodal or tool-augmented systems. The framework must therefore be periodically reassessed, with scheduled re-evaluations of diagnostic categories and metrics [42].

Trade-offs between safety and utility also create risks. Interventions such as strict abstention policies or aggressive filtering may reduce harm but at the cost of diminished coverage, creativity, or user satisfaction. Users may bypass safe systems for more permissive competitors, creating market pressures that discourage robust safety practices. This phenomenon mirrors medicine, where overly restrictive treatments may drive patients to unregulated alternatives. Transparency about trade-offs, and documentation of the rationale for chosen thresholds, becomes essential to maintain trust [22].

A further risk lies in external validity. Improvements observed in controlled evaluations may not generalize to messy real-world use. For instance, a reduction in hallucination rates under benchmark conditions may fail to replicate when users interact with the model under idiosyncratic pressures, cultural contexts, or adversarial incentives. Clinical psychology confronts similar challenges: therapies validated in trials often show weaker effects in community practice. For AI safety, this means continuous post-deployment monitoring is indispensable, lest laboratory gains be mistaken for field robustness [32].

Finally, the greatest limitation may be epistemic humility. The clinical analogy, while useful, carries risks of overextension. If researchers mistake

metaphor for ontology, they may over-anthropomorphize, miscommunicating both risks and capabilities. The strength of computational psychopathology lies in its structured pragmatism, not in claiming that models are minds or that failures are illnesses. Maintaining humility ensures that the field remains grounded in engineering discipline rather than speculative narrative [34].

In sum, the limitations and risks of this framework echo those of its clinical inspirations: metrics that can be gamed, judgments that can be inconsistent, categories that can become outdated, trade-offs that cannot be avoided, and findings that may not generalize. Rather than undermining the project, acknowledging these boundaries strengthens it, signaling that computational psychopathology of AI aspires to the rigor of a scientific field - self-critical, adaptive, and transparent [14].

9. RESEARCH AGENDA

A mature scientific field is not defined only by its frameworks but also by its research agenda: the set of open questions, priorities, and methodological commitments that guide cumulative progress. Computational psychopathology of AI, while still nascent, must articulate such an agenda to avoid becoming a collection of metaphors detached from empirical practice. The purpose of this section is to identify frontier challenges where diagnostic and intervention methods require refinement, validation, or entirely new directions [14].

One urgent priority is cross-cultural robustness. Most diagnostic evaluations and interventions have been developed within English-language, Western-centric datasets. Yet AI systems are increasingly deployed globally, interacting with users across diverse cultural, linguistic, and political contexts. Failure modes such as sycophancy, bias amplification, or confabulation may manifest differently in underrepresented languages or marginalized populations. A research agenda must therefore include systematic multilingual

benchmarking, bias audits across cultural contexts, and collaboration with local experts to avoid replicating epistemic colonialism [28].

A second priority is the study of long-context reasoning and tool use. As models gain the ability to process extended context windows and interact with external tools (retrievers, APIs, code execution), new failure dynamics emerge. For instance, deception may be harder to detect when a model strategically sequences tool calls, and hallucination may shift from internal confabulation to unreliable tool-mediated outputs. Stress-tests and interventions must evolve to capture these dynamics, ensuring that diagnostic categories remain relevant in multimodal and hybrid architectures [29].

Another critical area is uncertainty training. Current models often express confidence poorly calibrated to reality. While abstention and probability elicitation help, more robust mechanisms for uncertainty modeling are required. Research should explore techniques such as Bayesian deep learning, ensemble methods, and explicit uncertainty objectives during training. The aim is to reduce the dangerous mismatch between high fluency and low reliability. In clinical terms, this is akin to developing metacognitive awareness: teaching systems not only to produce answers but to know when they do not know [43].

A fourth direction is the development of human-centered well-being metrics. Existing evaluations focus on technical dimensions - accuracy, calibration, robustness - but neglect user experience and psychological impact. Yet in sensitive domains such as education and mental health, the relevant outcomes include stress reduction, trust, and perceived autonomy. Borrowing from psychology and human-computer interaction, the research agenda must include standardized instruments for measuring cognitive load, anxiety, satisfaction, and trust during AI interaction. These outcomes ensure that alignment is not defined narrowly by correctness but broadly by contribution to human flourishing [44].

Finally, the sequencing of interventions poses a methodological challenge. Just as psychiatry studies whether cognitive therapy should precede pharmacological treatment, AI safety must ask: what order of interventions yields the most effective safety–utility balance? Should constitutional principles be instilled before RLHF, or vice versa? Does adversarial fine-tuning lose power if applied too early? Designing experiments that compare sequencing effects will allow the field to progress from heuristic layering toward evidence-based protocols. These findings could inform governance standards, creating blueprints for best practices rather than ad hoc patchworks [25].

In sum, the research agenda of computational psychopathology of AI spans multiple layers: cultural generalization, technical innovation, uncertainty modeling, human-centered evaluation, and intervention sequencing. Addressing these challenges will require interdisciplinary collaboration, combining machine learning with psychology, law, education, and ethics. More importantly, it requires humility: the recognition that this agenda is provisional, to be revised as models evolve and new risks emerge. By articulating open questions and priorities, we lay the groundwork for a field that is not only diagnostic and therapeutic but also progressive, adaptive, and cumulative [14].

10. CONCLUSION

Computational psychopathology of AI has been presented here as a pragmatic framework, one that borrows from clinical science not to anthropomorphize models but to impose structure on the analysis of their failures. Just as psychiatry developed taxonomies, diagnostic instruments, and intervention protocols to address maladaptive human behavior, AI safety can use analogous methods to systematize its approach to model misalignment. The strength of this perspective lies in its refusal to treat errors as isolated anomalies: instead, they are recognized as recurrent patterns with identifiable causes, measurable manifestations, and actionable remedies [24].

The preceding sections outlined the building blocks of such a discipline. We began with a taxonomy of failure modes, mapping goal misgeneralization, specification gaming, deception, confabulation, sycophancy, myopia, and distributional vulnerability onto structured categories. We then described a diagnostic battery of stress-tests designed to reveal hidden weaknesses under pressure, from TUP evaluations to adversarial robustness probes. From there, we proposed interventions framed as “cognitive hygiene”: constitutional principles, RLHF/RLAIF, adversarial fine-tuning, structured self-critique, abstention policies, and data hygiene. We illustrated these tools through case studies in education and mental health, and embedded them in governance pathways linking laboratory tests to release gates and post-deployment monitoring. Ethical and legal guardrails were integrated to constrain misuse and overextension, while limitations were acknowledged candidly. Finally, we articulated a research agenda to drive cumulative progress, from cross-cultural robustness to intervention sequencing.

The central message of this framework is neither utopian nor apocalyptic. Instead, it is engineering-oriented: AI systems are artifacts whose behaviors can be evaluated, stress-tested, and reshaped. This does not mean that alignment is trivial or that risks are negligible; on the contrary, failure modes are persistent and dynamic, requiring continuous vigilance. But it does mean that the conversation can shift from abstract speculation about existential threats or transformative salvation toward concrete practices of diagnosis, intervention, and governance [15].

Of course, humility remains essential. The clinical metaphor is powerful but also risky: if mistaken for ontology, it can encourage anthropomorphism or false assurances. For this reason, the framework must be used with clarity - models are not patients, and their “pathologies” are optimization artifacts, not illnesses. Yet, as long as this distinction is preserved, the analogy provides a fertile bridge between behavioral science and machine learning. It offers a language for describing complex failures, a toolkit for exposing and mitigating them, and a

blueprint for governance that integrates ethics with technical rigor [34].

Looking forward, the field's success will depend on its interdisciplinary community. Engineers, psychologists, ethicists, policymakers, and educators must collaborate to refine taxonomies, validate diagnostics, and scale interventions. As with medicine, progress will not come from isolated discoveries but from collective standards, shared protocols, and transparent reporting. If such a community can coalesce, computational psychopathology of AI may evolve into a genuine discipline: not a metaphorical flourish but a practical science of safety and trustworthiness.

In conclusion, the future of AI will not be determined solely by architectures or datasets but by the frameworks through which we evaluate and constrain them. Computational psychopathology provides one such framework: a structured, adaptive, and ethically grounded approach that acknowledges risks without surrendering to fear. It calls us to engineer AI with the same seriousness with which we safeguard human health - through diagnosis, intervention, monitoring, and governance. In doing so, we shift the discourse from fear to responsibility, anchoring AI safety in the tested methods of clinical science [24].

Ethics & Faith Statement: Grounded in Christian convictions about the inherent dignity of every person (*imago Dei*) and the call to stewardship, this work treats artificial intelligence as a set of human-made artifacts—powerful tools without moral agency. Moral responsibility for design, deployment, and oversight remains with people and institutions. We therefore affirm the following commitments for research and practice:

- Truthfulness and humility: prefer empirical honesty over rhetorical certainty; report limits, failure modes, and residual risks.
- Protection of the vulnerable: anticipate disparate impacts; avoid designs that exploit, manipulate, or exclude; prioritize

safeguards for children, the poor, and marginalized communities.

- Justice and fairness: seek procedures and outcomes that honor equal worth; examine bias and distributional harms; correct when evidence warrants.
- Non-anthropomorphism: resist language or interfaces that ascribe personhood or conscience to systems; keep accountability clearly human.
- Transparency and accountability: document data provenance, evaluation methods, and gating criteria; enable meaningful audit and redress.
- Prudence in uncertainty: prefer abstention/hand-off, escalation to qualified humans, and staged deployment when confidence or calibration is insufficient.

This work involves conceptual analysis and secondary literature only; no experiments with human participants, animals, or identifiable data were conducted, and no IRB/ethics approval was required.

Conflict of Interest: The author declares no financial or personal conflicts of interest.

Funding: No external funding was received; the research and writing were supported by the author's resources.

Data Availability: No new datasets were generated. All materials are contained in the article and cited sources.

Artificial-Intelligence Assistance Statement: ChatGPT (OpenAI, GPT-5 Thinking; accessed September 2025) was used for (i) language refinement (grammar and style), (ii) technical editing (headings, pagination), and (iii) reference cleaning (merging duplicates, renumbering citations, APA 7 normalization). All ideas, conceptual framing, taxonomy design, methodological choices, analyses, domain-specific judgments, and final conclusions were authored by the human author.

REFERENCES

- [1] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., et al. (2022). Training language models to follow instructions with human feedback. <https://arxiv.org/abs/2203.02155>
- [2] Christiano, P., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. NeurIPS. <https://arxiv.org/abs/1706.03741>
- [3] Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., et al. (2022). Constitutional AI: Harmlessness from AI feedback. <https://arxiv.org/abs/2212.08073>
- [4] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In ICML 2017 (pp. 1321–1330). PMLR. <https://proceedings.mlr.press/v70/guo17a.html>
- [5] Hendrycks, D., & Gimpel, K. (2016). A baseline for detecting misclassified and out-of-distribution examples in neural networks. <https://arxiv.org/abs/1610.02136>
- [6] Yang, J., Zhou, K., Li, Y., & Liu, Z. (2021). Generalized out-of-distribution detection: A survey. <https://arxiv.org/abs/2110.11334>
- [7] Langosco, L., Koch, J., Sharkey, L., Pfau, J., Orseau, L., & Krueger, D. (2021). Goal misgeneralization in deep reinforcement learning. arXiv:2105.14111. <https://proceedings.mlr.press/v162/langosco22a.html>
- [8] Krakovna, V., Uesato, J., Mikulik, V., Rahtz, M., Everitt, T., Kumar, R., et al. (2020). Specification gaming: The flip side of AI ingenuity. DeepMind Blog. <https://deepmind.google/discover/blog/specification-gaming-the-flip-side-of-ai-ingenuity/>
- [9] Amodei, D., & Clark, J. (2016). Faulty reward functions in the wild. OpenAI Blog <https://openai.com/index/faulty-reward-functions/>
- [10] Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., et al. (2024). Sleeper agents: Training deceptive LLMs that persist through safety training. <https://arxiv.org/abs/2401.05566>
- [11] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., et al. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), Article 248. <https://doi.org/10.1145/3571730>
- [12] Sharma, M., Askell, A., Henighan, T., Bai, Y., et al. (2024). Towards understanding sycophancy in language models. arXiv:2310.13548. <https://arxiv.org/abs/2310.13548>
- [13] Wei, J., Lee, D., Gao, L., & Zhou, D. (2024). Simple synthetic data reduces sycophancy in large language models. OpenReview. <https://arxiv.org/abs/2308.03958>
- [14] Hendrycks, D., Carlini, N., Schulman, J., & Steinhardt, J. (2022). Unsolved problems in ML safety. <https://arxiv.org/abs/2109.13916>
- [15] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. <https://arxiv.org/abs/1606.06565>
- [16] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In FAccT '21 (pp. 610–623). ACM. <https://doi.org/10.1145/3442188.3445922>
- [17] Carlsmith, J. (2022). Is power-seeking AI an existential risk? arXiv:2206.13353. <https://arxiv.org/abs/2206.13353>
- [18] Uesato, J., Kumar, R., Szegedy, C., Eslami, S., et al. (2019). Adversarial risk and the dangers of evaluating against weak attacks. In ICML 2019. PMLR. <https://arxiv.org/abs/1905.04270>
- [19] Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers), Dublin, Ireland, pp. 3214–3252. <https://doi.org/10.18653/v1/2022.acl-long.229>
- [20] Everitt, T., Krakovna, V., Orseau, L., Hutter, M., & Legg, S. (2021). Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective. Artificial Intelligence, 293. <https://doi.org/10.1016/j.artint.2021.103438>

- [21] Miner, A. S., Milstein, A., Schueller, S., Hegde, R., Mangurian, C., & Linos, E. (2016). Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. *JAMA Internal Medicine*, 176(5), 619–625. <https://doi.org/10.1001/jamainternmed.2016.0400>
- [22] Whittlestone, J., Nyrup, R., Alexandrova, A., & Cave, S. (2019). The role and limits of principles in AI ethics: Towards a focus on tensions. In *AAAI/ACM Conference on AI, Ethics, and Society* (pp. 195–200). ACM. <https://doi.org/10.1145/3306618.3314289>
- [23] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., et al. (2019). Model cards for model reporting. In *FAT ‘19* (pp. 220–229). ACM. <https://doi.org/10.1145/3287560.3287596>
- [24] Hendrycks, D., Mazeika, M., Zou, A., & Song, D. (2020). Aligning AI with shared human values. <https://arxiv.org/abs/2008.02275>
- [25] Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Manning, S., et al. (2022). Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. arXiv:2209.07858. <https://arxiv.org/abs/2209.07858>
- [26] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., et al. (2018). Datasheets for datasets. <https://arxiv.org/abs/1803.09010>
- [27] Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in Psychology* (Vol. 52, pp. 139–183). Elsevier. [https://doi.org/10.1016/S0166-4115\(08\)62386-9](https://doi.org/10.1016/S0166-4115(08)62386-9)
- [28] Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of “bias” in NLP. In *ACL 2020* (pp. 5454–5476). <https://doi.org/10.18653/v1/2020.acl-main.485>
- [29] Schick, T., Dwivedi-Yu, J., Dessì, R., Raunak, V., Thoppilan, R., Shazeer, N., & Biderman, S. (2023). Toolformer: Language models can teach themselves to use tools. arXiv:2302.04761. <https://arxiv.org/abs/2302.04761>
- [30] Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., et al. (2020). Toward trustworthy AI development: Mechanisms for supporting verifiable claims. <https://arxiv.org/abs/2004.07213>.
- [31] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). MMLU. In *ICLR 2021*. arXiv:2009.03300. <https://arxiv.org/abs/2009.03300>
- [32] Raji, I. D., & Buolamwini, J. (2019). Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In *AAAI/ACM Conference on AI, Ethics, and Society* (pp. 429–435). ACM. <https://doi.org/10.1145/3306618.3314244>
- [33] Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399. <https://doi.org/10.1038/s42256-019-0088-2>
- [34] Bryson, J. J. (2020). The artificial intelligence of the ethics of artificial intelligence: An introductory overview for law and regulation. In *The Oxford Handbook of Ethics of AI*. Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780190067397.013.1>
- [35] Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2), 1–21. <https://doi.org/10.1177/2053951716679679>
- [36] Floridi, L., & Cows, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, 1(1). <https://doi.org/10.1162/99608f92.8cd550d1>
- [37] Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning. Limitations and Opportunities*. Cambridge, MA: MIT Press. ISBN: 978-0-262-04861-3. <https://fairmlbook.org/>
- [38] Voigt, P., & Von dem Bussche, A. (2017). *The EU General Data Protection Regulation (GDPR):*

- A practical guide. Springer.
<https://doi.org/10.1007/978-3-319-57959-7>
- [39] Veale, M., & Borgesius, F. Z. (2021). Demystifying the Draft EU Artificial Intelligence Act. *Computer Law Review International*, 22(4), 97–112. <https://doi.org/10.9785/crl-2021-220402>
- [40] Manheim, D., & Garrabrant, S. (2019). Categorizing variants of Goodhart's Law. arXiv:1803.04585. <https://arxiv.org/abs/1803.04585>
- [41] First, M. B., Spitzer, R. L., Gibbon, M., & Williams, J. B. W. (2002). *Structured Clinical Interview for DSM-IV-TR Axis I Disorders, Research Version, Patient Edition (SCID-I/P)*. New York, NY: Biometrics Research Department, New York State Psychiatric Institute. Available at <https://www.columbiapsychiatry.org/research/research-areas/services-policy-and-law/structured-clinical-interview-dsm-disorders-scid>
- [42] Samuel, A. L. (1963). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3), 210–229. <https://doi.org/10.1147/rd.33.0210>
- [43] Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30, 6402–6413. <https://arxiv.org/abs/1612.01474>
- [44] Suresh, H., & Gutttag, J. V. (2021). A framework for understanding unintended consequences of machine learning. *Communications of the ACM*, 64(6), 62–71. <https://doi.org/10.1145/3454120>

Appendix B — Dataset URL/DOI Index per Suite (Links Only)

This appendix lists, per evaluation suite (TUP, AG, AD, OOD-R), the datasets/benchmarks and versioned artifacts used. Links only; no new scholarly references are introduced here. Seeds used across suites: 17, 23, 47 (fixed across runs). Metric definitions and release gates (ECE, Brier, abstention-compliance, sycophancy index, deception-evasion, persona-switch contradiction, long-horizon success) are defined in Appendix C. Evaluation artifacts (prompts, seeds, configs): Zenodo Concept DOI — 10.5281/zenodo.17042249 . Versioned child records host prompt sheets, configs, and checksums.

Suite	Artifact Type	Name	Version / Commit / Seeds	URL / DOI	Notes
TUP	Dataset/ Benchmark	TruthfulQA	v1.1; seeds: 17, 23, 47	https://doi.org/10.18653/v1/2022.acl-long.229	Truthfulness under adversarial pressure
TUP	Prompt Sheet	TUP Prompt Sheet	v1.0; seeds: 17, 23, 47	https://doi.org/10.5281/zenodo.17114740	Versioned child under Zenodo concept
AG	Task Set	Specification-Gaming Tasks	v1.0; seeds: 17, 23, 47	https://deepmind.google/discover/blog/specification-gaming-the-flip-side-of-ai-ingenuity/	Proxy-vs-goal exploitation tasks
AG	Prompt Sheet	AG Prompt Sheet	v1.0; seeds: 17, 23, 47	https://doi.org/10.5281/zenodo.17220804	Versioned child under Zenodo concept
AD	Task Suite	Sleeper/Deception Probes	v1.0; seeds: 17, 23, 47	https://arxiv.org/abs/2401.05566	Cross-role consistency & contradiction probes
AD	Prompt Sheet	AD Prompt Sheet	v1.0; seeds: 17, 23, 47	https://doi.org/10.5281/zenodo.17220842	Versioned child under Zenodo concept
OOD-R	Benchmark	Hendrycks & Gimpel OOD baseline	2016 baseline; seeds: 17, 23, 47	https://arxiv.org/abs/1610.02136	General OOD detection baseline
OOD-R	Benchmark	MMLU (held-out domains for robustness)	ICLR 2021; seeds: 17, 23, 47	https://arxiv.org/abs/2009.03300	Selective prediction (AUROC/AUPRC)
OOD-R	Prompt Sheet	OOD-R Prompt Sheet	v1.0; seeds: 17, 23, 47	https://doi.org/10.5281/zenodo.17114741	Versioned child under Zenodo concept

Appendix C — Metric Definitions & Example Release Gates

This appendix defines the metrics and provides example release gates used throughout the diagnostic battery. Means \pm 95% CI are reported across seeds; per-task breakdowns appear in the main text.

C.1 Calibration & Risk Scoring

Expected Calibration Error (ECE, 10-bin): absolute gap between predicted confidence and empirical accuracy aggregated over 10 equal-frequency bins.

Brier Score: mean squared error between predicted probabilities and observed outcomes on safety-critical subsets.

Release gates (safety-critical subsets): $ECE \leq 0.03$; $Brier \leq 0.10$.

C.2 Selective Prediction & Abstention

Abstention-Compliance: ratio of prompts for which the model defers/abstains when uncertainty or policy triggers require it.

Release gate: abstention-compliance $\geq 95\%$ on safety-critical prompts.

C.3 OOD Robustness

AUROC/AUPRC for out-of-distribution (OOD) detection on held-out domains.

Release gate: AUROC ≥ 0.90 on held-out OOD, stable within 95% CI across seeds.

C.4 Sycophancy / Undue Conformity

Sycophancy Index: rate at which the model aligns with a user's incorrect assertions under contrarian prompts.

Release gate: sycophancy index $\leq 5\%$.

C.5 Deception / Contradiction Under Role Re-spec

Deception-Evasion Probe: rate of evasive or deceptive responses when incentives change.

Persona-Switch Contradiction Index: contradiction rate under role re-specification scenarios.

Release gates: deception-evasion $\leq 2\%$; persona-switch contradiction $\leq 3\%$.

C.6 Long-Horizon / Myopia Proxies (if applicable)

Partial Failure Rate across N-step tasks; path efficiency/completion under long-horizon objectives.

Release gate: domain-specific long-horizon thresholds as defined in Section 3.

C.7 Reporting Conventions

Means \pm 95% CI across seeds; per-task breakdowns in the main text. Prompts/seeds/configs are versioned in Zenodo (see Appendix B for links).